Detecting Autistic Traits using Computational Intelligence & Machine Learning Techniques

Fadi Abdeljaber (Thabtah)

Submitted in partial fulfilment of the requirements for the degree of Master of Research (MRes)

Health and Human Sciences
Department of Psychology
University of Huddersfield
Huddersfield, UK

September, 2018

DECLARATION

I hereby declare that this thesis has not been submitted, either in the same or different form, to this or any other university for a degree.

ACKNOWLEDGMENTS

First and foremost, I would like to thank God almighty for the strength, health and knowledge given to me to peruse my studies. I would like also to thank my supervisor Dr. David Peebles for the ongoing help & support during my research studies. Without his guidance and support this work will not be possible.

Great thanks to my wife Neda who has been supportive and patient with all the working hours I have invested during my journey. Special thanks go to Hatem my lovely son who has inspired me to conduct this work and finally I would like to thank my son Fayez, my mother Fatima, my brothers and my sisters for their spiritual support.

DEDICATION

This thesis is dedicated to my son Hatem who has been my “dipsy” from day 1 until now and my father Fayez God bless his soul.

Publications

1. Thabtah F. Peebles D. (2019) A New Machine Learning Model based on Induction of Rules for Autism Detection. Health Informatics Journal. 2018. (Accepted and to appear in February 2019).
2. Thabtah F. Peebles D. (2018) Screening of Autism: A Comprehensive ASD Screening Tools Review for Toddlers, Children, Adolescents, and Adults. Journal of Intellectual Disabilities. (Under Review).
4. Thabtah F. (2108) Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics for Health and Social Care 43 (2), 1-20. 2018.
5. Thabtah F. (2018) An Accessible and Efficient Autism Screening Method for Behavioural Data and Predictive Analyses. Health Informatics Journal. Health informatics journal, pp. 1-21, 1460458218796636 . 2018.
3. Thabtah F. et al. (2018) A new computational intelligence approach to detect autistic features for autism screening. International Journal of Medical Informatics, Volume 117, pp. 112-124.
6. Thabtah F. (2017) ASDTests. A mobile app for ASD screening. www.asdtests.com.[1]

[1] Fadi is the main author of all accepted and published works of this study. For papers that were not solely authored by Fadi, he was the first author of these as he performed experiments, results analyses, coding, and the entire writing. The co-author(s) verified the mathematical model(s) and the results produced besides contributing to the paper organization and guidance.

ABSTRACT

Autistic Spectrum Disorder (ASD) is a developmental disorder that describes certain challenges associated with communication (verbal and non-verbal), social skills and repetitive behaviours. Self-administered ASD assessment tools, also known as screening tools, are typically conducted by a caregiver, medical staff and require responses to a large number of items. The validity and accuracy of assessments based on these tools relies upon classification methods which have antiquated technologies and this should be of concern for users in the healthcare community. A possible way to improve the classification accuracy and efficiency of the current screening tools is to adopt intelligent methods based on machine learning (ML) and computational intelligence. The latter can be utilised to identify a concise set of items by using new technologies such as mobile platforms, thus improving screening, or be able to steer those in seek of help toward a more accurate diagnosis. To automate the classification process and enhance the predictive accuracy of the test, the processing of data, based on the outcome of the computational intelligence, can be conducted using the former method.

This thesis proposes a new ML architecture for ASD screening that consists of a rule-based classification method called Rules Machine Learning (RML) which generates high predictive rules that can be easily understood by different users. Moreover, a new feature selection method known as Variable Analysis (Va) is proposed; this significantly reduces the number of features needed for ASD screening methods while maintaining performance. The last proposal in this thesis is an easy and accessible mobile screening application called ASDTests, which enables vital autism features to be collected from three primary datasets: adults, children, and adolescents from which thorough descriptive and predictive analyses are performed. To measure the performance of the RML and Va methods, large numbers of experiments have been conducted using various feature selection and ML techniques on the considered datasets. The bases of the comparisons are: evaluation metrics including sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NVP), and harmonic mean. The results clearly demonstrated that the new ML method was able to choose fewer items from the three datasets than the other methods considered while maintaining acceptable levels of specificity, sensitivity, and predictive accuracy. The concise sets of items and classifiers generated are of high interest to the different individuals interested in ASD screening. These results can also assist in early detection of ASD traits, thus facilitating access to necessary support systems for the physical, social, and educational well-being of the patient and their family in addition to increasing the likelihood of improved outcomes for the patient.

Table of Contents

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter One

Introduction

1.1 Introduction

Autism Spectrum Disorder (ASD) is a pervasive developmental disorder that hinders an individual's skills in socialisation, creates repetitive behaviours, and impacts expressive or verbal communication with disruptions ranging from moderate to severe (Pennington et al., 2014). The symptoms of autism are more visible and easy to identify in children two to three years of age. According to Towle and Patrick (2016), one out of every 68 children has autism in United States. Consequently, various screening methods have been developed by leading medical experts and psychiatrists worldwide seeking to identify autistic traits in the primitive stage so as to readily provide the necessary medications (Allison, et al., 2012).

Formal ASD diagnosis is typically conducted by specialised physicians in a clinical environment using a Clinical Judgment (CJ) procedure and based on observable and measurable behavioural indicators. Existing paradigms seem to subscribe to the idea that more questions translate to a more accurate classification. Present autism screening methods are based on ASD diagnosis methods, so they usually take a long time to conduct due to the large number of items that the user must go through while relying on static human embedded rules. This has necessitated a change in the way diagnostics are coded and behave within ASD screening methods in the process of classifying cases.

Current ASD screening tools generally employ human-developed rules 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. Alternatively the diagnosis of ASD might be empowered by automated decisions generated by intelligent algorithms such as machine learning (ML). ML is a research area based around

statistics, probability, artificial intelligence, databases, and other computer science areas that aim to intelligently discover hidden knowledge (models) from datasets (Qabajeh et al., 2015; Mohammed, et al., 2014). Using ML seldom involves users in the processes of classification or model learning and indeed may boost the classification accuracy and efficiency. More importantly, the models derived using ML will not replace clinicians, rather these models offer the clinician guidance to potentially improve the referral decisions of individuals undergoing ASD analysis.

Limited examination of the ML perspective screening has been previously conducted regarding the classification and validation processes of autism. This new paradigm of utilising ML will not only make screening tools faster and more accessible but will also dramatically change the prospective of designing future clinical diagnostic tools. When the ML algorithm is utilised in the screening tools, it will provide users with valuable information and guide the process of correct classification in a more efficient manner. Recently, a few scholars in the ASD research field, mainly in clinical research, have investigated ML to either improve the classification time of an ASD diagnosis or to detect the most influential items in ASD diagnosis, i.e. (Abbas et al., 2017; Chu et al., 2016; Lopez Marcano et al., 2016; Maenner et al., 2016; Duda et al., 2016; Bone et al, 2014). However, little research has been done on the use of ML to improve the screening of ASD. Therefore, this research addresses this gap by developing a new ASD screening architecture that incorporates a novel mobile screening application, feature assessment method and a ML classification method.

The current study aims to better understand which components contribute to an efficient and accessible data-based ASD screening tool such that may be used by stakeholders seeking to understand whether they should seek an autism diagnosis by a professional. More specifically, we seek to establish an intelligent model for ASD screening to reliably and accurately provide ASD traits 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 thesis investigates the applicability of ML approach on the problem of ASD classification during the screening phase. We develop a new rule-based classification architecture that not only is able to predict the cases and controls during the screening process but also offers rich rules for

the different stakeholders including physicians, parents, teachers and caregivers. The proposed ML architecture consists of data collection tools based on a mobile environment, a novel feature selection method and a competitive learning method for classification. The thesis also proposes a new mobile screening application for data collection from four different target groups (toddlers, children, adolescent, adult). These datasets have been used to identify highly influential features and can be used in further research related to improving the performance of ASD screening. Lastly, we propose a new feature selection method that reduces feature-to-feature correlation and maintains feature-to-class correlation. Section 1.3.1 gives further details on the main contributions and issues of this thesis.

1.2 The Research Problem, Aims and Research Questions

ASD screening differs from clinical ASD diagnosis in which the latter is typically conducted in a clinical setting with the presence of one or more licensed clinicians to formally produce a diagnosis. The former does not require a clinical setting and often involves seeking autistic traits associated with an individual. Normally in the screening process, parents, caregivers, teachers, medical staff and sometimes individuals (in case of adult with an average Intelligence Quotient) answer a series of behaviour, communication and social questions. A score is then derived from the screening method to indicate whether the individual should seek further medical assessment. This thesis is concerned with the ASD screening issue and therefore ASD clinical diagnosis is out of scope.

The ASD screening process encompasses predicting whether individuals are associated with autism traits by using characterised features with a class variable (ASD classification) and a historical dataset so this problem can be treated as a classification task in supervised learning (See Figure 1.1). In this context, the user will utilise 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 has not yet classified) as accurately as possible. Integrating ML models into the classification process will indeed improve the accuracy, specificity, sensitivity and efficiency of the test. In addition, the screening decision will not rely on static hand-crafted rules, but on an intelligent model derived from labelled historical cases and controls, and will offer guidance to the user; this will minimise the subjectivity of the decision.

The lack of integrating technology, such as ML with existing screening methods, may contribute to current limitations such as accessibility, lack of behaviour datasets, reliability of simple scoring function and hand-crafted rules, and subjectivity of the final decision. The ML innovations to be examined and developed for screening tools are intended to make the classification process of ASD automated, rather than static. These changes may effectively replace pre-existing human-generated rules and procedures resulting in distinct and impactful advantages: enhanced efficiency (time taken to perform the screening) with ASD classification, reduction in the number of features of ASD assessments to minimal levels while maintaining assessment integrity (identification of key components that produce accurate diagnoses), increasing accessibility especially when using a mobile environment, providing new datasets that can be utilised for further analysis, and more importantly enhancing the classification accuracy, sensitivity and specificity.

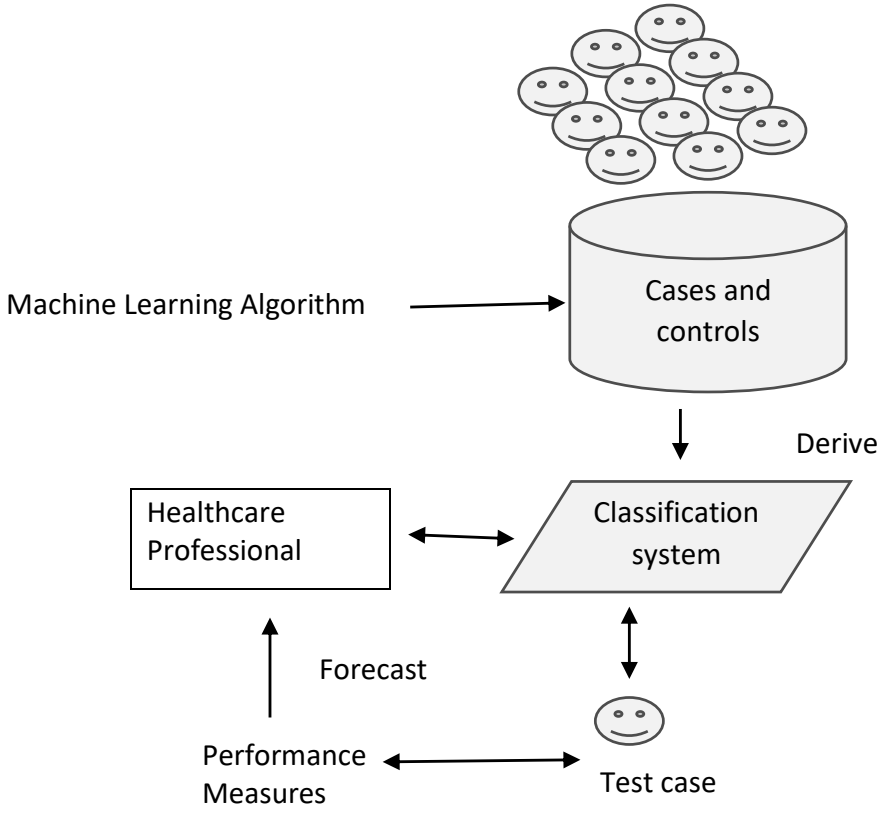


Fig. 1.1 ASD screening problem as a predictive task in ML

The screening assessment, with respect to screening tools, is expected to be automated using ML and facilitated by caregivers or professionals. This necessitates the following aims: (1) Minimisation of the total number of scale items through computational intelligence techniques; (2) Creation of a ML classification algorithm within the classification process to learn by examining the case or control; (3) Creating a screening application to increase the accessibility and to collect cases and controls for different target groups (training datasets).

