A New Machine Learning Model based on Induction of Rules for Autism Detection

Fadi Thabtah
School of Health, Psychology Dept.
University of Huddersfield
Queensgate, Huddersfield HD1 3DH, UK
Corresponding

David Peebles
School of Health, Psychology Dept.
University of Huddersfield
Queensgate, Huddersfield HD1 3DH, UK

Telephone Number: +64 211597399

Email address: f.thabtah2@hud.ac.uk

Address: School of Health, Psychology Dept.
University of Huddersfield
Queensgate, Huddersfield HD1 3DH, UK

A New Machine Learning Model based on Induction of Rules for Autism Detection

Abstract-

Autism Spectrum Disorder (ASD) is a developmental disorder that describes certain challenges associated with communication (verbal and non-verbal), social skills, and repetitive behaviors. Typically, ASD is diagnosed in a clinical environment by licensed specialists using procedures which can be lengthy and cost-ineffective. Therefore, scholars in the medical, psychology and applied behavioral science fields have in recent decades developed screening methods such as the Autism Quotient (AQ) and Modified Checklist for Autism in Toddlers (M-CHAT) for diagnosing autism and other Pervasive Development Disorders (PDDs). The accuracy and efficiency of these screening methods relies primarily on the experience and knowledge of the user, as well as the items designed in the screening method. One promising direction to improve the accuracy and efficiency of ASD detection is to build classification systems using intelligent technologies such as Machine Learning (ML). Machine Learning offers advanced techniques that construct automated classifiers that can be exploited by users and clinicians to significantly improve sensitivity, specificity, accuracy, and efficiency in diagnostic discovery. This paper proposes a new ML method called Rules-Machine Learning (RML) that not only detects autistic traits of cases and controls, but also offers users knowledge bases (rules) that can be utilized by domain experts in understanding the reasons behind the classification. Empirical results on three datasets related to children, adolescents, and adults show that RML offers classifiers with higher predictive accuracy, sensitivity, harmonic mean, and specificity than those of other ML approaches such as Boosting, Bagging, decision trees, and rule induction.

Keywords: Autism Diagnosis, Classification, Decision Making, Predictive Models, Rule-based Classifiers, Machine Learning

1. Introduction

Instances of Autism Spectrum Disorder (ASD) are rapidly increasing. One in every 68 children is diagnosed with ASD, a developmental condition that presents with certain challenges associated with communication (verbal and non-verbal), social skills, and repetitive behaviors. It is estimated that 1.5% of the entire world population is classified with autism (Towle & Patrick, 2016; Centers for Disease Control and Prevention (CDC), 2014). Unfortunately, the process of officially diagnosing individuals with autism is tedious, requiring clinical resources and diagnosis methods such as Autism Diagnostic Interview (ADI) and Autism Diagnostic Observation Schedule (ADOS) (Lord, et al., 1994; Lord, et al., 2000). Consequently, it is believed that many more people who are on the spectrum remain undetected (Fitzgerald, 2017).

Moreover, the time spent waiting for a formal diagnosis is lengthy; for instance, the average waiting time in the UK is over 3 years (Crane, et al., 2016). Therefore, scholars in psychiatric health, psychology, and the behavioral science fields have developed self-administered and parent-administered screening methods that at a preliminary phase provide individuals with the recognition of possible autistic traits. Examples of screening methods are: Screening Tool for Autism in Toddlers and Young Children (STAT), Childhood Autism Rating Scale (CARS-2), and Autism Spectrum Quotient (AQ) (Stone et al., 2000; Schopler & Bourgondien, 2010; Schopler et al., 1980; Baron-Cohen, 2001).

The accessibility and use of ASD screening tools are vital, as they may reduce waiting time for formal clinical evaluation and provide individuals on the spectrum, and their families, better understanding of the resources and services needed for support (special education, speech therapy, work environment, etc.). However, most existing screening tools are based on diagnostic methods that contain large numbers of items that the parent, caregiver, or the individual (in case of adult with an average Intelligence Quotient) are required to check. Therefore, these methods have been criticized as being too time-consuming (Allison et al., 2012; Wall et al., 2012b; Bone et al., 2014; Duda et al., 2016; Bone et al., 2016; Thabtah, 2017a).

ASD traits are often screened using recognizable and measurable behavioral indicators (e.g., social skills, engagement in age-appropriate play and leisure, behavior excesses, communication skills, etc.). These indicators are usually represented by items given in a questionnaire format for most current screening methods (i.e. AQ, STATS, CARS-2, etc.) The screening processes for individuals mainly rely on simple human rules with a scoring function that adds scores associated with the items in the questionnaire to calculate the outcome. Therefore, the quality of the classification outcome for individuals undergoing such screening is primarily based on a) the items designed in the method, b) the experience and knowledge of the user who is administering the screening, and more crucially c) the handcrafted rules linked with the scoring function.

Designing rules to compute the scores of the questionnaire components requires extensive knowledge and experience. Replacing the human rule with knowledge derived from previous cases and controls to improve the diagnostic outcome and classification process seems advantageous. Automated knowledge is not subjective, as are handcrafted human rules, because they are discovered using advanced learning approaches such as ML or data mining. Consequently, boosting specificity, sensitivity, and predictive accuracy as well as classification efficiency. There is an urgent need for some advanced intelligent methods that can offer automatic classification of ASD as well as the reason(s) for the classification. These intelligent methods can be utilized by clinicians, parents, teachers, caregivers, and family members, among others, to understand the outcome of the screening. In addition, the clinician can use that outcome to verify the result of the screening using his/her own knowledge and experience.

Recently, a few scholars in the ASD research field have investigated ML to either improve the classification time of an ASD diagnosis or to detect the most influential items in ASD diagnosis, e.g. Wall et al., 2012b; Pratap et al., 2014; Bone et al., 2014; Pancer & Derkacz, 2015; Kosmicki et al., 2015; Duda et al., 2016; Levy et al., 2017; Al-Diabat, 2018). ML is a research area based around statistics, probability, artificial intelligence, databases, and other computer science areas that aims to intelligently discover hidden knowledge from datasets (Mohammed, et al., 2014; Thabtah 2007). ML techniques, including support vector machines (Chan & Lin, 2011), decision trees (Quinlan, 1986), rule induction (Cohen, 1995), Boosting (Freund & Schapire, 1997), Bagging (Breiman, 1996), neural network (Mohammed, et al., 2016) and Covering (Qabajeh, et al., 2015), seldom involve users in the processes of classification or model learning (Thabtah, 2017c).

Since the ASD diagnostic process encompasses predicting whether individuals are on the spectrum, i.e. with or without ASD, using predefined features including a class (ASD classification) and a historical dataset, this problem can be treated as a classification task in supervised learning. In this context, the clinician will utilize labelled cases of individuals with and without ASD (training dataset) to construct a classification system (model) using an ML technique. The model is then employed to automatically forecast the class of a new case (cases that are not yet classified) as accurately as possible.

In this paper, a new classification method based on the Covering approach, called Rules-Machine Learning (RML), is proposed. This method offers automatic classifications systems (classifiers) represented as rule sets. The rule sets inside the classifiers can be used by health professionals to assist in the diagnosis process or to advise individuals and their families whether they should seek further evaluation. The rules offered by the proposed method can be easily interpreted by novice users as well as parents, teachers, caregivers, and family members.

The RML was evaluated against real datasets collected using a mobile application called ASDTests (Thabtah, 2017b) and recently published at the University of California Irvine Repository (UCI) (Lichman, 2013). The experimental tests showed that the RML derives classifiers that are highly competitive when compared to other existing learning approaches in ML such as Boosting, Bagging, decision trees, and rule induction (Section 4 provides further details on the results and analysis). The performance evaluation of ML algorithms was based on common metrics such as predictive accuracy, sensitivity, harmonic mean, knowledge derived, and specificity.

This paper is structured such that Section 2 discusses the problem, aims, and critically analyses related works on ML that are linked with ASD, while Section 3 presents the rule-based architecture and details related to the learning method. Finally, conclusions are presented in Section 4.

2. The Problem and Literature Review

2.1. The Problem and Aims

Official ASD diagnosis is typically conducted by specialist physicians in a clinical environment using a Clinical Judgment (CJ) procedure and based on observable and measurable behavioral indicators (Al-Diabat, 2018). Existing paradigms seem to subscribe to the idea that more questions translate to a more accurate classification. Current standardized diagnostic tools take a very long time to conduct due to the large number of items that the specialist must check while relying on static human embedded rules (Lopez Marcano, 2016; Thabtah, 2018). This has necessitated a change in the way diagnostics are coded and behave within ASD clinical tools in the process of classifying cases.