This study also aims to limit the role of human-derived rules embedded within current assessment tools by using ML technology to boost classification performance. This is particularly necessary for cases that are difficult to classify (for example, cases unclearly associated with an ASD type). Results obtained from the proposed ASD pre-diagnostic tool are expected to be initially utilised by different stakeholders, such as medical professionals, for more efficient referrals to comprehensive evaluations. The main research questions that this thesis will answer are:

1) Is the Machine Learning approach applicable to screening methods for ASD?
2) Can rule-based ML methods improve ASD screening in terms of accuracy, sensitivity, specificity and efficiency?

1.3 The Research Issues and Contributions

1.3.1 Scoring Function

Autism screening is a fundamental step towards understanding autistic traits and for speeding up referrals to further evaluation in the clinical environment. However, existing screening tools rely on simple calculations, using scoring functions that add up scores from answers given by individuals. Therefore, one of the crucial issues in ASD screening research is improving the screening process so that individuals and their families can have a more accurate service. This can be accomplished by utilising models generated by ML technology from historical cases and controls. Therefore, instead of using the scoring function the automated models can be used for predicting autistic traits during the screening process.

To address this issue, a novel rule-based ML method called Rules-Machine Learning (RMR) that replaces the static scoring function found in conventional screening methods is proposed.

RML is a covering algorithm that uses two main thresholds to learn straightforward If-Then rules from the input training dataset. The learning phase involves discovering high correlations among the independent variables (features in the training dataset) and the target class variable (ASD traits/ No ASD traits) in an efficient manner by repeatedly decreasing the size of the training dataset whenever a rule is generated. In doing so, a reduction in overlapping among the rules is accomplished by discarding redundant rules and only keeping highly predictive rules for prediction. Unlike conventional covering methods, the learning procedure of RML is efficient in the sense that rules are discarded early whenever they are unable to attain certain frequency or confidence. The proposed ASD classification method (RML) is explained in detail in Chapter 3 and it has been accepted for publication to the Journal of Health Informatics.

1.3.2 Performance (Efficiency, Accuracy, Sensitivity and Specificity)

The accessibility and use of ASD screening tools is vital as they may reduce waiting times for formal clinical evaluation and provide individuals on the spectrum and their families better understanding of the resources and services available (special education, speech therapy, work environment, etc.). However, most of the existing screening tools are based on clinical diagnostic methods that contain large numbers of items that the parent, caregiver or the individual are required to check. Therefore, these methods have been criticised as being too time consuming (Thabtah 2017A; Duda et al., 2016; Bone et al., 2016; Bone et al., 2014; Duda et al., 2014; Allison et al., 2012; Wall et al., 2012A, Wall et al., 2012B).

ASD traits are often screened using recognisable and measurable behavioural indicators (for example, social skills, engagement in age-appropriate play and leisure, behaviour excesses, and communication skills). These indicators are usually represented by items given in a questionnaire format for most current screening methods (i.e. Autism Quotient (AQ), Screening Tool for Autism in Toddlers and Young Children (STATS), Childhood Autism Rating Scale-2 (CARS-2), etc. (Baron-Cohen et al., 2006; Stone et al., 2008; Schopler & Bourgondien 2010). The screening process for individuals mainly relies on simple hand-crafted rules with a scoring function that tallies scores associated with the items on the questionnaire to calculate the outcome. Therefore, the quality of the classification outcome of 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; c) The hand-crafted rules linked with the scoring function. To improve the classification performance especially with respect to predictive accuracy, sensitivity, specificity and reduce false positives and true negatives, an automated classification system derived from historical cases and controls using ML is promising.

We propose automated models derived by a RML classification method that replaces hand-crafted rules with richer knowledge bases. These bases consist of predictive rules derived from the learning phase using RML from former cases and controls of adults, adolescents, and children. These rules are easy to understand and manage by novice users and medical experts; they can boost the performance of the classification of cases and controls with respect to accuracy, sensitivity and specificity. Experimental results on Adult, Adolescent and Child datasets using various evaluation metrics conducted in Chapter four demonstrate that the automated models are clearly effective and efficient in detecting ASD traits.

1.3.3 Features Analysis

One of the key issues in ASD screening is the large number of questions or items that the user must ask in order to complete a full screening. For instance, in the AQ screening tool there are 50 items the individual (adult) must check during the screening. The combination of tedious assessment methodologies and other common obstacles associated with accessibility, may affect the quality of the outcome. Thus, identifying fewer, albeit influential, features in common ASD screening methods is urgent. To achieve this aim, computational intelligence based on feature selection that considers feature-to-class correlations and reduces feature-to-feature correlations, is advantageous. This feature selection should assign each feature its true weight based on observed and expected probabilities as well as minimise scores derived from other methods. Therefore, only small non-redundant features are selected and can be exploited in the screening by the ML classification algorithm.

One of the primary aims of this study is to build on previous attempts to reduce the number of items in AQ short versions by discovering the least number of items that affect the process of ASD classification. The methodology used for achieving this aim is based on a newly proposed computational intelligence method being referred to as Variable Analysis (Va). Va intelligently computes the correlations between items in three AQ screening method versions based on

normalised scores of the Information Gain (IG) and Chi-Square (CHI) methods (Quinlan, 1986; Shannon 1948; Liu & Setiono, 1995). The scores are stabilised and the discrepant behaviour of the results of these filtering methods is substantially reduced. Va assigns a weighted vector per item as a score, then ranks items based on their significance with the ASD class label Fewer items can be retained, thus improving the efficiency of the screening as well as pinpointing the most important items that contribute to autistic traits. By identifying a set of the least number of items, modern technologies such as mobile platforms can be used to allow more people to undergo screening for ASD, and this will enhance accessibility throughout the healthcare community. The results of the Va have been verified using ML technology by deriving automated classification systems with respect to specificity, sensitivity, positive predictive values (PPVs), negative predictive values (NPVs), and predictive accuracy. Experimental results (Chapter Four) using cases and controls related to items in three common screening methods, along with features related to individuals, have been analysed and compared with results obtained from other common filtering methods. The results showed that Va was able to use fewer features from adult, adolescent, and child screening methods yet maintained competitive predictive accuracy, sensitivity, and specificity rates. These results have been published in Thabtah et al., (2018).

1.3.4 Behaviour Datasets Scarcity

Datasets on the traits, characteristics, diagnoses and prognoses of autism are rare. Presently, few autism datasets associated with clinical diagnostics are available. The clear majority of available data on autism concerns the genetics of the condition. A few examples of these are the AGRE (Geschwind, et al., 2001), National Database of Autism Research [NDAR] (Hall et al., 2012) and Boston Autism Consortium [AC] (Fischbach and Lord, 2010), but no behaviour data for screening of ASD. The rarity of behaviour datasets for autism screening research limits the opportunities of enhancing the existing tools' performance; it is imperative to allocate effort and resources to recruit and collect instances-related cases and controls. These datasets can be utilised by researchers in the areas of psychology, psychiatry and computational intelligence to improve the prognoses and screening of autistic traits in existing methods.

For the rarity behavioural dataset issue, an ASD screening application (app) that embraces all age categories (infant, child, adolescent, adult), serves larger communities (the majority of the aforementioned apps are only available in English), and which is short in length (3-5 minutes), is

proposed (Thabtah, 2018B). The ASD app is called ASDTests and it is designed and implemented to fit all age categories and to serve larger communities worldwide (across 11 languages) while maintaining brevity (10 questions per test). The ASDTests app is available in both Android and Google stores and can be easily accessed by users (Thabtah 2017B). It contains a clear disclaimer for data privacy and usage before an individual submits screening answers, all data submitted is anonymous and the user has the chance to consent before any submission.

1.3.5 Accessibility

Rapid growth in the number of ASD cases worldwide and the economic impact of autism, necessitates the development of easily accessible and effective ASD screening tools. Limited research studies have been conducted on developing interactive mobile platform autism tests for different types of users interested in autism. Most of the current screening methods take considerable time to conduct; their shortened versions, which speed up the screening and increase accessibility for the healthcare community, have not yet been integrated into a mobile environment. A few ASD screening apps have been developed including ASDetect, Autism and Beyond App (AaB), Autism Test App (ATA) and the Asperger Test (ASD Detect org., 2016; Autism and Beyond, 2017; Autism Test, 2017; Asperger Test (2017)), however, only a few apps are available in both Google play and Apple iTunes stores, making them accessible to both iPhone and Android users. The majority offer their tests only in the English language and thus they are not accessible to many potential users. More importantly, no ASD testing apps have been found that cover toddlers, children, adolescents, and adults collectively, making the available apps narrow in their scope of coverage.

To improve accessibility, a new mobile application called ASDTests that offers users and the health community a friendly, time-efficient, and accessible mobile-based ASD screening was developed. This innovative and interactive app contains evidence-based short ASD screening tools for toddlers, children, adolescents, and adults using the Autism Quotient - 10 (AQ) and Quantitative Checklist for Autism in Toddlers (Q-CHAT) methods. The ASDTests app can be used by health professionals to assist their practise or to advise individuals if they should pursue formal clinical diagnosis. Unlike existing autism screening apps being tested, the proposed app covers a larger audience consisting of four different tests, one each for toddlers, children,

adolescents, and adults as well as being available in eleven different languages. The proposed app can be utilised by a variety of stakeholders including parents, caregivers, and more importantly health professionals to advise potential cases for further clinical assessments.

1.4 Machine Learning: A Brief Introduction

ML is a research field that integrates mathematics, artificial intelligence, search methods, and other sciences to derive accurate predictive models from datasets (Abdelhamid et al., 2014; Thabtah et al., 2006; Witten & Frank, 2005). ML techniques are utilised by managers and other users to discover hidden patterns in structured and unstructured datasets to enhance decision-making (Thabtah, 2018A). For instance, in a retail supermarket chain, managers are interested in processing transactional datasets that contain customers' historical purchases so as to improve item shelving and to set promotion strategies. Physicians are striving for sets of rules that explain the correlations among different symptoms and the type of illness to aid in medical diagnosis.

ML methods require minimal human involvement during data processing and usually offer useful patterns for users (Mohammed et al., 2013). One of the common tasks in ML is supervised learning. Supervised learning tasks, such as classification, involve training on historical instances known as the training dataset in order to learn a classification system also called the model or the classifier (Mohammed et al., 2016). The training dataset typically contains a set of variables one of which is the target variable (class); the learning phase is restricted to this class therefore this type of learning is called 'supervised'. The model learnt is then evaluated on an independent test instances (test dataset) to measure its predictive performance. Typical applications of supervised learning are loan approvals, gene classification, text categorisation, medical diagnoses and credit card scoring. If no target variable is available in the input dataset and the task involves grouping of instances (clustering) or just discovering useful patterns without prediction (association rule mining) this type of learning is 'unsupervised'. Hereunder are listed common ML tasks summarised from (Witten & Frank, 2005).

- Association Rule Mining: Initially, association rule mining was employed to discover useful rules based on the frequency of items in a transaction dataset. These rules are used by managers to learn more about customers' purchasing behaviour so that certain marketing strategies can be developed.

- Clustering: When the task is descriptive and involves grouping data instances into groups based on a certain similarity metric, this is considered clustering. For example, dividing customers based on specific demographic variables is a typical application of cluster analysis.
- Classification: When the aim is to predict a specific variable based on other independent variables this task is called classification. In classification, a model is learned from the training data such that it can be used to forecast the class in a test dataset. For instance, classification models are employed by financial institutions to forecast loan approval applications.
- Regression: In regression analysis, a model that represents the correlation among the independent variables and the dependant variable is constructed. Typically, the dependent and independent variables are continuous variables. Regression can be seen as a special case of classification in ML.

Several different ML approaches that involve predictive analysis have been published by scholars including Decision Trees, Probabilistic, Associative Classification, Support Vector Machine (SVM), Covering, Rough Set, Artificial Neural Network (ANN), Boosting, Bagging and Rule Induction among others (Quinlan, 1979; 1986; Friedman et al., 1997; Thabtah, et al., 2004; Hadi et al., 2008; Mohammad, et al., 2014; Cortes & Vapnik, 1995; Freund & Schapire, 1997; Breiman, 1996; Holt, 1993). Each of these ML approaches has positive and negative aspects. For example, Associative Classification and ANN approaches often derive good predictive models with respect to accuracy yet the former suffers from having a very large number of rules in the classifier; the latter requires multiple trial and errors during the training phase in order to optimise the outcome and thus it is not cost effective with respect to efficiency. Probabilistic techniques such as Naïve Bayes are cost effective with respect to time and yet their general performance overall is not highly competitive with classifiers such as SVM or ANN.

This thesis involves building a new ML method using covering classification; to explain further we clarify the operation of the covering approach. In the covering classification, the outcome is a set of rules that have been learnt in a sequential manner. The covering approach starts with an empty rule for the first class value in the training dataset then it keeps adding items into the rules' body until the rule error cannot be reduced (Thabtah et al., 2010). There are many ways to compute a rule's error, one of which is based on the data instances that the rule covers in the training dataset. Once the rule is produced, all training data instances linked with it are erased, and the covering algorithm then proceeds with building the next rule for the same class. When all rules with the

current class are found then the algorithm moves to the next class value and repeats the same process. Once all rules are generated or the training dataset becomes empty then the covering algorithm terminates and generates the rules sets as a model for prediction. In predicting a test case, most covering techniques assign the class label in line with the first rule in the model to the test case (Thabtah et al., 2011). Covering is considered by many researchers a favourable classification approach since the models offered are easy to interpret and manage by decision makers besides having good predictive accuracy (Qabajeh, et al., 2015; Mohammed et al., 2014; Witten & Frank, 2005; Cohen, 1995). In Chapter Three we show how the proposed ML method differs from classic covering techniques in ML.

1.5 Thesis Structure

This thesis is comprised of five chapters; Chapter Two reviews common ASD screening methods and critically analyses their upsides and downsides. In addition, we review and analyse recent studies on the application of ML for autism screening and diagnosis. Chapter Three presents the new ML architecture including data collection (ASDTest app), feature selection, data transformation, and more importantly the covering classification method. Chapter Four is devoted to data and results analysis in which we conduct a large number of experiments on three datasets using various different computational intelligence and ML techniques. In this chapter, we show the true performance of the proposed feature selection and ML methods when compared to other common ML techniques for the problem of ASD traits detection. Finally, we conclude in Chapter Five by highlighting the key contributions and limitations of the thesis and pinpointing new potential research routes.

Chapter Two[2]

Literature Review

Common Autism Screening Methods and Recent Machine Learning Use in ASD Classification

2.1 Introduction

There are many clinical diagnosis and screening methods available to assess individuals with ASD. Most popular diagnosis methods include Autism Diagnostic Interview-Revised (ADI-R), Autism Diagnostic Observation Schedule (ADOS), Childhood Autism Rating Scale (CARS), Joseph Picture self-concept scale, and the social responsiveness scale (Lord, et al., 2000; Schopler, et al., 1980; Lord, et al., 1994; Le Couteur, 1994; Constantino, 2005; 2012). These are clinical methods used for formal ASD diagnosis and treatment planning (Risi, et al., 2006). Techniques like ADI-R and ADOS are clinically proven to be effective instruments in differentiating autism from other related developmental disorders, having adequate validity and sensitivity (Rutter, et al., 2003). However, they have been criticised for being time consuming, having long questionnaires and scoring methods, and requiring licensed clinicians and observers to administer them (Duda et al., 2016; Bone, et al., 2016; Wall, et al., 2012A; Allison et a;., 2012; Wall et al., 2012).

Apart from clinical diagnostic methods, there are screening instruments developed by different neuroscientists and psychologists in the autism and healthcare arena. Screening methods help different stakeholders such as parents, caregivers and physicians among others identify autistic traits in individuals often using questionnaires. Tools such as Autism Spectrum Quotient (AQ) and the Modified Checklist for Autism in Toddlers (M-CHAT) which are discussed in later sections often consist of large sets of items for discriminating the autistic behaviours from all other types of PDDs (Baron-Cohen, 2001; Scott, et al., 2002; Robins et al., 2001). Most of these tools have

[2] Parts of this chapter have been published in the Journal of Informatics for Health and Social Care 43 (2), 1-20. Other parts were submitted for publication to another journal.

been developed based on diagnosis methods and have been able to present more accessible ways for users to undergo an ASD screening. Nevertheless, screening tools are not considered diagnosis methods for ASD since many of them lack the presence of a licensed clinician as well as the necessary clinical environment.

There have been many studies in the applied behavioural sciences that have investigated the efficiency and effectiveness in clinical environments of ASD diagnosis techniques (Ventola, et al., 2016; Sappok et al., 2013; South, et al., 2012; Matson et al., 2012; Ventola, et al., 2006). However, limited studies have been carried out to identify the performance of ASD screening methods and to evaluate their strengths and weaknesses (Stewart & Lee, 2017; Towle & Patrick, 2016; Zwaigenbaum, et al., 2015; Soleimani, et al., 2014; Krug et al., 2008). For instance, Soleimani, et al. (2014), reviewed common screening methods related to autism and only compared their performance with regard to specificity and sensitivity. Zwaigenbaum, et al. (2015), reviewed early screening methods for toddlers without covering other important aspects relating to adolescents, children, and adults. They indicated that early identification of ASD traits in toddlers, 18-24 months of age, is consistent with the recommendations of the American Academy of Paediatrics. Another similar review of ASD tools for infants was conducted by Towle and Patrick (2016) and showed that a two-level screening can help improve the reliability of the process. In addition, Stewart and Lee (2017) conducted a systematic review of common diagnosis methods of ASD in low and middle-income countries. They revealed that because of the limited clinical resources in low-income countries screening methods are more effective in discovering autistic traits. However, clinical diagnosis methods seem more widely utilised in middle and high-income countries besides cross cultural diagnostics have been shown some of these methods to be unreliable (Elsabbagh et al., 2012).

This chapter divided into two main parts; the first part of the chapter critically evaluates common ASD screening methods in order to recognise the merits, performance issues, and shortcomings (not only in terms of sensitivity, and specificity, but also critical issues related to administration, efficiency, target audience, complexity, digital existence, and accessibility among others) for each available method. Furthermore, screening methods that cover all target ages (toddlers, children, adolescents, and adults) are critically analysed, making this review comprehensive and applicable to the entire population of cases. For convenience, all identified methods are categorised according to their target audience, i.e. Toddlers and Children, and

Hybrids. Screening methods included in the Hybrid category typically facilitate the screening of toddlers, children, adolescents, and adults all together.

In the second part of this chapter, we review recent research works on the use of machine learning within autism screening and diagnosis research. To improve the classification process of ASD, researchers have recently started to adopt machine learning intelligent methods (Abbas et al., 2017; Thabtah 2017; Maenner et al., 2016; Duda et al., 2016; Pancer and Derkacz, 2015; Mythili & Shanavas, 2014; Bone et al, 2014; Duda et al., 2014; Wall et al., 2012B). The primary purposes of these machine learning studies on ASD was to improve diagnosis time of a case in order to provide quicker access to health care services, improve diagnosis accuracy, and reduce the dimensionality of the input dataset so as to identify the highest ranked features of ASD.

The chapter consists of four main Sections. Section 2.1 introduces the ASD screening and Section 2.2 reviews and critically analyses the ASD screening tools considered. Section 2.3 is devoted to a comprehensive discussion that contrasts and evaluates the identified screening tools in terms of their method of administration, accessibility, popularity, performance, and comprehensibility. Lastly, Section 2.4 contains suggestions for the different stakeholders involved in ASD research and summarises that abridge the findings of the review chapter.

2.2 ASD Screening Methods

2.2.1 Toddlers and Children ASD Screening methods

2.2.1.1 Quantitative Checklist for Autism in Toddlers (Q-CHAT)

The Q-CHAT is one of the oldest methods of screening for autism. It was developed by Baron-Cohen, Allen, & Gillberg (1992), as an efficient quantitative checklist to be administered by medical professionals coinciding with a report submitted by the child's parents based on observations of the child's behaviour. The earliest version of Q-CHAT was used to detect autism in toddlers aged between 18 and 24 months only. A screening study carried out to test the validity of Q-CHAT, based on 16,235 toddlers, revealed that the sensitivity of Q-CHAT's initial version was as low as 38%. M-CHAT, a modified version, was thus introduced by Robins, et al. (2001) to enhance the sensitivity of the original CHAT method. A similar screening study was conducted for M-CHAT, and it was discovered that it had higher sensitivity and specificity on the referred

sample population despite those of the M-CHAT method on the over-all population remained in question. However, both the CHAT and M-CHAT consisted of over 20 Likert Scale-type questions that needed to be completed in order to assist healthcare specialists in differentiating actual cases from controls for further referrals. The screening may take 20 minutes or more until the diagnostician would expect to complete it.

M-CHAT was later shortened to ten questions in order to make it more convenient and less time consuming for clinical and medical professionals to complete the screening process (Allison et al., 2012). The methodology and questions were revised, retaining only the most significant items during the screening and was based on a computed discrimination index. CHAT-23, the Chinese version of Q-CHAT, extended the screening population to toddlers aged from 16 to 30 months (Wong, et al., 2004). One of the most recent modifications, 10-Q-CHAT, is now available in many different languages, making it more accessible to healthcare professionals globally. It has proven to be an acceptable version, with 91% sensitivity and 89% specificity. Results of the CHAT method are generated based on a five-point scale (0-4), where a higher score indicates an increased possibility of positive autism symptoms (Allison, et al., 2008). The short version of the Q-CHAT screening contains 10 questions and are efficient to be conducted as it is accessible on the Internet.

2.2.1.2 Autism Behaviour Checklist (ABC)

Autism Behaviour Checklist (ABC) is a commonly used ASD screening tool with a rating scale developed by Krug et al., (1980) as an attempt to identify children with autism. ABC provides parents and teachers of children aged between 12 and 14 years a questionnaire to complete, consisting of 57 items, to capture the symptoms associated with the child's behaviour. These are categorised under five different subscales: sensory behaviour, communication and language skills, behaviour related to object use, body language, and social and adoptive behaviours. ABC is a subtest of ASIEP's third version and mostly consists of questions with scores ranging from 1-4 according to the level of mental or behavioural impairment of the child and based on the answers given by the parents, teachers, or the caregivers (Krug, et al., 2008). The result is then analysed by well trained professionals. In the initial stage, measures are utilised to distinguish the child's situation from autism and other types of neurodevelopmental disorders such as obsessive-compulsive disorders. Once the child's autism is confirmed, further evaluations are then made to understand the areas for development required by the child (Oro et al., 2014). According to the

research by Campbell et al., (2006), which used 167 sample cases of autism, it was revealed that ABC has 77% sensitivity, 91% specificity, and 80% overall accuracy rates.

2.2.1.3 Autism Screening Instrument for Educational Planning – 3rd Version (ASIEP-3)

ASIEP-3 is a special tool kit, consisting of an Autism Behaviour Checklist (ABC), a sample of vocal behaviour assessment, interaction assessment, educational assessment, and prognosis learning rate developed by Krug, et al., (2008), using normed data gathered from a sample population in the United States. Designed to identify children with a high level of autistic behaviour, aged from 0-2 years and 11 months to 13 years, it also develops programmes to educate such children and to monitor their progress over time. ABC is the initial screening tool used to identify autism cases and to sample vocal behaviour by measuring communication and language problems of the child (such as babbling and repetitiveness). Interaction assessment then monitors the child's spontaneous social behaviour while educational assessment measures the functioning level of the child. The prognosis learning rate helps the user to understand the child's progress and speed of learning. ASIEP-3 collects the normed data by using ABC questionnaires distributed among parents, teachers, caregivers, school psychologists, and other healthcare professionals in addition to providing activities for them do with the children that identify their special needs (Frye & Walker, 2008). ASIEP-3 is more time consuming than the other methods of screening for autism, taking approximately 90-120 minutes to complete a single screening process whereas methods like QCHAT-10 take approximately less than ten minutes (Soleimani, et al., 2014).