There is a need to re-examine features within ASD diagnostic tools to fulfil smaller item sets while maintaining the sensitivity and validity of the test. However, very limited examination of the ML perspective on autism has been previously conducted regarding the classification and validation processes of autism (Bone et al., 2016; Levy et al., 2017;). This new paradigm of utilizing ML will not only make pre-diagnostic tools faster and more accessible but will also dramatically change the prospective of designing clinical diagnostic tools. When the ML algorithm is embedded in self-assessment tools, it will provide users with valuable concealed knowledge and guide the process of correct classification selection decisions in a more efficient manner (Duda et al., 2016; Thabtah, et al., 2018).

To address this global dilemma, the proposed research paper will take a new direction in the development of an ASD screening tool that incorporates rule-based architecture. Furthermore, the current study aims to better understand what components contribute to an efficient and accessible data-based ASD screening tool such that may be used by health professionals and other stakeholders seeking to understand whether they should seek an autism diagnosis by a professional. More specifically, we seek to establish a method that can be embedded within a self-administered ASD screening method to reliably and accurately provide feedback to patients, caregivers, and medical professionals regarding the potential need for professional diagnostic services. This investigation is vital for the standardization of efficient ASD diagnostic tools worldwide, serving to support long-term research goals and potentially impacting society directly.

This study also aims to limit the role of human-derived rules embedded within current assessment tools by using ML technology to increase classification accuracy, sensitivity, and specificity. This is particularly necessary for cases that are difficult to classify (e.g., cases unclearly associated with an ASD type). Results obtained from the proposed ASD pre-diagnostic tool are expected to be initially utilized by medical professionals for more efficient referrals to comprehensive evaluations. The main research questions that this study will answer are:

1) Is Machine Learning applicable to self-administered screening methods for ASD?

and more specifically,

2) Can rule-based ML methods help ASD screening and diagnosis in terms of the efficiency, accuracy, and knowledge presented to the user?

2.2. Literature Review

There have been a few studies investigating ML in autism screening research in recent years, e.g. Thabtah 2017c; Bone, et al., 2016; Duda et al., 2016; Lopez Marcano, 2016; Pratap, et al., 2014; Bone et al., 2014; Ruzich, et al., 2015; Wolfers, et al., 2015; Wall et al., 2012a; Wall et al., 2012b. These studies have concentrated mainly on the following two aspects of ASD diagnosis:

1) Accelerating the diagnostic process by identifying the least number of items required to be checked during the screening

2) Improving the sensitivity and specificity of the diagnostic decision by adopting a ML algorithm (neural network or decision tree) instead of the scoring function embedded in the diagnostic methods.

Wall et al., (2012a; 2012b) conducted a comparative study using several ML algorithms, particularly decision tree-based, on a dataset that contained cases and controls collected using the ADOS-Revised-Module 1 diagnostic method (Lord et al., 2000). The dataset was imbalanced with respect to class labels and contained many missing values and was stored in the Autism Genetic Resource Exchange (AGRE) repository (Geschwind, 2001). The aim of the study was to calculate the most influential items in the ADOS-Revised-Module 1 that can be utilized by clinicians to reduce the time associated with the diagnosis. The authors utilized the WEKA platform to conduct the experiments of the different ML algorithms (Hall et al., 2009).

Results obtained claimed that the ADOS-R-Revised can be replaced with merely eight items. The eight items were identified in classifiers generated by the Alternating Decision Tree algorithm (ADTree) by simply looking at what items appeared in the classifiers. A more suitable approach is to investigate the impact of feature selection methods such as wrapping or filtering on the ASD dataset and then analyze common features. A later study by Bone et al., 2014, reported serious conceptual and implementation issues associated with the Wall et al., (2012a; 2012b) studies.

Duda et al., (2014) reported conceptualization and implementation issues linked with the Wall et al., (2012a; 2012b) studies. The authors stressed that ASD prediction based on ML requires careful investigation especially when dealing with diagnostic methods that strictly follow procedures within a clinical environment. The claim that the ADOS-Revised diagnostic method can be minimized to eight items is misleading since to produce the decision, the entirety of activities must be conducted by the clinician on a test case before the classification system is constructed.. Consequently, there is no time saving.

More importantly, the activities of the ADOS-Revised must be performed in a clinical setup and not self-administered, as claimed by Wall et al., (2012a; 2012b). Therefore, the eight items proposed by Wall et al., cannot replace the original items of ADOS-Revised. The full ADOS-R test must be conducted before building a classifier using the ADTree algorithm in WEKA. Lastly, discarding, by the authors, data that is on the border between ASD and No ASD simplified the problem. This is because these cases are hard to detect by the ML algorithms. Therefore, including them prior to the data processing phase will impact the sensitivity, specificity, and accuracy of the results.

Duda et al., (2016) conducted an empirical analysis comparing several intelligent algorithms to discriminate between ASD and attention deficit hyperactivity disorder (ADHD). Six algorithms have been contrasted on a dataset with 65 items adopted from the Simplex Simon Collection (SSC) version 15 (Fischbach & Lord, 2010). The dataset was collected using a parent administered questionnaire diagnostic method called Social Responsiveness Scale (SRS) (Rutter et al., 2003). A preprocessing phase was applied by the authors to a) discard instances that had four or more missing values, b) balance the dataset using under sampling technique, and c) reduce the data dimensionality using feature selection methods. Empirical results reported that Logistic Regression produced classifiers with almost 95% classification accuracy.

Chu et al., (2016) investigated efficient ways to differentiate between ADHD and obstructive sleep apnea (OSA). The authors utilized the information of 217 children who had been classified by physicians as having ADHD, OSA, and a mixture of ADHD and OSA according to DSM IV standards (American Psychiatric Association, 2000). The data was collected using different diagnostic tools. Three ML algorithms were adopted to derive classifiers that could assist clinicians and physicians in improving the diagnostic decision. Reported results indicated that 17 features show substantial difference among three classes of Pervasive Developmental Disorders (PDDs,) particularly in the Child Behavior Checklist (CBCL) (Achenbach, 1991). A decision tree algorithm called CART was faster to derive the classifiers than the neural network and CHAID algorithms

Wolfers et al., (2015) investigated issues related to PDDs including small sample sizes, external validity, and ML algorithmic challenges without a clear focus on ASD. Lopez Marcano (2016) reviewed the applicability of different algorithms such as neural network and decision tree methods (Random Forest) to shorten the time of the ASD diagnostic process. Maenner et al., (2016) investigated the Random Forest algorithm (Breiman, 2001) on an autism dataset from Georgia Autism and Developmental Disabilities Monitoring (ADDM) Network utilizing phrases and words obtained in children's developmental evaluations. The dataset consists of 5,396 evaluations for 1,162 children of whom 601 are on the spectrum. The Random Forest classifiers were evaluated on an independent test dataset that contained 9,811 evaluations of 1,450 children. The results reported that Random Forest achieved around 89% predictive value and 84% sensitivity.

Thabtah, (2017c) critically analyzed pitfalls associated with experimental studies that adopted ML for ASD classification. The authors pinpointed issues related to datasets and learning algorithm methodologies used. These issues included: interpreting the classifiers content derived by the learning algorithm, noise in autism datasets, feature selection process, missing values, class imbalance, and embedding the classification algorithm within an existing screening method.

3. Proposed ASD Classification System

3.1 Rule-based Architecture for Detecting ASD (RML)

One of the least studied classification approaches in ML is Covering. Covering techniques normally discover simple chunks of information from historical datasets structured in the If-Then format, which makes their outcome highly favorable to novice users. In this section, we propose a new ASD detection method based on the architecture shown in Figure 1. Our method (RML) is based on Covering classification which employs a search method for rule discovery. The RML then evaluates the discovered rule and discards any redundancies. Hence, only rules that have been classified as training instances are kept. The evaluation phase performed not only reduces the number of discovered rules but also shrinks the search space of data items, which improves the efficiency of the training process.

In Figure 1, data is collected by a mobile application called ASDTests (Thabtah, 2017b) that implements four different ASD screening methods (AQ-Adult-10, AQ-Adolescent-10, AQ-Child-10, Q-CHAT-10) (Allison, et al., 2012). For the purposes of this research project, focus was on the child, adolescent, and adult modules and researchers utilized three datasets collected between September 2017 and January 2018. Once the raw data was obtained, several pre-processing operations were applied, including missing values replacement and discretization for certain continuous attributes such as the age of individuals.

Feature selection was used to remove features that were redundant and may have created biased results. Two features were eliminated, including the final score obtained by the screening method and the scoring method type (See Section 4 for further details on data features).