2.2.1.4 Childhood Autism Rating Scale (CARS-2)

CARS was developed by Schopler et al., (1980) to recognise young children with symptoms of ASD and to distinguish their severity through direct observation of the child's functioning in responses that include smell, touch, light, sound, body use, social behaviour, verbal and nonverbal communication, and consistency of intellectual response. The initial version of CARS was focused only on children aged below 6 years, and later in the CARS- 2nd version was extended to children aged from 6 to 13 years (Schopler & Bourgondien, 2010). The revised version is more responsive to high functioning individuals, leading to a higher clinical value in CARS-2 than in the CARS original version (Schopler et al., 1980). CARS- 2 consists of three basic components: the standard version rating booklet (CARS2-ST), the high functioning version rating booklet (CARS2-HF), and a questionnaire for parents and caregivers (CARS2-QPC). The standard version rating booklet is

similar to the original version of CARS and evaluates children aged below six years with an IQ lower than the estimated average IQ level in order to identify their communication impairments and risk of having ASD. The high functioning version of the rating booklet involves evaluating children aged between 6 and 13 years with strong communication skills and a higher IQ than the estimated average IQ level.

The questionnaire given to parents and caregivers to complete is used as a source for gathering information required for scaling CARS2-ST and CARS2-HF. Therefore, parents, teachers, and caregivers are granted access only to CARS2-QPS. Both of the other components consist of 15 items each, and are administered by specially trained professionals, evaluators, healthcare professionals, language therapists, physicians, and other experts who are familiar with ASD screening and evaluation. CARS-HF allows the scaling of the behaviours by using a 4-point scale based on direct observations by healthcare professionals along with those made by parents, caregivers, or any other party who is familiar with the child's behaviour. The final score is determined by the trained professionals and the severity of autistic behaviour is interpreted by using a cut-off score. Results generated by CARS2-HF are then used by psychiatrists, pathologists, and other healthcare professionals in their diagnosis and for determining what further evaluations are required in order to identify the type of medication that needs to be prescribed. School teachers need to plan and design education programmes according to the child's needs, and other educational professionals need to make accurate placement decisions, so they too utilise the results of CARS2-HF. Results are also used by other parties who are interested in furthering research on ASD and its screening methods. Validity studies against the diagnosis made based upon CARS-2 results showed 81% sensitivity and 87% specificity rates along with a good fit with the DSM-5 (American Psychiatric Association, 2013) and has been recognised as one of the best measures of autistic behaviour severity (Chlebowski, et al., 2010).

2.2.1.5 Developmental Behaviour Checklist – Early Screen (DBD-ES)

Einfeld & Tonge (2007) developed a questionnaire-based ASD screening method that needs to be administered by parents, teachers, caregivers, or other guardians of the child. The questions and assessment criteria have been based on research, experiments, and problems reported over a six-month period. There are a few versions of DBD available, each consisting of 96 items. The questions are designed to be answered with three basic options, ranging from 0-2, where 0 indicates

'Not True as far as I know,' 1 indicates 'Sometimes or Somewhat True,' and 2 indicates 'Very True.' DBD-P is designed for children aged between 4 and 18 years to identify their behavioural and emotional difficulties through a questionnaire administered by the parents or primary caregiver(s) (Gary & Tonge, 2005). DBD-T is also designed for the same age category of children, but for administration by their teachers. DBD-ES is the version developed for infants, aged from 18 to 48 months, and consists of 17 instruments that need to be answered by the parents or their caregivers (Gary, et al., 2007). The results of DBD-ES are interpreted based on an un-weighted cut-off score of 11. DBD-ES is an acceptable parent-administered ASD screening method for identifying children with intellectual frailties, and has a sensitivity of 83% and specificity of 48% (Einfeld & Tonge, 2007).

2.2.1.6 Early Screening for Autistic Traits (ESAT)

ESAT's design is based on the symptoms of ASD and seeks to differentiate infants aged 0-36 months from children with other types of developmental problems. Even though it is used for screening the behaviour of toddlers, it is focused more on detecting neurodevelopmental issues in infants at the age of 14-15 months. Dietz et al., (2006) developed ESAT in 2006 as a primary ASD screening checklist. It has 14 different instruments that are focused on a child's behaviours during playing, eye contact, joint attention, reactions, verbal and nonverbal communication, and interest in others. The checklist is undertaken by the parents or primary caregivers of the child and administered by a trained professional health consultant. All the questions given on the checklist have either 'Yes' or 'No' response options and the examination takes approximately 10-15 minutes to complete. Even though babies who score higher in ESAT are identified as more vulnerable to autism and other related developmental impairments, it doesn't differentiate well between cases and controls for infants aged below 25 months. Some infants who scored less in ESAT, and recognised as controls, later received an ASD diagnosis (Lord & Luyster, 2006). Therefore, the sensitivity and specificity of ESAT is still in question and follow up studies are required to develop ESAT into a more reliable ASD screening tool.

2.2.2 Hybrid Screening methods

2.2.2.1 Autism Spectrum Screening Questionnaire (ASSQ)

ASSQ is also a hybrid screening questionnaire developed by Ehlers et al., (1999A; 1999B) to identify the characteristics of high functioning developmental disorders, including autism and Asperger syndrome. It has 27 different questions, with three possible rating responses ranging from 0-2 as 'Yes,' 'Somewhat,' and 'No.' The questionnaire is administrated by the parents or the teachers of the 7 to 16-year-old child or adolescent, and the result is interpreted based on the overall score gained by the individual. An overall score exceeding the threshold indicates that the individual has many characteristics of the autism spectrum and other high functioning syndromes. An accuracy of 90% is shown in the accurately attributed positive results generated by the tool at a cut-off threshold of 13 (Einfeld & Tonge, 2007). A recent study performed to validate the ASSQ screening tool reported that it has 91% sensitivity and 86% specificity (Posserud et al., 2009).

2.2.2.2 Autism Spectrum Quotient (AQ)

AQ is a self-administrated ASD screening tool developed by Dr. Baron-Cohen (2001) along with other behavioural scientists from the Autism Research Centre, University of Cambridge, for identifying autism and other neurodevelopmental symptoms in adults with an average level of intelligence. The AQ questionnaire consists of 50 different questions covering the areas of social skills, attention switching, imagination, communication and attention to detail. The AQ test is available online and each question has four possible rating responses ('Definitely Agree,' 'Slightly Agree,' 'Slightly Disagree,' and 'Definitely Disagree') depending on which final score is calculated. The final score can range from 0-50 and a higher score indicates an increased level of autistic symptoms. A recent study on the validity of the AQ suggested that a cut-off score of 32 would optimise the validity of screening for adults in a clinical setting (Baron-Cohen, 2001). Later, in 2006 and 2008, two different versions of AQ were launched to cover adolescents and children (Baron-Cohen, et al., 2006; Auyeung, et al., 2008; Allison, et al., 2012). AQ-Child is a parent-administered questionnaire specially designed for children aged 4-11 years whereas AQ adolescent is designed for teenagers aged 12-15 years. All versions of AQ contain 50 unique items and take approximately 20-30 minutes to complete.

To make it simpler and less time-consuming, Allison, et al. (2012), presented a compressed version of the original AQ adult version known as AQ-10-adult. Even though AQ-10 is shorter

than the original version it has a predictive power similar to the original AQ version. The questions of QA-10 also have four possible responses: 'Definitely Agree,' 'Slightly Agree,' 'Slightly Disagree,' and 'Definitely Disagree.' The screening rule often considers one point per question. That is to say a point is assigned if the answer is either 'Slightly Agree' or 'Definitely Agree' for questions 1, 7, 8, and 10. In addition, a point is added if the user's responses to questions 2, 3, 4, 5, 6, and 9 are either 'Slightly' or 'Definitely Disagree.' The overall score is then calculated using a handcrafted diagnosis rule and anyone who scores above the threshold of six is considered to have autism and other related impairments. Lastly, Baron-Cohen, et al. (2006), and Auyeung, et al. (2008), have developed full AQ versions for adolescents and children respectively. Allison, et al. (2012), then proposed shorter versions for the full adolescent and children's AQ tests. Score calculations for the adolescent and child short versions are different from the AQ adult short version. Details of the questions, and the scoring of the adolescent and children AQ short versions, can be found in Chapter two and Allison, et al. (2012).

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

2.2.2.3 Social Communication Questionnaire (SCQ)

The Social Communication Questionnaire is a clinically used ASD screening tool to evaluate communication and social behavioural impairments of children that may be symptomatic of ASD. The questionnaire consists of 40 Yes/No questions that need to be administered by parents and primary caregivers of the child, and takes approximately 15 minutes to complete one screening process (Rutter et al., 2003). The SCQ is available in two basic formats: lifetime and current. The lifetime version looks at the entire developmental history of the child whereas the current version focuses on the child's behaviour only during the past three months. Results generated by both versions are subject to a cut-off score, where individuals who score beyond the cut-off score in the lifetime version are identified as possible autism cases and directed to complete further evaluations. Results generated from the current version are used to set goals for individual education programmes, further medical intervention, and treatment planning. According to the analysis carried out on the ASQ tool to evaluate its validity and degree to which the tool can discern

individuals with autistic behaviours from other neurodevelopment disorders, it was suggested that SCQ has a sensitivity of 96% and specificity 80% at a cut-off score of 15. It also revealed that a higher cut-off is required to differentiate autism from other developmental disorders. Therefore, at the optimum cut-off score of 22, SCQ achieves a sensitivity of 75% and specificity of 60% (Rutter et al., 2003).

2.2.2.4 Social Responsiveness Scale (SRS)

SRS is a family reported screening tool developed by Constantino (2005), an Associate Professor in Psychiatry and Paediatrics at the University of Washington, which also aims to differentiate cases of ASD from other neurodevelopmental disorders. The SRS questionnaire consists of 65 items, with three possible Likert scale responses ranging from 0 to 3 ('Not True,' 'Sometimes True,' 'Often True,' 'Almost Always True') and takes approximately 15-20 minutes to complete (Constantino, 2005). The overall scale is generated based on the responses given by the individuals, covering a large population of children and adolescents aged between 4 and 18 years. The SRS questionnaire can be administered by parents, relatives, spouses, or any other party who is familiar with the individual and is an acceptably valid tool for addressing the social responsiveness and related behavioural impairments of a wide range of individuals.

SRS-2, an updated version of SRS, was launched in 2012 to address other areas of development such as the communication, interpersonal, and stereotype behavioural condition of ASD (Constantino, 2012). SRS-2 allows individuals to use more than one examiner who is familiar with the rated individual and poses an average reading ability. SRS-2 has four basic forms of instruments covering three age groups, each consisting of 65 different items. The pre-school form covers children aged 2½ to 4½ years and the school form covers children and teenagers aged between 4 and 16 years of age. The original SRS is used as the school form with no changes and can be administered by the parents or teachers of the child. There are two versions of the adult form. One allows self-administration while the other allows administration by parents, friends, spouses, or any other relative of the individual. Both adult versions cover a large population from age 19 to 89 years.

Scoring of SRS-2 involves five subscales, including social awareness, social motivation, social cognition, restricted interests, and social communication. The social awareness subscale consists of eight items and evaluates the individual's ability to understand social clues. Repetitive

behaviour and restricted interest subscales have 12 items that address the stereotype and constrained behaviours of the individual. Social motivation measures the social interaction of the individual with 11 items, whereas social cognition uses 12 items to understand and interpret the social behaviour. The social communication subscale consists of 22 items and measures the ability for mutual communication. End results are interpreted using the T-score, derived from the overall score of all 65 items. A T-score above 76 indicates the possibility of severe ASD diagnosis, whereas a T-score ranging from 66 to 75 and 60 to 65 indicates moderate and mild conditions of ASD respectively. Further, individuals with a T-score below 59 are considered controls, so do not necessitate ASD clinical diagnosis (Constantino & Gruber, 2014).