Once the raw data is preprocessed, then a learning algorithm is applied to discover rules sets that represent correlations between the variables in the training dataset and the class variables (ASD or No ASD). The datasets are then evaluated to remove useless and redundant rules, storing only rules that have classified training instances.

The outcome of the rule evaluation phase is the classification system (classifier) that will be used to predict the value of the class for unseen cases (individuals who have not yet been classified). When the classifier is tested, various evaluation metrics are derived to reveal the effectiveness of the rules in predicting cases and controls. These metrics, as well as the rules in the classifier, are shared with the health professional and the users involved in the screening. Therefore, not only does the new architecture provide users with decisions related to ASD detection, it also offers rich information on the reasons behind that decision as well as the quality of the outcome.

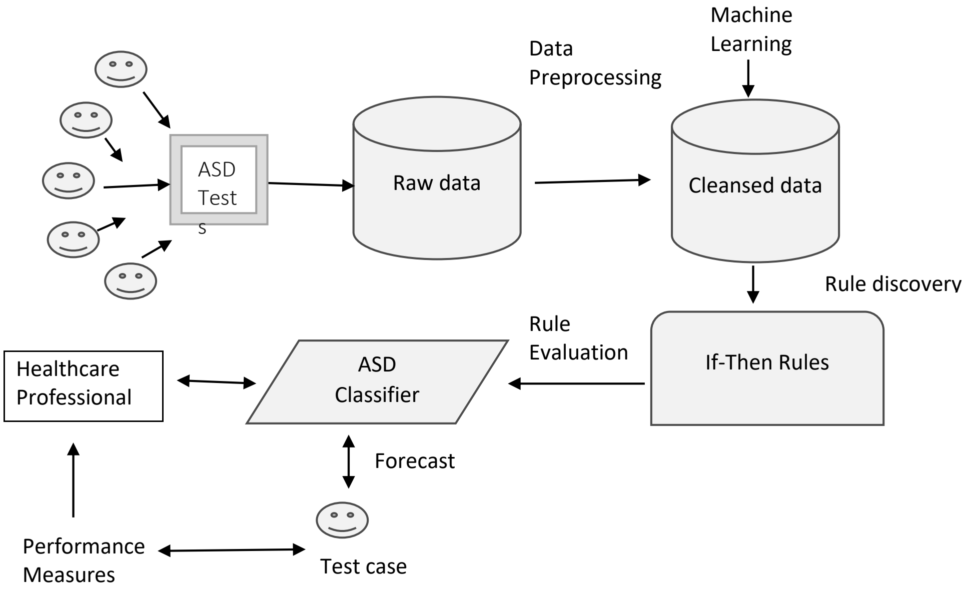


Figure 1: The Proposed ML Architecture for ASD Classification

3.2 The Learning Covering Algorithm

Researchers here developed a new learning mechanism based on the Covering approach called RML. The learning method pseudocode is shown in Algorithm 1. The algorithm utilizes two thresholds named the Minimum Frequency (Min_Freq) and Rule Strength (R_S) as other Covering approaches such as Dynamic Rule Induction (DRI) (Qabajeh et al., 2015; Thabtah et al., 2016) to find and extract the rules (Definitions 2 and 3 respectively). The Min_Freq threshold is used as a cutoff point for variables and class values in the training data (items).

An item is represented as (Variable Value, class Value) (Definition 1), and any item in the training data with a frequency equal to or above the Min_Freq threshold is qualified to be part of the rule's body during the process of building a rule. On the other hand, each rule is linked with a calculated strength (Rule Strength), which denotes the rule's items plus class frequency divided by the items frequency (see Definition 5).

A rule is represented as $(A_1, v_1) \wedge (A_2, v_2) \wedge \ldots \wedge (A_k, v_k) \rightarrow C_n$ where the antecedent is a conjunction of variables values (rule body), and the consequent is a class value (ASD, NO ASD). When the computed rule strength for a rule such as R is larger than or equal to the R_S threshold, R can then be generated, otherwise R will be removed. The computed rule strength for any given rule acts as a quality assurance metric that ensures only mathematically fit rules (that have proper data representation) are generated.

The Min_Freq and R_S thresholds are like minimum support and minimum confidence parameters in association rule mining (Agrawal et al., 1993). However, minimum support and minimum confidence are used to differentiate frequent items from infrequent items by considering the items' frequencies in the transactional data, whereas Min_Freq and R_S thresholds consider the target class when counting attribute values. More importantly, whenever the rule is derived by RML, the dataset is amended and therefore the frequency of rules is updated.

Definition 1: 1-Item in the training dataset (T), i.e. $[(A_1, v_1), C_n]$is an attribute plus a class. K-Item is a combination of attributes values plus a class, i.e. $([(A_1, v_1), (A_2, v_2), \ldots, (A_k, v_k)], C_n)$.

Definition 2: *Min_Freq* is a user threshold used to separate weak items from strong items.

Definition 3: R_S is a user threshold used to form rules.

Definition 4: A strong item, i.e $[(A_1, v_1), C_n]$, is recognized when $\frac{|[(A_1, v_1), C_n]|}{|T|} \geq$ *Min_Freq*

Definition 5: A rule such as *r* is formed when $\frac{|([(A_1, v_1), (A_2, v_2), \ldots, (A_k, v_k)], C_n)|}{|[(A_1, v_1), (A_2, v_2), \ldots, (A_k, v_k)]|} \geq$ *R_S*

The Learning algorithm initially scans the training dataset and discards any 1-item that has failed the Min_Freq threshold test (Lines 4-5). All remaining items with computed frequencies above the Min_Freq threshold are considered and saved into a data structure. To build a rule such as R, the algorithm attaches the best item in terms of computed frequency to the rule's body and repeats the process until the rule's accuracy cannot improve any further (Lines 6-7). When this occurs, the rule is then saved into the classifier (Line 9) and all training instances linked with R are erased from the original training dataset (Lines 10-11). When this happens, the strong items' frequencies are updated in the data structure. Consequently, some items may become weak and thus discarded by the learning algorithm.

In other words, items that share training instances with R are affected by ''s data removal, therefore frequencies of these items are normally reduced. The update procedure ensures that rules learnt are indeed non-redundant and often cover a larger portion of the training dataset. Continuing, this procedure can be considered as a quality measure, as items' frequencies are continuously updated since the training dataset is shrinking whenever a rule is generated. Resulting from this repetitive learning process, some manageable models with small yet effective rules are formed, which then can be exploited for decision-making by users in the autism screening application.

The RML algorithm guarantees that the search space of items is constantly reduced during the training phase and thus results in more efficient data processing. In addition to that, data instances that might overlap among items are removed, ensuring that rules extracted are not similar. Recall that RML keeps appending items in the rule's body until it processes with zero error so that the rule can be derived. However, in scenarios when rules are associated with some errors, RML allows the generation of such rules if they have computed strengths larger than or equal to the R_S threshold set by the end-user. This mechanism offers rules with slightly acceptable margins of error but minimizes the chance of models' overfitting.

The RML assumes that the variables in the training dataset are categorical (they are associated with a finite set of possible values). Continuous variables (integers and decimals) should be discretized before data processing. Lastly, missing values are dealt with as any other values in the training dataset.

To evaluate the rules sets generated by the learning algorithm, a test procedure that assigns test data the appropriate class, is utilized. Whenever a test case is present, the test method allocates the class label linked with the best ranked rule that matches the test case. This method necessitates that all items of the selected rule's body are presented in the test case in for the rule to be used for prediction. In cases when there are no rules in the classifier fulfilling this condition, the test method then allocates the class label of the first partially matching rule to the test case. When no rules are partially or fully matching the test case then a default class is allocated. The default class is basically a rule that represents the class with the largest frequency in the training dataset.

Hereunder are the key features of the ASD rule-based model:

1) The learning method produces non-redundant rules in the format 'If-then' that are easy to understand by different users such as clinicians, physicians, family members, caregivers, teachers, and others
2) Efficient procedure for learning the rules that requires one data scan and keeps reducing the search space of items during the training process
3) Straightforward metrics are utilized to derive the rules
4) Classifiers derived have fewer rules which make them more manageable by the different users.

5) Better sensitivity, specificity, and classification accuracy than the classical process-based scoring functions in current screening methods (See Section 4 for further details on the results).

```
Input: dataset with cases and controls T, R_S, Min_Freq thresholds    // Rule Strength = R_S. Minimum Frequency = Min_Freq.
Output: A Model with rules ( $RS$ )

1.  $E\_S\_R \leftarrow \{\}$
2.  $r_1 \leftarrow \{\}$
3.  $Temp \leftarrow T$
4.  Do    {
5.  If [(p(Ai, vi) | $c_i$ = l ) /| $Temp$ |] >=  min_freq  {
6.  If [(p(Ai, vi) | $c_i$ = l ) /|(p(Ai, vi)] >=  R_S                    {
7.  $r_i \leftarrow (Ai, vi)$
8.  Repeat steps 5-7 until $r_i$ accuracy cannot improve
      }}
9.  $E\_S\_R \leftarrow r_i$
10. $Temp \leftarrow (Temp - \ Locate\ (r_i\ , Temp))$

11. $Train \leftarrow (Train - Temp)$
12. Repeat steps 2-11
13. Exit when T has no more instances OR all p(Ai, vi) have been tested
14.  }
15.  Generate $E\_S\_R$
16.  Classify Test (Test, $E\_S\_R$)
```

Algorithm 1. The RML Classification Method

4. Data Features and Empirical Results

4.1 Data Description

The data has been collected using a recently developed mobile application called ASDTests (Thabtah, 2017c). ASDTests implements four screening methods for toddlers, children, adolescents, and adults based on Q-CHAT-10, AQ-10-child, AQ-10-Adolescent and AQ-10 adult respectively (Allison, et al., 2012). Figures 2A and 2B depict the landing page and one question page related to toddler test of the ASDTests app. During the data collection, there was no direct access to participants; the ASDTests mobile application provides clear information to the users about their participation and the use of their data in a disclaimer. In addition, the webpage also clearly states the use of the data is for research purposes only and informs the users about the use of data. The participants read this before submitting their answers. Anonymity has been imposed in the mobile app used to collect the data. Participants' identities are anonymous since no names or sensitive information are involved (See variables in Table 1.

While using the ASDtests application, the user can choose the age category test, which includes ten questions presented in a simple graphical user interface. Each question is associated with multiple answers that are easily selected in a mobile environment using a smart phone (IOS and Android) or a tablet. Once all the questions have been answered by the user, a review screen appears so the user can review and verify their answers before a data submission page with a disclaimer is invoked. The datasets used have been recently published by the authors at the University of California Irvine Data Repository (Lichman, 2013).

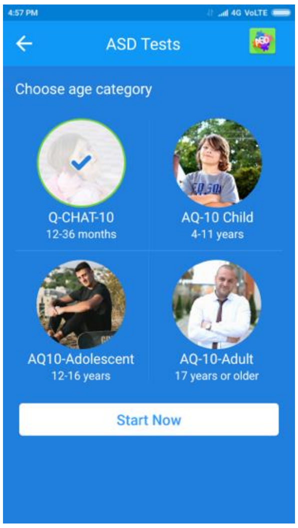
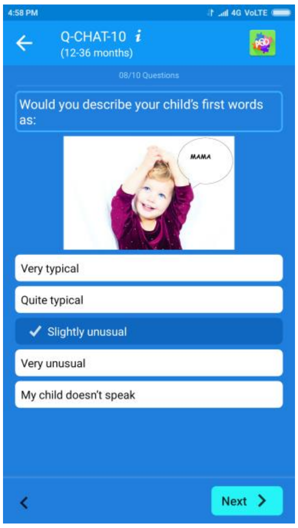


Figure 2A: Age Selection Screen (Thabtah, 2017b)



Would you describe your child's first words as:

- Very typical
- Quite typical
- ✓ Slightly unusual
- Very unusual
- My child doesn't speak

Figure 2B: A Sample Question: Toddler's (Thabtah, 2017b)

Table 1 shows the key features in the dataset used plus the type of screening, i.e. class label. The class label was assigned in an automatic manner during the process of data collection by the AQ-10 scoring method based on the final score obtained by the individual after taking the screening. There are two possible values of the class, i.e. '0' indicates that the individual has no ASD traits, and '1' indicates that the individual does have ASD traits. The '0' label is assigned when the final score based on the AQ-10 methods' scoring function is more than 6. More details on the score calculation can be found in Alison, et al., 2012. Overall, there are 20 features in Table 1 including the class label.

Table 1: Features Collected, their Descriptions and Mapping to the Actual AQ-10 Questionnaire

Feature	Type	Description
Age	Number	Adults (year), i.e. 17 years +.
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). In our case only Adult data has been used
A1	Binary (0, 1)	The answer code of: I often notice small sounds when others do not
A2	Binary (0, 1)	The answer code of: I usually concentrate more on the whole picture rather than the small details
A3	Binary (0, 1)	The answer code of: I find it easy to do more than one thing at once
A4	Binary (0, 1)	If there is an interruption, I can switch back to what I was doing very quickly
A5	Binary (0, 1)	The answer code of: I find it easy to 'read between the lines' when someone is talking to me
A6	Binary (0, 1)	The answer code of: I know how to tell if someone listening to me is getting bored
A7	Binary (0, 1)	When I'm reading a story I find it difficult to work out the characters' intentions
A8	Binary (0, 1)	I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant, etc)
A9	Binary (0, 1)	I find it easy to work out what someone is thinking or feeling just by looking at their face
A10	Binary (0, 1)	The answer code of: I find it difficult to work out people's intentions
ASD Score	Integer	The final score obtained based on the scoring function of on AQ-10-Adult. This was computed in an automated manner.
Class label	Boolean	The decision of the screening based on the scoring score of AQ-10-Adult method. Possible values '0' (no ASD traits or '1' (ASD traits)

The ten questions: A1-A10, have been transformed into binary attributes based on the values assigned to them by individuals during the screening process. To be exact, values in the A1-A10 variables in the dataset have been mapped to '0' or '1' depending on the actual values given during the screening process by the participants. In other words, during the screening using the AQ-10 screening 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 allocated when 'Definitely' or 'Slightly Disagree' was chosen for questions 2, 3, 4, 5, 6, or 9. The binary representation for features allows more efficient data processing when adopting learning algorithms in addition to making interpretation easier.

Table 2 displays 10 sample cases of individuals who experienced the AQ-10 adult screening, for presentation purposes. The dataset size is 704 cases collected over a period of four months. The dataset is imbalanced with respect to class labels, with 515 cases belong to 'No ASD Traits' and 189 cases with 'ASD'. This can be attributed to the fact that most people being screened through the app have no autistic traits. There are missing values in some cases, especially in two features, i.e. ethnicity, and who_is_taking_the_test. Slightly more male than female individuals have taken the ASD screening using the app. The three most popular ethnicities in the dataset belong to white, Asian and Middle Eastern. The computed mean for age was 29.2 with the youngest person to have taken the screening being 17-years-old, and the oldest 64-years-old. Lastly, more adults in the dataset have taken the test independently. The adolescent and child datasets contain 104 and 292 instances respectively.

Table 2: Sample of 10 Data Cases Collected for Adults Using ASDTests App based on AQ-10 Adult Screening Method

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q19	age	sex	ethnicity	jundice	family austim	contry_of _res	used_app _before	ASD Score	Who is taken the test	Class/ ASD
1	1	1	1	0	0	1	1	0	0	26	f	White	no	no	USA	no	6	Self	NO
1	1	0	1	0	0	0	1	0	1	24	m	Latino	no	yes	Brazil	no	5	Self	NO
1	1	0	1	1	0	1	1	1	1	27	m	Latino	yes	yes	Spain	no	8	Parent	YES
1	1	0	1	0	0	1	1	0	1	35	f	White	no	yes	USA	no	6	Self	NO
1	0	0	0	0	0	0	1	0	0	40	f	Middle Eastern	no	no	Palestine	no	2	Self	NO
1	1	1	1	1	0	1	1	1	1	36	m	Others	yes	no	USA	no	9	Self	YES
0	1	0	0	0	0	0	1	0	0	17	f	Black	no	no	USA	no	2	Parent	NO
1	1	1	1	0	0	0	0	1	0	64	m	White	no	no	New Zealar	no	5	Self	NO
1	1	0	0	1	0	0	1	1	1	29	m	White	no	no	UK	no	6	Self	NO
1	1	1	1	0	1	1	1	1	0	17	m	Asian	yes	yes	China	no	8	Health care professional	YES

4.2 Experimental Settings

This section presents the experimental settings of the proposed rule-based model and other common ML algorithms based on rule induction, Bagging, Boosting, and decision tree approaches on the child, adolescent, and adult datasets. We used six different algorithms in addition to RML to reveal the performance of the rule-based model. RIPPER, RIDOR, Nnge, Bagging, CART, C4.5, and PRISM algorithms have been adopted in the experimental results (Cohen, 1995; Gaines, 1995; Martin, 1995; Breiman, 1996; Breiman et al., 1984; Quinlan, 1993; Cendrowska, 1987). The main reason for choosing these algorithms, aside from them all producing rule-based classification models (classifiers,) is the fact that they employ different learning schemes in processing the dataset.