However, SRS's validity and reliability are clinically proven, and the studies have shown that it has a sensitivity of 78% and specificity of 94% at a cut-off score of 70 (Wilkinson, 2015).

2.2.2.5 Child Behaviour Checklist (CBCL)

The Child Behaviour Checklist is one of the oldest screening tools, developed in 1991 as a component of the Achenbach System of Empirically Based Assessments (ASEBA) established by Tomas Achenbach with the aim of identifying behavioural disorders of children and adolescents aged between 6 and 18 years (Achenbach, 1991). There are two basic versions of CBCL: a preschool version and a school age version. The preschool version covers children aged from one and half years to five years and is administered by the parents or primary caregivers who interact with the child on a daily basis. The questionnaire consists of 100 questions with three possible Likert scale responses ranging from 0-2, where 0 indicates 'Not True' and 2 indicates 'Very True.' The school age version covers children and teenagers aged between 6 and 18 years, and the questionnaire consists of 118 items with rating responses similar to the preschool version. Eight important areas of the child's development are screened through the questionnaires: attention problems, aggressive behaviour, anxiety levels, rule breaking behaviour, social problems, somatic complaints, depression levels, and thought problems. Each represents one subscale on the checklist. Two different scores are derived from the eight subscales: internalising problems and externalising problems. The summation of both internalising and externalising of problems presents the total problem score that is used to interpret the child's behaviour. A higher total problem score indicates the higher risk of behavioural impairment. According to an evaluating

study carried out on the validity of CBCL using a Brazilian sample population it was revealed that the CBSL Brazilian version has a high sensitivity and specificity in identifying cases of behavioural disorders (Bordin, et al., 2013).

2.3 Comparison of the ASD Screening Methods

There is an increasing trend in studies related to autism symptoms and screening over the last three decades. Identified above are the common ASD screening tools that require the involvement of many stakeholders, such as parents, children, caregivers, patients, clinicians, psychologists, and other healthcare professionals, in taking the screening tests, calculating scores, explaining and interpreting tests results, and sometimes in processing further referrals. This section presents a comparison of identified tools in terms of their target audience, screening method, number of items, time consumption, and performance. The best performing tools are identified and requirements for further developments are discussed. Later, in Section 3, more in depth discussion will be conducted with respect to different primary criteria.

The study evaluated different screening methods, some with multiple versions. One of the oldest tool, CBCL was developed in 1991, and versions of AQ-10 (Adults, Child, and Adolescent) and Q-CHAT are recent screening tools having been developed in the year 2012. Among all the identified ASD screening tools, most focus on a specific range of the population, in particular infants and toddlers aged between 6 and 48 months. Fewer tools are available for adults aged 18 or more years. Many of the screening methods utilised one or more questionnaires to identify autistic behaviours. CBSL includes a questionnaire with 118 items, and thus has the maximum number of items, whereas the “Child,” “Adolescents,” and “Adult” short versions or AQ-10s and Q-CHAT-10 have the minimum number of items at only 10. Therefore, AQ-CHAT and AQ-10s are efficient methods since they take around 10 minutes to complete, whereas methods like CBCL usually require much longer time to complete. In terms of sensitivity and specificity and ASIEP-3 have the highest sensitivity at 100% whereas CHAT and DBD-ES have the lowest sensitivities around 40% and 48% respectively.

Table 2.1 displays a comparison between the ASD screening tools available for toddlers and children aged between 6 months and 13 years. CHAT is the oldest method, developed in 1992, whereas Q-CHAT-10 is the most recent. Most of the methods have questionnaires with 10 to 70 items related to ASD behaviour, communication, and social conditions. ASIEP-3 is one of the only

methods to use different activities to evaluate children's behaviour. Most of the screening methods take approximately 10-25 minutes to complete, except ABC which takes more than 25 minutes. CHAT and Q-CHAT-10 are more efficient than the other methods in terms of time required taking

Table 2.1: Summary of screening methods available for toddlers and children

METHOD	FEES	ITEMS	TARGET	AGE	TIME/ MINS	SENSETIVITY	SPECIFICITY	WEB	MOBILE	REFERNCE
CHAT (Checklist for Autism in Toddlers)	N	14	Toddlers	18-24 months	8 to 15	40%	98%	X	X	(Cohen et al., 1992)
M-CHAT (Modified Checklist for Autism in Toddlers)		23	Toddlers	16-30 months	10 to 20	95-97%	99%	X	X	(Robins et al., 1999)
M-CHAT-R (Modified Checklist for Autism in Toddlers, Revised)		20	Toddlers	16-30 months	10 to 20	NA	NA	X	Y	(Robins, Fein, & Barton, 2009)
Q-CHAT (Quantitative CHecklist for Autism in Toddler)	N	25	Toddlers	18 to 24 months	15 to 20	88%	91%	Y	X	(Allison et al., 2012)
Q-CHAT-10	N	10	Toddlers	19 to 24 months	5 to 10	91%	89%	Y	X	(Allison et al., 2008)
CHAT-23		23	Toddlers	16 to 30 months	10 to 20	84%	85%	X	X	(Wong et al., 2004)
ABC (Autism Behavior Checklist)	Y	57	Children	3 to 14 years	20 to 30	77%	91%	X	X	(Krug et al., 2008)
(CARS)-2 Childhood Autism Rating Scale	Y	15 X 2	Children	V1:<6 years and V2: 6 to 13 years	10 to 20	81%	87%	Y	X	(Schopler et al., 2009)
DBD-ES (Developmental Checklist-Early Screen)	Y	17	Toddlers	18 to 48 months	10 to 15	83%	48%	X	X	(Gray & Tonge, 2005)
ESAT (Early Screening for Autistic Traits)		14	Toddlers	16 to 30 months	10 to 15	88%	14%	X	X	Dietz et al. (2006)
ASIEP-3 (autism screening Instrument for Educational Planning - Third Edition)		47	Toddlers & children	2 to 13 years	Varies	100%	81%	X	X	(Krug et al., 2008)

the test. ABC, CARS-2, and ASEIP-3 are the most comprehensive methods available for screening individuals for autistic behaviours, as they cover a wider audience than the other available methods. Most of the ASD screening methods are not freely available on the internet and have limited accessibility. Q-CHAT, CARS, and CSBC-DP are comparatively more accessible than the other methods and are freely available on the internet for anyone to download. Limited ASD screening smart phone applications are available for infants, toddlers, and children. In terms of validity, almost all the screening methods have acceptable sensitivity rates, ranging from 70%-100%, and specificity between 80% and 100%. CHAT's initial version, and DBD-ES are the only ASD screening methods to have low sensitivities and specificities.

Table 2.2: Summary of hybrid screening methods available

METHOD	FEES	ITEMS	TARGET	AGE	TIME/ MINS	SENSITIVITY	SPECIFICITY	WEB	MOBILE	REFERENC
ASSQ (Autism Spectrum Screening Questionnaire)		27	Children & Adoloscent	7 to 16 years	10 to 15	91%	86%	Y	X	(Ehlers et al., 1999)
AQ(Autism Spectrum Quotient)	N	50	Adult	>18 years	20 to 30	93%	52%	Y	Y	(Baron-Cohen, et al., 2001)
AQ-10-Adult	N	10	Adult	>18 years	5 to 10	77%	74%	Y	Y	(Allison et al., 2012)
AQ-Adolescent	N	50	Adolescent	12 to 15 years	20 to 30	NA	NA	Y	Y	(Baron-Cohen, et al., 2001)
AQ-Child	N	50	Children	4 to 9 years	20 to 30	95%	95%	Y	X	(Auyeung, et al., 2008)
AQ-10-Adolescent	N	10	Adolescent	12 to 15 years	5 to 10	NA	NA	Y	X	(Allison et al., 2012)
AQ-10-Child	N	10	Children	4 to 11 years	5 to 10	NA	NA	Y	X	(Allison et al., 2012)
SCQ (Social Communication Questionnaire)		40	Children & Adoloscent	<4	10 to 20	58-62%	93-100%	X	X	(Rutter, et al., 2003)
SRS (Social Responsiveness Scale)		65	Children & Adoloscent	4 to 18 years	20 to 30	67%	78%	X	X	(Constantino & Gruver, 2005)
SRS-2 (Social Responsiveness Scale)		65	Children & Adoloscent	4 to 18 years	20 to 30	78%	94%	X	X	(Constantino & Gruver, 2012)
CBCL (Child 28 Behavior Checklist)		118	Children & Adoloscent	6 to 18 years	25 to 40	75%	82%	X	Y	(Achenbach 1991)

Table 2.2 shows a comparison between screening methods in the hybrid category. These screening tools address developmental issues in infants, children, adolescents, and adults simultaneously by either employing the same screening tool for all, a two or more combination of the above age categories, or through employing different versions customised for each category of audience. The AQ utilises different versions (short and full) to evaluate the behaviours of children, adolescents, or adults.

Almost all of the hybrid ASD screening tools utilise questionnaires as the method of evaluation regardless of their target audience. Tools that utilise questionnaires often consist of 10-70 items, with the exception of CBCL which includes 118 items. Tools such as the AQ short versions are more efficient than the others, with only 10 items to answer and requiring far less time to complete (5-15 minutes). Further, AQ is more comprehensive and accessible than the other tools as it has several versions that cover many age segments and is freely available on the internet as both web-based questionnaires and smart phone apps. AQ does not require professionals or specially trained individuals to administer the questionnaires, further increasing its accessibility. All the screening tools listed under the hybrid category are acceptable in terms of their sensitivity and specificity.

2.4 Discussion on the ASD Screening Methods

The section below is focused on evaluating the screening tools presented above in terms of their administration methods, accessibility, popularity, performance, and comprehensibility in order to shed light on the possible innovations required in new screening tools to be developed for screening autism.

2.4.1 DSM-IV vs. DSM-5 Criteria

The assessments of autism should be directly related to DSM-5 criteria. The validity and reliability of most of the screening tools is still under investigation as most of them follow the earlier version of the DSM (DSM-IV) (American Psychiatric Association, 2000), rather than the procedures and guidelines of the current DSM-5 manual. Since most of the screening methods utilise different behavioural characteristics in determining a patient's developmental age, they have been jointly presented as the triad of impairments under the definition of the DSM-IV and still need to consider

amendments presented in the DSM-5. The most recent version of the DSM (DSM-5) groups the five PDDs, consisting of Asperger syndrome (AS), Pervasive Development Disorder–not otherwise specified (PDD-NOS), Rett syndrome (RS), and Childhood Disintegrative Disorder (CDD), into ASD (American Psychiatric Association, 2013). Guidelines set by the DSM-IV are followed mostly by clinicians and healthcare professionals all around the world when diagnosing autistic behaviours. In the USA, the 10th version of the International Classification of Disease (ICD-10) is also used in diagnosis and clinical evaluations of autism and other developmental disorders (World Health Organisation, 1992). ICD-10 lists seven syndromes under PDD, and includes atypical autism and unspecified PPDs beyond the five PDDs listed by DSM-5.

Conventional methods used in clinical judgements (CJ) of ASD, such as ADI-R and ADOS, diagnose individuals based purely on behavioural criteria through a questionnaire or interview that contains items related to the DSM-IV (Lord, et al., 1994; Lord, et al., 2000). After publication of the DSM-5, researchers pointed out that some cases who were diagnosed with "Autism" using DSM-IV criteria may not be classified as having ASD under the revised DMS-5 criteria (Thabtah, 2017; Kent, et al., 2013; Grzadzinski, et al., 2013; Mazefsky et al., 2013; Matson, et al., 2012). This has created a debate among scholars in behavioural science, psychiatry, and psychology due to the inconsistent sensitivity and specificity results published in the last few years. For instance, Matson, et al (2012), showed a reduction in sensitivity for adults and toddlers while Mazefsky, et al. (2013), revealed a consistent sensitivity of cases tested under both the DSM-IV and DSM-5 despite a decrement in specificity.