C4.5 and CART construct decision tree classifiers that get converted into rules sets; PRISM is a greedy algorithm that seeks for rules that have 100% expected accuracy. C4.5 uses pessimistic error estimation for pruning the trees before converting these trees into rules sets, whereas PRISM uses expected accuracy to measure the usefulness of adding an item into the rule's body while constructing a rule. On the other hand, RIPPER and RIDOR implement optimization and pruning procedures to test rules. For instance, RIPPER uses growing and pruning datasets to evaluate the worthiness of attributes' values prior to appending them into the rule's body. So, if adding an attribute value decreases the rule's predictive power, RIPPER ignores adding the attribute value and generates the rule. Lastly, Bagging and Boosting employ weak classifiers that in turn are merged to derive rules. This has been accomplished by deriving N classifiers and then utilizing them in predicting the class label of test instances using a voting mechanism, i.e. the class that belongs to the majority classifiers gets allocated to the test instance.

The considered algorithms are well investigated on different real-world applications and have proved their merits in terms of performance, such as predictive power and efficiency. Different evaluation metrics have been adopted to reveal the ML algorithm's true performance in detecting ASD traits from the datasets. To be exact, classification accuracy, specificity, and sensitivity among others (see Equations 1-5) were used (Witten & Frank, 2005).

The proposed rule-based model has been developed in the Java programming language and integrated within the WEKA platform version 3.9.1 (Hall, et al., 2011). WEKA is a known environment for implementing methods related to learning, classification, prediction, variable analysis, visualization, and dimensionality reduction. WEKA consists of packages that contain large numbers of ML and data mining techniques. Hence, all empirical runs have been conducted in WEKA for fair comparison. In testing the classifiers generated by the learning algorithm considered, a ten-fold cross validation method has been adopted (Abdelhamid & Thabtah, 2014; Witten & Frank, 2005).

In using ten-fold cross validation, the training dataset is partitioned into ten subsets. The classification algorithm randomly utilizes nine data subsets to learn the classifier and then tests the classifier on the remaining data subset. The same process is repeated ten times to generate an average error rate (Thabtah, 2006). The cross-validation procedure is embedded in WEKA platform and can be selected prior the learning phase. Lastly, all experimental runs have been conducted on a personal computer that has a 2.0 GHz processor and 8 RAM of memory.

The ASD screening process is a binary classification problem since individuals are classified to either having ASD traits or No ASD traits using characterized quantifiable variables. Therefore, performance evaluation methods that align with the binary classification problem in ML have been used. The confusion matrix (Table 3) can be used to derive different evaluation metrics including classification accuracy, error rate, sensitivity, and specificity to report the performance of the learning algorithms. Using the confusion matrix, a test case will be assigned a predicted class in the classification step of the screening.

Based on Table 3, classification accuracy (Equation 2) is a common metric in classification that computes the number of test data that was correctly classified from the total number of test data. Opposite to accuracy is the error rate (Equation 1). On the other hand, sensitivity (Equation 3) computes the percentage of the test cases that are truly positive (with ASD class) and specificity (Equation 4) denotes the percentage of the test cases that are negative (cases with no ASD).

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)

$$One\_error(\%) = 1 - Accuracy \tag{1}$$

$$Accuracy(\%) = \frac{|TP + TN|}{|TP + TN + FP + FN|} \tag{2}$$

$$Sensitivity(\%) = \frac{|TP|}{|TP + FN|} \tag{3}$$

$$Specificity(\%) = \frac{|TN|}{|TN + FP|} \tag{4}$$

$$\Pr ecision = \frac{|TP|}{|TP + FP|} \tag{5}$$

$$F1 = 2 \times \frac{|\Pr ecisoin \times \mathrm{Re}\, call|}{|\Pr ecisoin + \mathrm{Re}\, call|} \tag{6}$$

4.3 Results Analysis

Multiple experimental runs have been performed using different ML algorithms on Child, Adolescent and Adult datasets ASD screening datasets to reveal the true performance of the proposed model.

Figure 3 depicts the error rate results in % of the considered algorithms on the Child, Adolescent and Adult datasets. The figures show that Bagging, Boosting, rule induction and decision tree classifiers were able to accurately classify most of the cases and controls as their error rates for the Adult dataset were between 5.68 – 8.23%. However, the enhanced Covering algorithms such as our model (RML) outperformed the remaining algorithm in terms of error rate, i.e. an error rate less than 5.6%.

For the Adult dataset, RML derived a classifier with lower errors rates of 4.41%, 2.7%, 0.15%, 2.14%, 3.7%, 3.27%. 1.57% and 1.83% respectively than PRISM, CART, AdaBoost, Bagging, Nnge, RIDOR, C4.5 and RIPPER algorithms. For the smaller datasets (Adolescent, Child), RML maintained higher predictive rates than most of the considered algorithms. For instance, for the Child dataset, RML achieved 5.82%, 4.11%, 0.69%, 2.4%, 1.03%, 5.82% and 2.4% less error rate than PRISM, CART, AdaBoost, Bagging, Nnge, RIDOR, and RIPPER algorithms. Only C4.5 slightly achieved 0.34% higher than RML on this dataset. Nevertheless, RML, outperformed C4.5 on the Adolescent dataset by 7.69%. This, if limited, shows that RML not only performs well on datasets with enough data instances, such as the Adult dataset, but also with datasets with a limited number of instances, such as the Adolescent dataset. In addition, the superiority of the proposed algorithm is clear in the case of small datasets, whereas the considered ML algorithms suffered from low predictive rates due to not having enough instances. For example, rule induction algorithms such as RIPPER and tree-based algorithms such as C4.5 and CART, derived classifiers with 20%, 7.69% and 13.46% less error rates respectively than that of the RML model on the Adolescent dataset.

The reduction in the error rate can be attributed to the procedure employed by RML in the rule generation phase in which only non-redundant rules are produced and redundant rules that have no data coverage are discarded. Our model ensures that each rule has data coverage and eliminates any overlapping among rules on training instances, hence deriving an accurate classifier. In building the classification systems for detecting ASD, the RML algorithm ensures that whenever a rule is generated all its data instances are removed before learning the next rule from the training dataset. Additionally, it amends candidate item frequencies during the learning phase whenever training instances associated with the generated rules are erased. These amendments may result in potential rules becoming weak and thus discarded at the preliminary phase, which reduces the search space and improves the efficiency of the training phase.

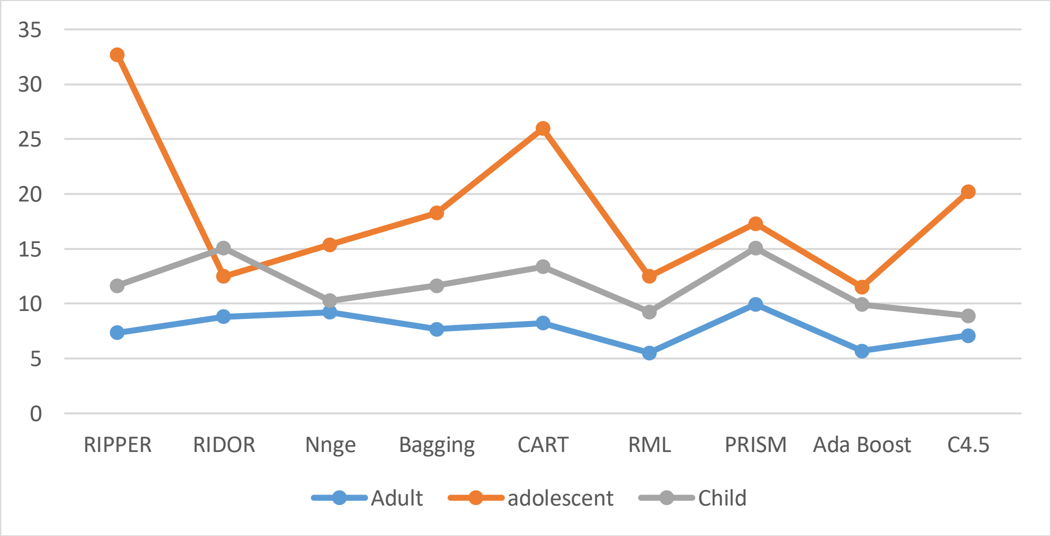


Figure. 3: Error Rates Derived by the Considered ML Algorithms on the Child, Adolescent and Adult Datasets

Figure 4A displays the specificity and sensitivity rates derived by the RIPPER, RIDOR, Nnge, Bagging, AdaBoost, CART, RML, C4.5 and PRISM algorithms on the Child, Adolescent and Adult datasets. Usually, acceptable specificity and sensitivity rates in autism research should be at least 80% (Towle

& Patrick, 2016). The results of the specificity and sensitivity rates generated by the considered algorithms on the two datasets (Adult, Child) have shown acceptable levels. Moreover, the Covering approach represented by RML produced classification systems with higher sensitivity and specificity rates than most of the remaining algorithms on these datasets. For example, for the Adult dataset, RML derived 1.9%, 3.3%, 2.0%, 2.8%, 3.2%, 1.7%, 0.2% and 1.7% higher sensitivity rates than RIPPER, RIDOR, Nnge, Bagging, CART, PRISM, AdaBoost, and C4.5 algorithms respectively. On the other hand, and for the same dataset, RML achieved 2.52%, 3.49%, 1.55%, 2.72%, 2.72%, 1.94%, 5.02% and 2.72% higher specificity rates than RIPPER, RIDOR, Nnge, Bagging, CART, PRISM, AdaBoost, and C4.5 algorithms respectively.

For the Child dataset, RML achieved 2.4%, 5.9%, 1.1%, 2.4%, 4.2%, 0.78% and 0.7% higher sensitivity rates than RIPPER, RIDOR, Nnge, Bagging, CART, PRISM and AdaBoost algorithms respectively. The rates get larger for the Adolescent dataset since most of the ML algorithms are unable to perform well on small datasets with lesser number of instances as RML. To be exact, the sensitivity rate of RML is 20.2%, 2.9%, 5.8%, 13.5%, 0.6% and 7.7% higher than RIPPER, Nnge, Bagging, CART, AdaBoost and C4.5 algorithms respectively. C4.5 slightly outperformed RML with respect to sensitivity rate on the Child dataset and by 0.3%. For the Child dataset, the specificity rate of RML was higher than most of the considered ML algorithms. To be exact, RML derived 19.8%, 2.7%, 6.0%, 13.2%, 0.4%, 7.5% higher specificity rate than RIPPER, Nnge, Bagging, CART, PRISM, and C4.5 algorithms respectively. Only RIDOR and AdaBoost slightly outperform RML in terms of specificity rate on the Adolescent dataset, and by just 0.2% and 0.8% respectively. Overall, the results reported higher sensitivity and specificity rates for RML on the three datasets when compared with the considered ML algorithms; these results are consistent with the error rates produced earlier and can be attributed to the non-redundant rules sets generated by RML.

The researchers investigated the confusion matrix results produced by the classifiers to understand the sensitivity, accuracy, and specificity results. For the Adult dataset, it was observed that the PRISM algorithm had the largest number of false negatives, followed by the CART and Bagging algorithms. Specifically, PRISM predicted 38 instances of individuals with No ASD traits that should have been classified on the spectrum. As a result, the sensitivity rate for class 'ASD' for this algorithm was low, at least on this dataset. On the other hand, PRISM had a high specificity rate having only 12 false positives. In other words, PRISM only predicted 38 adults without ASD traits that potentially supposed

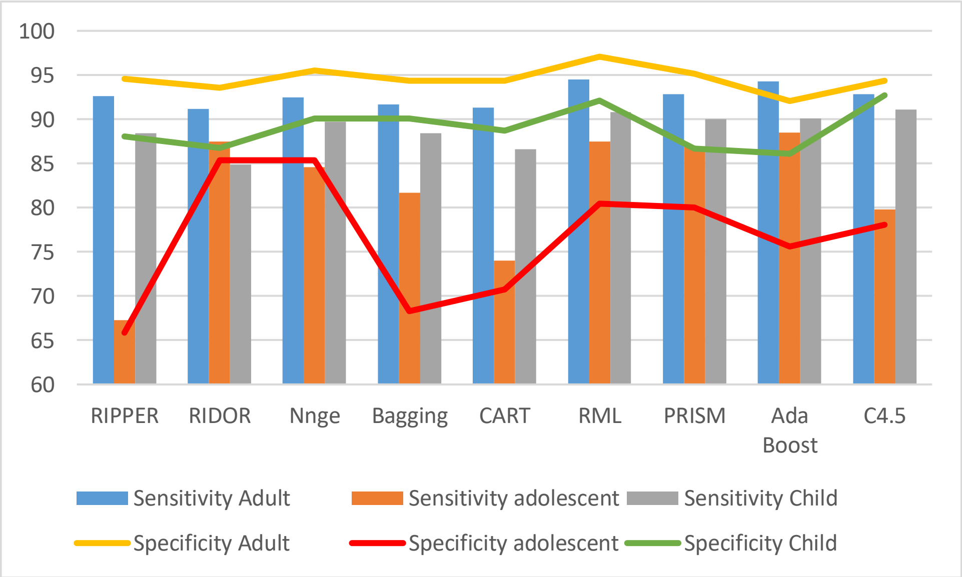


Figure 4A: Specificity and Sensitivity Rates of the ML Algorithms on the Adult, Adolescent and Child datasets

to be on the spectrum, and 12 individuals with ASD that are supposed to be classified as 'No ASD'. The 'No ASD' class has much higher data representation in the training dataset than the ASD class, which means that the PRISM algorithm is sensitive to the number of data items linked to class labels; for RML there were 15.

Figure 4B depicts the predictive accuracy of the considered ML algorithms including RML derived from the Child, Adolescent, and Adult datasets. The figure clearly shows that the RML algorithm generated classifiers with higher accuracy than the considered algorithms on the Adult and Child datasets. The AdaBoost algorithm slightly outperformed RML on the Adolescent dataset, yet RML has derived a competitive classifier on the same dataset.

The RIDOR algorithm has the largest number of false positives, wrongly predicting 33 instances having ASD who are supposed to be without ASD traits. Overall, there were higher classification rates for class 'No ASD' than 'ASD' with most of the considered algorithms. A probable reason for that fluctuation is that more instances representing class 'No ASD' are present in the training dataset. When the learning algorithm starts the training process more rules are then derived for class 'No ASD' in the classifier and therefore test instances that are supposed to be 'No ASD' will have less misclassifications.

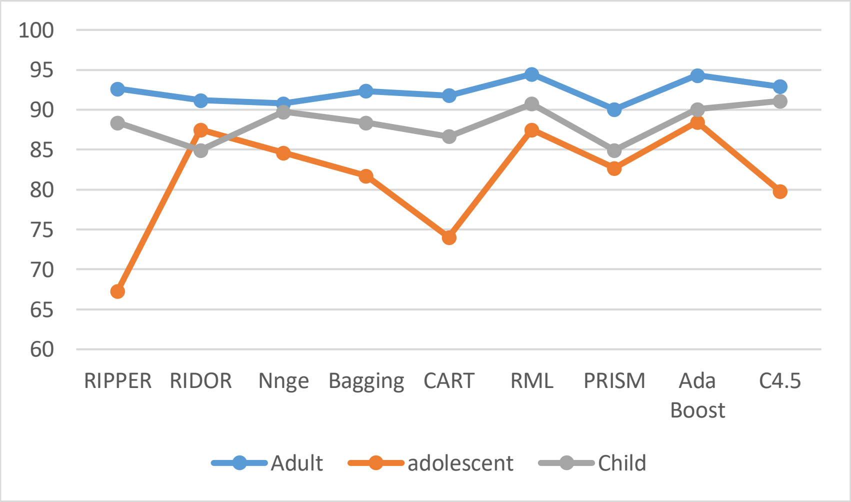


Figure 4B: Predictive Accuracy Rates in % Derived by the Considered ML Algorithms on the Adult, Adolescent and Child Datasets

Since the adult autism dataset is imbalanced with respect to class variable, researchers here included a metric called the harmonic mean (F1) that considers both recall (sensitivity) and precision (Equation 6). The F1 rates produced by the classifiers and shown in Figure 5 are high for the Covering (RML) and Boosting algorithms. This indicates that RML and AdaBoost perform well in datasets with imbalanced class labels and higher than the decision trees, Bagging, and rule induction algorithms represented by Nnge, RIDOR, RIPPER, CART, C4.5 and Bagging. For instance, on the Adult dataset, RML outperformed RIPPER, RIDOR, Bagging, CART, PRISM and C4.5 with respect to F1 rate by 1.7%, 3.1%, 3.6%, 2.2%, 2.6%, 1.8% and 1.6% respectively.

The results produced by the ML algorithms with respect to error rate, sensitivity, specificity, and F1 reveal a promising direction for autism screening. The results also pinpointed that Covering algorithms, such as RML, work well in ASD detection for at least the adults. Furthermore, the performance of ML may be impacted when the number of instances for a class label is low, i.e. class ASD. However, when a class is representative, such as 'No ASD', then the performance improves.