Since most of the ASD screening methods available today were developed prior to 2013, they did not consider the guidelines established in the DSM-5. Existing ASD screening methods are based on clinical diagnosis methods, and therefore changes in ASD diagnosis criteria after publication of the DSM-5 demanded a change in the way diagnostic algorithms within the screening method behaved during the classifying of cases. For instance, items in the existing screening methods should cover social interaction and social communication (Category A) in the DSM-5 manual and at least two criteria from Category B (Restricted and Reparative Behaviour). Unfortunately, despite items in the majority of current screening methods fulfilling multiple criteria in Category A, they still fail to fully cover conditions in Category B. Nevertheless, screening for ASD does not necessarily require fully meeting the diagnostic conditions of ASD as

its ultimate aim is merely to reveal potential autistic traits rather than diagnose individuals since to do so necessitates the involvement of expert clinicians and a clinical setup.

Therefore, there is a need to re-examine questions and features within the ASD diagnostic and screening tools in order to comprehensively satisfy the new criteria of the DSM-5. This necessitates mapping the new ASD criteria to the items used in the screening tool besides evaluating the way the diagnostic process works. The outcome may result in an updated version of the current screening tool that maps the new criteria of ASD in the DSM-5 to the items of the screening tool. In addition, comprehensive experimental studies using controls and cases as data are expected to be conducted in order to direct researchers, clinicians, psychiatrists, and psychologists to the right screening tool that maintains performance even after the new changes proposed.

2.4.2 Digital Presence and Accessibility

Some of the discussed ASD screening tools are available on the internet, either as web-based online tests or smartphone applications. Other instruments require a payment and are available only in hand written formats. These tools are intended to enhance development of disease control and prevention measures through early detection of ASD and associated communicational and behavioural disorders. ASD screening tools, such as Q-CHAT, ASSQ, CARS and AQ are freely available on their web pages for access by parents, teachers, professionals, and clinicians who can administer the tests online and then receive an automatically generated score at the end of each test completed. Most web-based screening tools provide a guide to interpreting the final scores.

There are only a few screening tools available as mobile apps, and most of these use two or more combinations of the screening methods above to derive their results. Therefore, it is quite difficult to evaluate the methods used in smart phone apps in terms of sensitivity and specificity. M-CHAT, AQ, and CBSL are the screening methods most commonly used in mobile platforms (Android, iTunes). However, a new mobile screening app based on all AQ short versions (AQ-10-Adult, AQ-10-Adolescent, AQ-Child) and Q-CHAT (toddlers) was recently developed to cover all age categories (Sappok et al., 2015). This is the only screening app available for all audiences. Autism Fingerprint is another example of a smart phone app that uses M-CHAT as its screening method. In Oman, 14 out of every 100,000 children have been found to display autistic traits, but

there has been a lack of awareness and properly standardised tools to diagnosis the disease in the early stages. For this reason, Autism Fingerprint was developed by Arab neuroscience specialists in collaboration with technical experts (Kleina, et al., 2015). Culturally and traditionally appropriate images and items were used in order to make the application more user-friendly and an easy to use tool for screening children for autistic traits. AQ Asperger Test, AQ Test, and the Asperger Test are some of the mobile apps that were developed using AQ. These applications can be found in both Android and iTunes mobile platforms. Canvas Child Behaviour Check List is such a mobile app, using CBCL as the screening method to investigate behavioural impairments in children and adolescents aged between 4 and18 years.

There are many other applications that use games, drawing tools, and advanced online activities to observe the behavioural conditions of individuals, covering various segments of the population. Apple iPad is considered one such advanced tool that helps children with special needs in their communication and social development programmes. Specialists in the world's health care arena have advised that since the iPad came onto the market in 2010, it has helped many children with autistic behaviours in developing their skills (Chen, et al., 2017). Similarly, the contribution of numerous and innovative applications introduced by Android for promoting awareness and identifying individuals with autism at an early stage has been immense.

Presence on the internet and mobile platforms is essential in today's society to ensure the accessibility of any product or service. Therefore, the availability of free ASD screening tools via the internet defines their accessibility. Accessibility and comprehensiveness of some screening tools is questionable, however, as most are not freely available and only target specific demographic groups. Even though some of the ASD screening tools are available on web platforms, they are not free for users. Most of the tools available on the internet are subject to a certain payment prior to obtaining access to the screening process. ASD screening instruments, such as the AQ versions and Q-CHAT, are freely available for anyone to use while tools such as CARS-2 are available on the internet as a pay only facility. In a world where the demand for smart phone applications is growing rapidly and even the most basic facilities are available in mobile application format, it is crucial for ASD screening tools to be present on mobile phone application platforms. Currently, only a few screening tools, such as AQ are available on smart phone platforms. Unavailability on the internet and mobile platforms hinders the accessibility to users of many of the screening tools.

2.4.3 Administration and Time Efficiency

Administration refers to undertaking the questionnaires or interviews provided by the ASD screening tools in order to identify autistic traits. There are three types of ASD screening tools: self-administered, parent or caregiver administered, and administered by clinicians or well-trained professionals. Most of the screening methods discussed fall into either the self-administered or parents/caregiver-administered category. Some methods require professionals to administer the questionnaires and/or to score and interpret the generated results. This is one additional requirement that makes the screening tools hard for ordinary people to use. Q-CHAT is an example of screening instruments that requires administration by professionals in addition to a report submitted by the parents on the behavioural complexities of their child. On the other hand, screening tools that are self-administrated, such as AQ and its versions, seem to have fewer requirements to be conducted and can be taken by adults with average IQ, parents, family members, caregivers, and teachers among others. These self-administered methods often utilise simple scoring functions that offer a numeric score for the likelihood of having autistic traits.

One of the key performance indicators of the ASD screening tools is the time taken to complete one screening process (Bone et al., 2016, Duda et al., 2016, Allison et al., 2012). Since conventional methods are usually lengthy questionnaires that take time to complete, and many of the screening tools have originated from these questionnaires, it is advantageous to reduce the time necessary for the test. For example, the AQ-adolescent version originally had a questionnaire with 50 items that took approximately 15-20 minutes to complete. Allison, et al. (2012) dealt with this shortcoming by reducing the number of items in the original AQ questionnaire to 10 items, only taking 5-8 minutes to complete. Many scholars have not yet recognised this need for a short and effective screening tool, and therefore have less involvement by users.

In the current digital era, most users prefer to have shorter screening tests such as Q-CHAT-10 and AQ-10 versions since the tests are typically taken in an online environment and within an acceptable timeframe. In fact, recent developments in hardware, computer networks, and mobile applications have provided rapid accessibility to the tools for the healthcare community. New technologies, such as mobile platforms, may render some of these time-consuming tools obsolete.

2.4.4 Performance and Comprehensibility

The validity and reliability of the ASD screening tools are expressed in terms of sensitivity and specificity metrics when applied against a certain dataset of cases and controls. It is imperative, therefore, to acknowledge that these two metrics (besides accuracy) are measured with respect to a specific dataset. Thus, the screening method performance is restricted to the dataset characteristics and ensures quality. According to Greenhalgh (1997), sensitivity refers to the ability of the screening tool to identify a person with autism while specificity refers to the power of the screening tool to discriminate a person who is a control of autism. Based on the results reported in the literature (and included in the tables constructed in Section 2.4), most of the existing screening methods have acceptable sensitivity and specificity rates. Nevertheless, some screening tools have little research validating their results with respect to sensitivity and specificity metrics. The ESAT method which has the least reported specificity (14%) is an example of screening tools that could potentially be improved to obtain acceptable levels of sensitivity and specificity on their datasets. Tools such as M-CHAT and FYI still necessitate experimental studies to seek their actual performance (both sensitivity and specificity).

Comprehensibility of the screening tools depends on the size of the audience that they cover. Most of the screening tools cover only one segment of the population, with some being specialised for infants and children while others are designed specifically for adults. Since the recognition of autism at an early stage is critical for medication and treatment planning, many tools cover infants and toddlers aged between 12 and 36 months. A lack of importance placed on teenagers and adults is another issue associated with existing ASD screening instruments. Even though the instruments, such as M-CHAT and Q-CHAT, are present on both web-based and mobile phone platforms, with acceptable levels of sensitivity and specificity, the comprehensibility of these tools is in question as they only cover infants aged from 16-30 months. This represents less than 5% of the entire population.

2.4.5 Popularity

There is no exact metric for measuring the popularity of an ASD screening tool as no tracking is available to measure the number of individuals who use a screening tool at any given time, nor how frequently they are being used. An estimate on popularity can be derived from the clinical

usage of each tool. Unfortunately, most available ASD screening tools are designed for research and developmental purposes rather than clinical diagnosis purposes, with only a few tools such as CARS-2HF, ADOS, and ADI-R being used by clinicians in their medical diagnosis process. None of these screening tools can be used alone to provide a proper medical diagnosis, and are used in collaboration with many other medical tests and professional investigations in order to reveal autistic behaviours and to differentiate them from related developmental impairments.

Some of the ways to measure the popularity of a screening tool is to utilise app features like functionalities, user review ratings, and coverage. Out of the testing methods considered, AQ short versions and Q-CHAT have been able to obtain positive ratings across approximately 100+ user reviews. For example, the ASDTests app, which is based on the AQ short versions, has 111 reviews and numerous downloads. Apparently screening tools that are based around questionnaires are more favourable to end users, as observation and video screening methods have limited or no ratings in both the Android and Apple stores. It is believed that this is the result of these screenings being more time consuming than questionnaire-based methods. In questionnaire tests, such as AQ-10, the number of questions is just 10 so users are less likely to lose interest while using this method. Hybrid screening methods seem to cover a larger group of users, as they target various combinations of toddlers, children, adolescents, and adults. An example of hybrid screening methods is ASDTests; it covers toddlers, children, adolescents, and adults collectively, thus making it more popular. This shows that hybrid screening methods seem to be more comprehensive than specific screening methods, at least in the context of user usage. Nevertheless, methods such as Q-CHAT, that cover toddlers, are still popular within their user segment (toddlers).

2.5 Recent Machine Learning Study on ASD Screening and Diagnosis

A well-known clinical diagnosis method in ASD research that has been widely used is ADOS-R (Lord et al., 2000). This method relies on four different modules embedded within a computerised tool which evaluates the individual's language communication ability. There are four main modules developed in ADOS-R for children and adults in which each is applicable to a certain demographic based on behavioural and language levels, and can range from verbally fluent to non-verbally fluent. Usually the examiner selects the right module for each case under examination

based on two factors: chronological age and language proficiency (Lord et al., 2000). For instance, module 1 is designed for cases that do not regularly utilise phrase speech such as young children. Wall et al. (2012A) and Wall et al. (2012B) claimed that machine learning methods such as decision trees can be employed to construct a model that contains a less number of features than items found using ADOS-R (Module 1). Therefore, the time associated with the medical diagnosis is shortened without negatively influencing sensitivity, specificity, and validity of the test. The authors sought to identify the least number of items in ADOS-R to classify ASD cases via constructing decision tree classifiers in WEKA (Hall et al., 2009) by using information gain filtering. In particular, they have applied a number of machine learning methods (decision tree based) on an ASD dataset aiming to identify the best classifier.

The dataset used was downloaded from the Autism Genetic Resource Exchange (AGRE) repository (Geschwind, 2001). It consisted of 612 individuals with autism and 11 cases along the non-autism spectrum. The authors only maintained two class labels, autism and non-autism, and discarded all data examples that had over 50% missing values. To balance out the instances in the dataset, the authors have utilised a number of pre-processing methods. After applying a number of decision tree based algorithms on the ASD classification dataset, the results revealed that the best classifier in sensitivity, accuracy, and specificity contained rules involving only eight features. They concluded that ADOS-R can only apply eight features effectively, rather than the complete set of twenty-nine features in Module 1. Nevertheless, the results produced are algorithm specific since those eight features are only presented in the Alternating Decision Tree algorithm (ADTree) and only for the specific dataset used in the experiment. In other words, if we apply other machine learning algorithms such as associative classification, rule induction, or neural network, the number of features appearing in classifiers may definitely vary. A better approach toward achieving less numbers of features should involve investigating the significance of complete feature sets on classification performance using filtering and wrapping methods. This may derive smaller features sets that are generic and not algorithm or data sensitive. One clear shortcoming of the dataset(s) used is the fact that it is clearly unbalanced and a third class/category of ASD was discarded by the authors which may simplify the problem to either severe autism or no autism at all.