Overall, most algorithms generated acceptable sensitivity, specificity, and F1 results with more superiority to the Covering and Boosting classification approaches. These algorithms are more tolerant toward data with noise, i.e. imbalanced datasets. A possible direction to boost the performance is to have more data for the low frequency class.

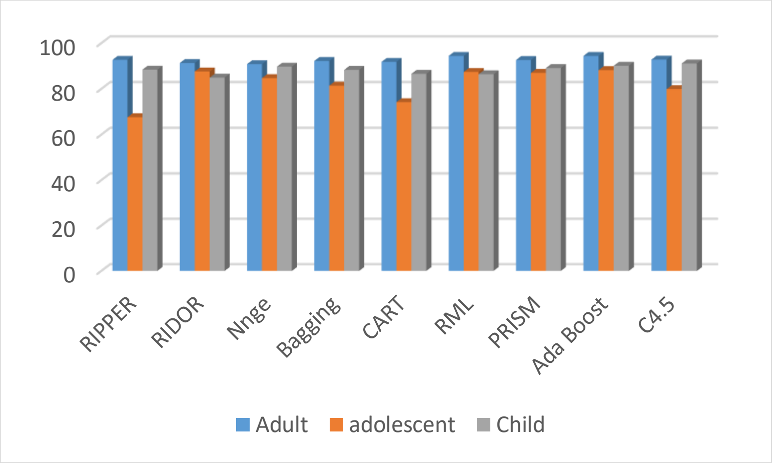


Figure 5: F1 Rates in % Derived by the Considered ML Algorithms on the Adult, Adolescent and Child Datasets

The researchers investigated the classifier content generated by the Covering, decision tree, Bagging, and rule induction algorithms to seek important knowledge that could help in detecting ASD. Figure 6 shows the number of rules generated by the considered algorithms on the adult dataset. The figures clearly show that the PRISM and Nnge algorithms generate the highest number of rules. The reason for extracting too many rules by PRISM is the fact that this algorithm has no rule pruning strategy, so it keeps building up rules, whereas Nnge is an algorithm that adopts the nearest neighbor search using non-nested generalized exemplars.

The number of rules results pinpointed that decision tree-like algorithms, such as CART and C4.5, derive classifiers larger in size than the rule induction and Covering approaches. The rule induction approach, represented by RIPPER and RIDOR, generate slightly smaller classifiers than the Covering approach. This can be attributed to the rigorous pruning procedures adopted by RIPPER and RIDOR in evaluating the rules.

Table 4 contains the common rules related to ASD detection that have been extracted by the Covering and rule induction approaches respectively (RML, RIPPER). It seems that items in the AQ-Adult-10 screening methods have high influence on the class labels, particularly items 5, 9, 8, and 4. Additionally, items 7 and 2 appeared in multiple rules in RIPPER and RML classifiers. It seems that the

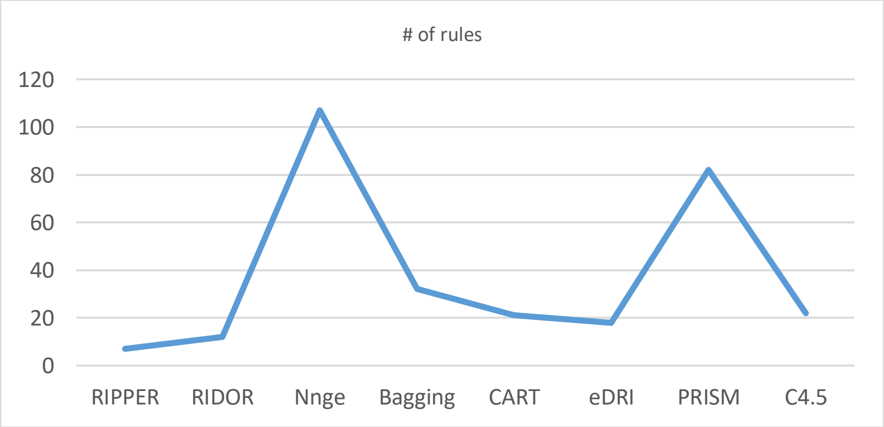


Figure 6: Number of Rules Derived by the Considered ML Algorithms on the Adult Dataset

Table 4: Common Rules Derived by RML and RIPPER Algorithms on the Autism Datasets

RML rules (Freq, R_S)	RIPPER rules
1. (238, 1.00) Label = NO when A5_Score = 0, A3_Score = 0 2. (054, 1.00) Label = NO when A5_Score = 0, A8_Score = 0 3. (037, 1.00) Label = NO when A1_Score = 0, A10_Score = 0 4. (019, 1.00) Label = NO when A4_Score = 0, A5_Score = 0 5. (023, 1.00) Label = YES when A6_Score = 1, A7_Score = 1, family_with_autism = yes 6. 07 - (056, 1.00) Label = YES when A6_Score = 1, A7_Score = 1, A9_Score = 1, A1_Score = 1 7. (039, 1.00) Label = NO when A4_Score = 0, A2_Score = 0, A9_Score = 0 8. (030, 1.00) Label = YES when A6_Score = 1, A3_Score = 1, A1_Score = 1, A2_Score = 1 9. (026, 1.00) Label = NO when A8_Score = 0, A9_Score = 0, born_with_jaundice = no 10. 14 - (023, 1.00) Label = YES when A9_Score = 1, A2_Score = 1, A8_Score = 1 11. (018, 0.86) Label = YES when A7_Score = 1, A4_Score = 1, A5_Score	**Adult Dataset** 1. If (A9_Score = 1) and (A5_Score = 1) and (A6_Score = 1) and (A10_Score = 1) => Class/ASD=YES (102.0/3.0) 2. If (A9_Score = 1) and (A3_Score = 1) and (A1_Score = 1) and (A5_Score = 1) => Class/ASD=YES (40.0/4.0) 3. If (A4_Score = 1) and (A6_Score = 1) and (A7_Score = 1) and (A8_Score = 1) => Class/ASD=YES (18.0/1.0) 4. If (A5_Score = 1) and (A2_Score = 1) and (A10_Score = 1) and (A8_Score = 1) and (A3_Score = 1) => Class/ASD=YES (16.0/2.0) 5. If (A9_Score = 1) and (A4_Score = 1) and (A1_Score = 1) and (A8_Score = 1) and (A2_Score = 1) => Class/ASD=YES (8.0/0.0) 6. If (A7_Score = 1) and (A5_Score = 1) and (A4_Score = 1) and (A8_Score = 1) and (A10_Score = 1) => Class/ASD=YES (11.0/2.0) **Adolescent Dataset** 1. (A4_Score = 0) and (A10_Score = 0) => Class/ASD=NO 2. (A5_Score = 0) and (A7_Score = 0) => Class/ASD=NO 3. (A3_Score = 0) and (A2_Score = 0) => Class/ASD=NO **Child Dataset** 1. (A4_Score = 1) and (A7_Score = 1) and (A10_Score = 1) => Class/ASD=YES (79.0/1.0) 2. (A8_Score = 1) and (A1_Score = 1) and (A10_Score = 1) => Class/ASD=YES (35.0/6.0) 3. (A9_Score = 1) and (A5_Score = 1) and (A2_Score = 1) => Class/ASD=YES (17.0/1.0) 4. (A9_Score = 1) and (A5_Score = 1) and (A8_Score = 1) and (A3_Score = 1) => Class/ASD=YES (9.0/2.0) 5. (A4_Score = 1) and (A5_Score = 1) and (A1_Score = 1) and (A3_Score = 1) and (A2_Score = 1) => Class/ASD=YES (5.0/0.0) 6. (A9_Score = 1) and (A4_Score = 1) and (A10_Score = 1) and (A3_Score = 1) => Class/ASD=YES (4.0/0.0) Class/ASD=NO (143.0/2.0)

Table 5: Common Features Mapping with AQ-adult-10 Screening Method

Item	Description
5	I find it easy to 'read between the lines' when someone is talking to me
9	I find it easy to work out what someone is thinking or feeling just by looking at their face
8	I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant, etc.)
4	If there is an interruption, I can switch back to what I was doing very quickly

Table 6: Time to Build the Models in Milliseconds (ms)

	RIPPER	RIDOR	Nnge	Bagging	AdaBoost	CART	RML	C4.5	PRISM
Child	2	2	3	2	3	4	0	0	1
Adolescent	5	3	5	4	5	5	0	1	2
Adult	7	4	8	4	6	7	1	3	3

items that have frequently appeared within the rules cover certain autistic behaviors within the DSM-5 manual. For instance, item 8 covers repetitive behavior, item 4 is aligned with communication and lastly items 5 and 9 are aligned with social behavior (Table 5).

Finally, Table 6 depicts the runtime in milliseconds (ms) for the considered ML algorithms in processing the three autism datasets. It is obvious from the table that RML is more efficient than the remaining ML algorithms in building the models and for all datasets considered. Overall, ML techniques showed good efficiency in deriving the screening models from the child, adolescent and adults datasets respectively.

5. Conclusions

Autism screening is a fundamental step toward understanding autistic traits and for speeding up referrals to further evaluation in a clinical setting. However, existing screening tools such as AQ, Q-CHAT and many others rely on simple calculation, using scoring functions that tally the scores of answers given by individuals. During the screening process these scoring functions have been developed based on hand-crafted rules and thus can be criticized for being subjective. Therefore, one of the crucial issues in ASD screening research is improving the screening process so that individuals and their families can have a faster and more accurate service. This can be accomplished by utilizing automated methods based on ML that build accurate classification systems from historical cases and controls. This paper proposes a new ML method called RML that not only boosts the sensitivity, specificity, and predictive accuracy of the ASD screening process, but also offers automatic classification beside rich rules sets for clinicians, caregivers, patients and their families and teachers.

The proposed method generated non-redundant rules in a straightforward manner utilizing Covering learning. Empirical evaluation on different autism datasets using rule induction, Bagging, Boosting and decision trees algorithms reported the superiority of the RML model with respect to different evaluation metrics including specificity, sensitivity, harmonic mean, and error rate. The results also showed that the RML derived classifiers contain useful rules for understanding the reasons behind the ASD classification. Lastly, the classifier's content revealed some influential items in the autism screening that are aligned with social and communication behaviors yet not fully fulfilling the Statistical Manual of Mental Disorders (DSM-5) criteria for ASD diagnosis.

In conclusion, this paper clearly revealed that ML approaches, especially Covering, showed promising results in detecting ASD cases especially for adults. In future, the researchers intend to design and implement a new autism screening tool based on rules sets derived by our model for toddlers, children, and adolescents.

One of the limitations of this paper is not including instances related to toddlers as these are rare and difficult to obtain. In addition, RML possibly needs a method to deal with datasets that are imbalanced with respect to class labels, to further improve its predictive performance. Soon, we are going to build a new screening mobile application that will embed the rule based classifiers to help diagnosticians access a rich knowledge base for improving screening and diagnostic decisions related to autism.

References

[1] Abbas, H., Garberson, F., Glover, E., & Wall, D. P. (2017). *Machine learning approach for early detection of autism by combining questionnaire and home video screening*. arXiv preprint arXiv:1703.06076.

[2] Abdelhamid N. & Thabtah F. (2014). Associative classification approaches: Review and comparison. *Journal of Information and Knowledge Management (JIKM), 13*(3).

[3] Achenbach, T. (1991). *Manual for the Youth Self-Report and 1991 Profile*. Burlington, VT: University of Vermont Department of Psychiatry.

[4] Agrawal, R., Imielienski, T., & Swami, A. (1993). *Mining association rules between sets of Items in large databases.* In: Proc. Conf. on Management of Data, 207–216. New York: ACM Press

[5] Al-Diabat, M. (2018) Fuzzy data mining for autism classification of children. *International Journal of Advanced Computer Science and Applications (IJACSA), 9* (7), 2018.

[6] Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief “red flags” for autism screening: The short autism spectrum quotient and the short quantitative checklist for autism in toddlers in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child Adolescent Psychiatry 51*(2), 202–12Y.

[7] American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders*. Washington: American Psychiatric Publishing.

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

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

[11] Breiman, L. (1996) Bagging predictors. *Machine Learning, 24* , pp. 123-140.

[12] Breiman, L. (2001) Random forests. *Machine Learning, 45*(1):5-32, 1300

[13] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and regression trees*. Wadsworth International Group, Belmont, California.

[14] Centers for Disease Control and Prevention (2014). Prevalence of autism spectrum disorder among children aged 8 years - Autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *MMWR 63*(SS02): 1–21.

[15] Cendrowska, J. (1987) PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies, 27* (4), 349-370.

[16] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* (2), 27.

[17] Chu, K. C., Huang, H. J., & Huang, Y. S. (2016). Machine learning approach for distinction of ADHD and OSA. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on* (pp. 1044-1049). IEEE.

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

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

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

[21] Fischbach, G. D., Lord, C. The Simons simplex collection: A resource for identification of autism genetic risk factors. *Neuron* (68), 192–195.

[22] 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

[23] Freund, Y. & Schapire, R.E., (1997) A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System*.

[24] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An Update. *SIGKDD Explorations 11*(1).

[25] Modeling Large Databases. *Intelligent Information Systems, 5*(3): 211-228.

[26] Geschwind, D. H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., … Jones, P. (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.

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

[28] Levy, S., Duda, M., Haber, N., Wall, D. (2017). Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism.

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

[30] Lopez Marcano, J. L. (2016). *Classification of ADHD and non-ADHD using AR models and machine learning algorithms* (Doctoral dissertation, Virginia Tech).

[31] Lord, C., Risi, S., Lambrecht, L., (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.

[32] Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders* (24), 659–685.

[33] Maenner, M. J., Yeargin-Allsopp, M., Van Naarden Braun, K., Christensen, D. L., Schieve, L. A. (2016) Development of a machine learning algorithm for the surveillance of autism spectrum disorder. *PLoS ONE 11*(12): e0168224. https://doi.org/10.1371/journal.pone.0168224

[34] Martin, B. (1995). *Instance-based learning: Nearest neighbor with generalization*. Hamilton, New Zealand.

[35] Mohammad, R., Thabtah, F., McCluskey, L. (2016) *An improved self-structuring neural network*, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Auckland, New Zealand, 2016, pp. 35–47.

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

[37] Nakai, Y., Takiguchi, T., Matsui, G., Yamaoka, N., & Takada, S. (2017). Detecting abnormal voice prosody through single-word utterances in children with autism spectrum disorders: Machine-learning-based voice analysis versus speech therapists. *Perceptual and Motor Skills*, 0031512517716855.

[38] Pancers, K. Derkacz, A. (2015). *Consistency-based pre-processing for classification of data coming from evaluation sheets of subjects with ASDS*. Federated Conference on Computer Science and Information Systems, 63-67.

[39] Pratap, A., Kanimozhiselvi, C. S., Vijayakumar, R., Pramod K. V. (2014). Predictive assessment of autism using unsupervised machine learning models. *International Journal of Advanced Intelligence Paradigms 6*(2),113-121.

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

[41] Quinlan J. 1986. Induction of Decision Trees. *Mach. Learn. 1, 1,* 81-106.

[42] Qabajeh, I., Thabtah, F., Chiclana, F. (2015) A dynamic rule-induction method for classification in data mining. *Journal of Management Analytics 2* (3), 233-253

[43] Rutter, M., Bailey, A., & Lord, C. (2003). *The social communication questionnaire manual*. United States of America: Western Psychological Services.

[44] Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & 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).

[45] 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.

[46] Schopler, E., & Bourgondien, M. (2010). *(CARS™2) Childhood Autism Rating Scale™, Second Edition*. WPS.

[47] Stone, W., Coonrod, E., & Ousley, O. (2000). *Brief Report: Screening Tool for Autism in Two-Year-Olds*

[48] Thabtah, F. (2018). Machine learning in autistic spectrum disorder behavioral research: A review. To appear in *Informatics for Health and Social Care Journal*.

[49] Thabtah, F., Kamalov, F., Rajab, K. (2018) A new computational intelligence approach to detect autistic features for autism screening. *International Journal of Medical Informatics 117,* 112-124

[50] 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.

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

[52] Thabtah, F., Qabajeh, I., Chiclana, F. (2016) Constrained dynamic rule induction learning. *Expert Systems with Applications 63*, 74-85.

[53] Thabtah, F. (2006) Rule preference effect in associative classification mining. *Journal of Information & Knowledge Management 5* (01), 13-20.

[54] Thabtah, F. (2007). Review on associative classification mining. *Journal of Knowledge Engineering Review*, *22*:1, 37-65. Cambridge Press.

[55] Towle, P., & Patrick, P. (2016). *Autism spectrum disorder screening instruments for very young children: A systematic review*. New York: Hindawi Publishing Corporation.

[56] Wall, D. P., Kosmiscki, J., Deluca, T. F., Harstad, L., Fusaro, V. A. (2012a). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry* (2).

[57] Wall, D. P., Dally, R., Luyster, R., Jung, J. Y., Deluca, T. F. (2012b). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PlOs One 7: e43855*.

[58] Witten, I. H. & Frank, E. (2005). Data mining: Practical machine learning tools and techniques.

[59] Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev*. *57*:328–349. [PubMed]