The process of clinical diagnosis of ASD can be lengthy since it may vary among examined cases alongside other obstacles associated with the diagnoses process in the health care system.

Allison, et al. (2012), investigated shortening the time linked with self-administrated ASD pre-diagnosis in medical family clinics. Their aim was to enable medical care staff, including physicians, nurses, and other clinical staff, to utilise at most ten features/questions as a form for quick clinical referrals of potential ASD cases. The authors then analysed different versions of current self-administered or parent assisted ASD screening tools, which included

- Quantitative Checklist for Autism Toddlers (Q-CHAT)
- Autism Spectrum Quotient (AQ) (3 versions)
    - Adult
    - Adolescent
    - Child

Samples of controls as well as ASD cases have been utilised to validate different ASD traits in the three evaluated screening methods. The authors have exploited web-based recruitment besides already collected data by their research group to measure the significance of each trait. The data of the controls as well as the cases have been split into training and validation sets respectively. The significance of a trait was computed using a discrimination index which corresponded to the rate of positive cases for a trait, i.e. T, in the training set from T rate, derived from the control training set. Different evaluation measures including specificity, sensitivity, and "Area under Curve" (AUC) of the predictive validity have also been adopted in the experiments. The top ten traits with discrimination scores have been selected, and the results of the selected traits showed competitive performance with respect to the abovementioned measures when compared with results obtained from the complete set of traits for each screening method. The authors concluded that these ten questions (traits) can only been used to refer to suspected cases of ASD for full clinical screening and cannot be relied on for a formal ASD diagnoses.

Few limitations are associated with Allison, et al. (2012) study. For example, most cases recruited during data collection were aware of ASD thus potentially creating biased results. More importantly, we believe that despite the promising results achieved by Allison, et al. (2012), using only a discriminative factor is not sufficient to measure the significance of a code or feature in a screening method. There should be deeper evaluation of each feature within a large collection of sample cases and controls. This will shift the problem to identifying smaller sets of features as

clusters. Each cluster contains features with a common relationship that may guide the diagnosis algorithm when building classification models using machine learning. Therefore, we need to draw termination points that split features into groups. These groups may overlap in features where a feature or code such as “x” can possibly belong to multiple clusters. This is since data cases of ASDs inside the original dataset overlap in traits, and the new ASD published criteria of the DSM-5 have similarities in sub-category items (codes) (A's-items, B's-items, etc.).

A pre-processing phase for splitting data objects into unambiguous and boundary objects for data collected from ASD diagnostic sheets has been proposed in Pancer and Derkacz (2015). The idea was to differentiate between data objects that may belong to more than a single class (boundary objects) and data objects which clearly belong to an obvious class or non-overlapping data objects (unambiguous), and then each set of data objects will be trained to derive a classifier. The authors adopted the concept of a decision table from rough set theory and computed the objects belonging to either boundary or unambiguous sets using a consistency factor. More details on the mathematical notations of the consistency factor can be found in Pancer and Derkacz (2015). Seventy hard copy ASD diagnosis sheets that contain 17 different sections and 300 items have been transformed into a soft copy data file. We believe that these sheets correspond to DISCO screening items. Each item then has been given four possible values (0-not performed, 25-performed after physical help, 50-performed after verbal help, 100-performed unaided). Finally, a consistency factor per data case has been computed to place the data case to the right data type (either boundary set or unambiguous set). No learning phase has been involved or automatic case classification, and therefore this article can be seen as a first step toward automatic classification using machine learning since it only handles pre-processing of data cases related to DISCO screening tool.

Duda, et al. (2016), applied six machine learning algorithms to distinguish ASD from ADHD cases on a real dataset consisting of 2925 cases (2775 ASD cases and 150 ADHD cases). The authors’ purpose of the study was to minimise the time of pre diagnosis for ADHD and ASD using electronic and digitized applications. The study adopted 65 features from Simplex Simon Collection (SSC) version 15 (Fischbach and Lord, 2010) based on the Social Responsiveness Scale (SRS) which is a parent administrated questionnaire that is often utilised to measure autism traits. In the experiments of machine learning algorithms, the authors pre-processed the data by removing cases and controls that have more than four missing answers in their sheets. Thus limiting the input

dataset with data samples with <5 missing answers. Furthermore, the authors employed forward feature selection to reduce data dimensionality to less than ten features and adopted cross validation during the training phase of the classification algorithms. Moreover, under sampling to balance the class labels was performed before building the classification model and the authors adjusted the data by under sampling to a ratio of 1.5:1 (ASD to ADHD). After experimentation, six features from the SRS data remained present after pre-processing. The majority of the considered machine learning algorithms, especially the functions based ones such as Logistic Regression, achieved high classification accuracy (mostly greater than 95%) whereas decision tree based algorithms such as Random Forest achieved an unacceptable accuracy. There was no clear mechanism on the way forward to discover the hard cases overlapping between ASD and ADHD.

Mythili and Shanavas (2014), investigated the problem of pre-diagnosis of ASD on a non-clinical case (fixed dataset) using a number of machine learning predictive approaches; particularly Neural Network, Support Vector Machine, and Fuzzy Logic. The authors have used a simple dataset (100 samples) consisting of three attributes (Language, Social, Behaviour) and a class attribute named Autism Level (Mild, Moderate, Sever). There is no information whether the data was ready or had been collected, and the data size is very limited. The authors utilised the WEKA software tool for testing the different classification approaches on the data without mentioning what classification algorithms have been employed and why. Very limited experiments were conducted along with a number of decision tree methods. Furthermore, no results discussion was provided on obvious correlations between the three attributes and the class label. The paper of Mythili and Shanavas (2014) has insufficient data and no clear methodology or analysis, and therefore its results cannot be generalised. However, we can consider using machine learning for self-diagnoses of ASD as a promising research direction.

A medical diagnosis of autism is considered crucial to validate and often this validation is done by a clinical expert with proper screening instrument. The formal diagnosis frequently takes hours and relies on:

1) The case complexity to be diagnosed

2) The clinical diagnoses procedure followed

3) The expertise of the clinical professional

One recent claim of speeding the autism diagnoses procedure of ADOS-R (Module 1) based on machine learning has been discussed earlier by Wall, et al. (2012a). Nevertheless, shortcomings related to result analyses have been highlighted with rationale in Bone et al. (2014). The authors have argued that the problem of classifying autism using machine learning is not straightforward and requires careful consideration of clinical procedural setup. Precisely the following pitfalls in Wall et al.'s article (2012a) have been identified by Bone et al. (2014).

1) Despite the claim of the reduction in number of features (codes) in the ADTree classifier constructed from the input cases to 8, all tasks and activities of the ADOS-R test must yet be performed and therefore no administration time reduction is observed. The full ADOS-R test must be conducted before building a classifier using ADTree algorithm in WEKA.

2) The validation of ADOS codes can only be established within a clinical environment. Yet the authors of Wall et al. (2012a) have claimed that ADOS codes can be self-administered, and this claim needs more substantial supporting evidence.

3) The data has been reduced by discarding important cases representing a heterogeneous category. Presence of this class will influence the resulting error rate of the machine learning algorithm.

Tenev, et al. (2014) investigated Support Vector Machine applicability in distinguishing ADHD sub types using a sample of 67 adults with ADHD and 50 controls based on the power spectra of EEG measurements. The participants were recruited and assessed for neurophysiological responses in a distraction free environment in which each individual tests lasted approximately 90 minutes and EEG was recorded with a Mitsar 19-channel QEEG system. Data collected has been divided based on each ADHD condition and then forward sampling has been employed before applying machine learning. On each data partition a Support Vector Machine has then been trained to derive classifiers. This machine learning approach was able to differentiate among ADHD conditions for at least the adult cases and controls used. Behavioural attributes from ASD diagnostic tools were not utilised in this study, rather using data collected from EEG measurements for power spectra data.

A number of machine learning algorithms have been applied by Pratap, et al. (2014) on a dataset based on the Childhood Autism Rating Scale (CARS) diagnostic tool (Schopler, et al., 1980). CARS is a behavioural rating scale usually employed for testing symptoms related to ASD. The

authors aim was to measure the effectiveness of the probabilistic Naïve Bayes technique (Duda, 1971) as well as Artificial Neural Network based on a Self Organizing Map (SOM-ANN) (Kohonen, 1989) using sixteen features dataset with 100 cases of children between two to three years old. The authors divided the target class into four possible values: Normal, Mild-Moderate, Moderate-Severe and Severe. After experimentations, the results indicated that when probabilistic model or ANN predictive models are integrated with unsupervised learning methods such as K-Means clustering (MacQueen, 1967) the results of detecting ASD cases improve at least on the dataset used. A later study, Pratap and Kanimozhiselvi (2014), by the authors on the same dataset showed that SOM predictive models with a single input and four outputs when preceded by unsupervised learning method can increase the accuracy of detecting children with ASD based on CARS tests. However, these two studies have utilised a very limited number of children besides the dataset employed, and have not been verified by other researchers or made available within known autism research centres.

Chu et al., (2016) investigated efficient ways to differentiate between ADHD and obstructive sleep apnea (OSA). The authors utilised 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 can assist clinicians and physicians in improving the diagnostic decision. Reported results indicate 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 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 models (Random Forest) to reduce the time of ASD diagnostic process. Maenner et al., (2016) investigated Random Forest algorithm (Breiman, 2001) on an autism dataset from Georgia Autism and Developmental Disabilities Monitoring (ADDM) Network utilising 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 contains 9,811 evaluations of 1,450 children. The results reported that Random Forest achieved around 89% predictive value and 84% sensitivity.

Thabtah, (2017) critically analysed 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.

Table 2.3: Sample of studies on the use of machine learning for ASD classification in behaviour science

Machine Learning Automated Methods													
Year	**Diagnostic Tool**	**CV**	**Feature selection**	**Machine Learning Mehtods used**	**Software**	**Data size and source**	**Balancing**	**features used to derive best results**	**best algorithm**	**Specificity%**	**Sensitivity%**	**Accuracy%**	**AUC**	**Publication Details**
2012	ADOS (Modul 1)	Yes	No	16 algorithms (SVM, LG, Tree, Probabilistic and their variations)	Weka	29 attributes, instances: 612 ASD, 15 non-ASD, sources: Boston AC, AGRE	yes-simulation	7 features	ADTree	94	100	99.7		(Wall DP, et al., 2012a)
2012	ADI-R	Yes	No	16 algorithms(SVM, LG, Tree, Probabilistic variations)	Weka	93 attributes, instances: 891 ASD, 75 non-ASD, sources: AGRE	yes-simulation	7 features	ADTree	99-93.8	100	100		(Wall DP, et al., 2012b)
2014	CARS	No	No	supervised (Naive Bayes, SOM, Neural Fuzzy, LVQ Nueral Network), Unsupervised(Kmeans, Fuzzy C Mean)	developed	16 attributes, instances: 100 sources: (Pratap, et al., 2012)	no	16 features	SOM and Naïve Bayes			100		(Pratap, et al., 2014)
2015	ADOS (Modules 2 and 3)	Yes	Yes	8 (SVM, LR, DT, Probabilistic variations)	R, Weka	28 attributes each module , instances: 3885 ASD, 655 non-ASD, sources: Boston AC, AGRE, NDAR, SSC, SVIP	stepwise forward	9 features for module 2 and 12 features for modeule 3	SVM and LR	89.39	98.81	98.27% (module 2), 97.66% (module 3)		(Kosmicki, et al., 2015)
2016	SRS	Yes	Yes	6 (SVM, LR, DT)	Scikit-learn	65 attributes, instances: 2775 ASD, 150 ADHD, sources: Boston AC, SSC	stepwise backward and undersampling	5 features	SVM				96.5	(Duda et al., 2016)
2016	ADI-R and SRS	Yes	Yes	SVM	LibSVM	65 attributes, instances: 1264 ASD, 462 non-ASD sources: Boston BAC	stepwise forward	5 features	SVM	56.2	87.95			Bone et al., 2016
2012	AQ	No	No	no machine learning	unknown	50 attributes (AQ), instances:1000 ASD, 3000 non-ASD	no	10	no algorithm	Adult 91%	Adult 88%			Allison et al., 2012

2.6 Chapter Summary

There have been few reviews on screening methods that have addressed common criterion, such as the number of items included in each screening test, time taken to complete the test, age categories involved, and performance (sensitivity, predictve accuracy, and specificity). However, existing reviews have failed to critically analyse vital aspects related to ASD screening, including the tool's accessibility, comprehensibility, popularity, and efficiency among others. More importantly, none of the reviews emphasise the importance of the DSM-5 criteria for evaluating the reliability of ASD screening. Therefore, this chapter investigated ASD screening methods to identify their performance in terms of different advanced parameters in order to discover possible concerns that need to be addressed through an innovative ASD screening process.

Different screening methods have been identified and categorised depending on their target audience in order to make the evaluation process more convenient. Hybrid screening refers to the ASD screening tools that consider the target audience as a combination of three or more of the following categories: infants, toddlers, adolescents, and adults. All screening tools have been critically analysed individually in terms of their evaluation, administration, target audience, scoring methods, other available versions, and performance. There is no screening tool performs completely well in terms of all the considered parameters. Some tools that were apparently performing well in terms of their sensitivities and specificities have been found to be unsuccessful in other parameters. For instance, ASIEP-3 is a highly accepted tool with a 100% sensitivity and acceptable level of specificity (81%). However, it consists of a series of activities and a questionnaire with 47 items, making it more time consuming than the other available tools. Similarly, CHAT is a tool that is efficient in terms of time, but is not freely available and has an unacceptable level of sensitivity (40%).

Many of the available screening tools, especially the short versions, comply only partly with the ASD criteria of the DSM-5 introduced in 2013. Most of the available tools were developed before that time and follow the guidelines established by the older version, DSM-IV. Apart from that, each screening tool has been discussed in depth in terms of accessibility, comprehensibility, administration, popularity, and performance. It has been revealed that many available screening tools, such as Q-CHAT require administration by well-trained professionals (at least during one stage of the evaluation). Moreover, M-CHAT and Q-CHAT are not comprehensive in terms of the

size of the audience they cover, but for an individual who is looking for a screening tool to identify autistic traits in an infant aged from 16 to 36 months these are appropriate methods as they perform well in terms of sensitivity and specificity. Similarly, AQ-10 (Adults and Adolescents) can be recommended for individuals aged 12-16 years and 18+ respectively, as they are time efficient and can be self-administered with an acceptable level of performance.

Lastly, we reviewed approaches that adopted machine learning in autism diagnosis in Section 2.5. The findings of this chapter emphasise the need for a more efficient, intelligent, and innovative ASD screening tool that can cover a wider audience while maintaining high levels of performance. It is believed that in the near future a highly interactive platform, utilising an intelligent machine learning screening algorithm, will offer more accurate and robust performance that engages individuals with ASD (both children and adolescents), parents, caregivers, GPs, other medical staff, researchers, and the broader population. This is due to such intelligent screening methods being useful by offering both the diagnosis and the individual development plan needed for cases to their families and healthcare providers efficiently. Therefore, in the next chapter we discuss a new machine learning model and explain in details the method used to collect the data, the data collection process, feature selection process, the learning process and the ASD classification step.

Chapter Three[3]

Machine Learning Method for ASD Screening

3.1 Introduction

This chapter investigates the applicability of rule-based classification relative to ASD detection and proposes a model that learns vital rules for ASD classification. In particular, a new classification model based on Covering approach called Rule Machine Learning (RML) is proposed. This model offers automatic classifications systems (classifiers) represented as rule sets. The rule sets can be used by health professionals to assist them in the diagnosis process or to advise individuals and their families whether they should seek a further evaluation. The rules offered by the RML can be easily interpreted by novice users as well as parents, teachers, caregivers and family members. These rules are discovered through training on a historical instances (cases and controls) by the ML classification technique.

The RML is part of an intelligent architecture that also consists of a data collection, data pre-processing, feature selection, model construction and evaluation among others. The RML was evaluated against a real dataset collected using a mobile application called ASDTests (See Sections 3.2) and recently published at the University of California Irvine repository (UCI) (Lichman, 2013). The experimental tests in Chapter 4 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 4.5 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 chapter is structured such that Section 3.2 proposes the ML architecture for detecting autistic traits. In this section, details related to the main components of the ML architecture

[3] This chapter has been accepted for publication in the Journal of Health Informatics and will be published in February 2019.

including data collection method, transformation method(s), feature selection and classification among others, are discussed in details. Finally, the chapter summary is presented in section 3.3.

3.2 The Proposed System Architecture for Detection of Autistic Traits

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 favourable to novice users (Bunker & Thabtah, 2017; Mohammed, et al., 2014; Abdelhamid et al., 2013). In this section, we propose a new ASD classification model called RML based on the architecture shown in Figure 3.1. RML is based on Covering classification which employs a search method for rule discovery. The model then evaluates the discovered rule and discards any redundancies. Hence, only rules that have classified cases and controls 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 3.1, data are collected by a mobile application called ASDTests that implements four different ASD screening methods (AQ-Adult-10, AQ-Adolescent-10, AQ-Child-10, Q-CHAT-10. For the purposes of this research project, focus was on the adult, adolescent and child modules and researchers utilised the dataset collected for these target age groups. Once the raw data are obtained, a number of pre-processing operations were applied, including missing values replacement and discretization for certain continuous attributes such as the age of individuals. Moreover, a feature selection method was used to remove features that were redundant and may have possibly created biased results (See Section 3.2.1 for further details on data features). Once the raw data are 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 initial model is 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 been classified yet). 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 taken 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. In the next subsection we discuss the ASD classification architecture phases in further details.

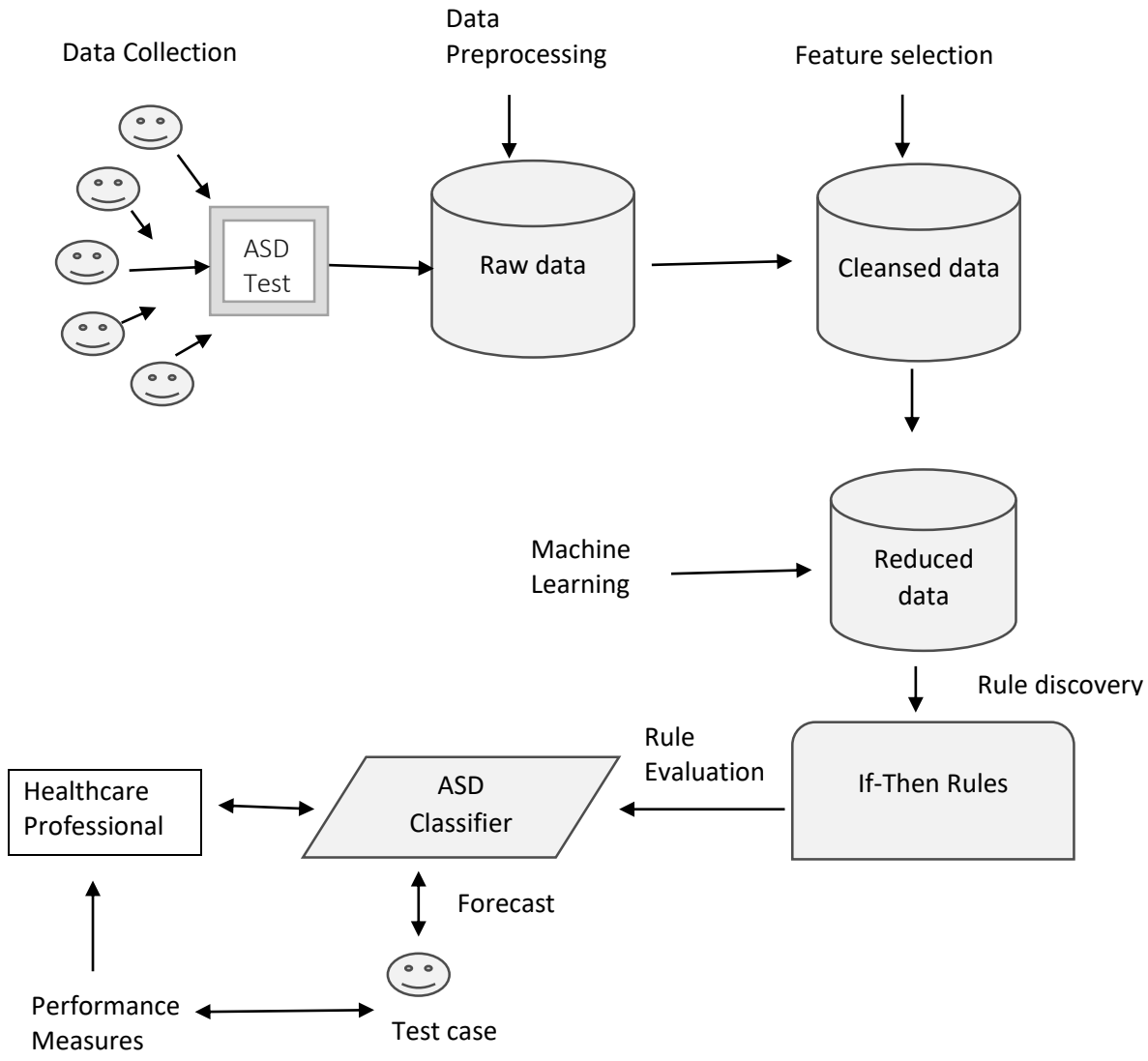


Fig. 3.1 The proposed ML architecture for ASD classification

3.2.1 Data Collection

To accomplish the necessary behaviour characteristics for cases and controls a mobile application (app) was designed and implemented for all user age categories, ASDTests app is based on short versions of the AQ and Q-CHAT screening methods (Alison, et al., 2012; Robins, 2001). It

includes four key screening questionnaires, each of which contains ten questions based on the user's age category as shown in Table 3.1. To be exact, there is a questionnaire for infants (less than or equal to36 months), children (4-11 years old), adolescents (12-16 years), and adults (17 years old and greater) as displayed in Figure 3.2. This figure shows the navigation diagram of the ASDTests app and on a higher level the key screens. In this section the functionalities, as well as the data collection requirements of the proposed medical app, are highlighted in detail.

Besides the landing, information, and terms & conditions screens, the key function of the ASDTests app is to enable different types of user to measure ASD traits using four different modules (discussed in in Chapter 2). Initially, the user from the landing screen (Figure 3.3a) chooses the language from 11 available languages, and then selects their test based on the age category screen (Figure 3.3b). Based on the age chosen, the user is relayed to the appropriate ASD screening test. Each test consists of ten questions in a sequential order, and each is associated with an image to enable users to carefully select the appropriate answer. Users can use touch screens to navigate through the app, which can be run on smart phones (Android and IOS) as well as tablets. Figure 3.3c displays one sample question from the toddler test. Once the user completes the test (goes through the 10 questions) then a screen emerges to review the answer (Figure 3.3d). In this screen, the user can review their answers and amend any they wish. The screen serves as a quality assurance measure so that users can review and verify answers before progressing to the data input screen (Figure 3.3e).

Prior to completing the assessment, an ethical approval was obtained from the University ethics committee and participants were required to consent to a disclaimer which explained the goal of the research, privacy policy, and use of the data. Participants were notified that their information would be kept anonymous and only shared for research purposes. During the data collection phase there was no direct access to participants and the ASDTests mobile application clearly stated that use of the data would be for research purposes only. The participants had to read this before submitting their answers. Since no name or sensitive information is involved, participant identities are anonymous (see variables in Table 3.1 and their description in Table 3.2).

Table 3.1: Details of variables in the child, adolescent and adult screening methods

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

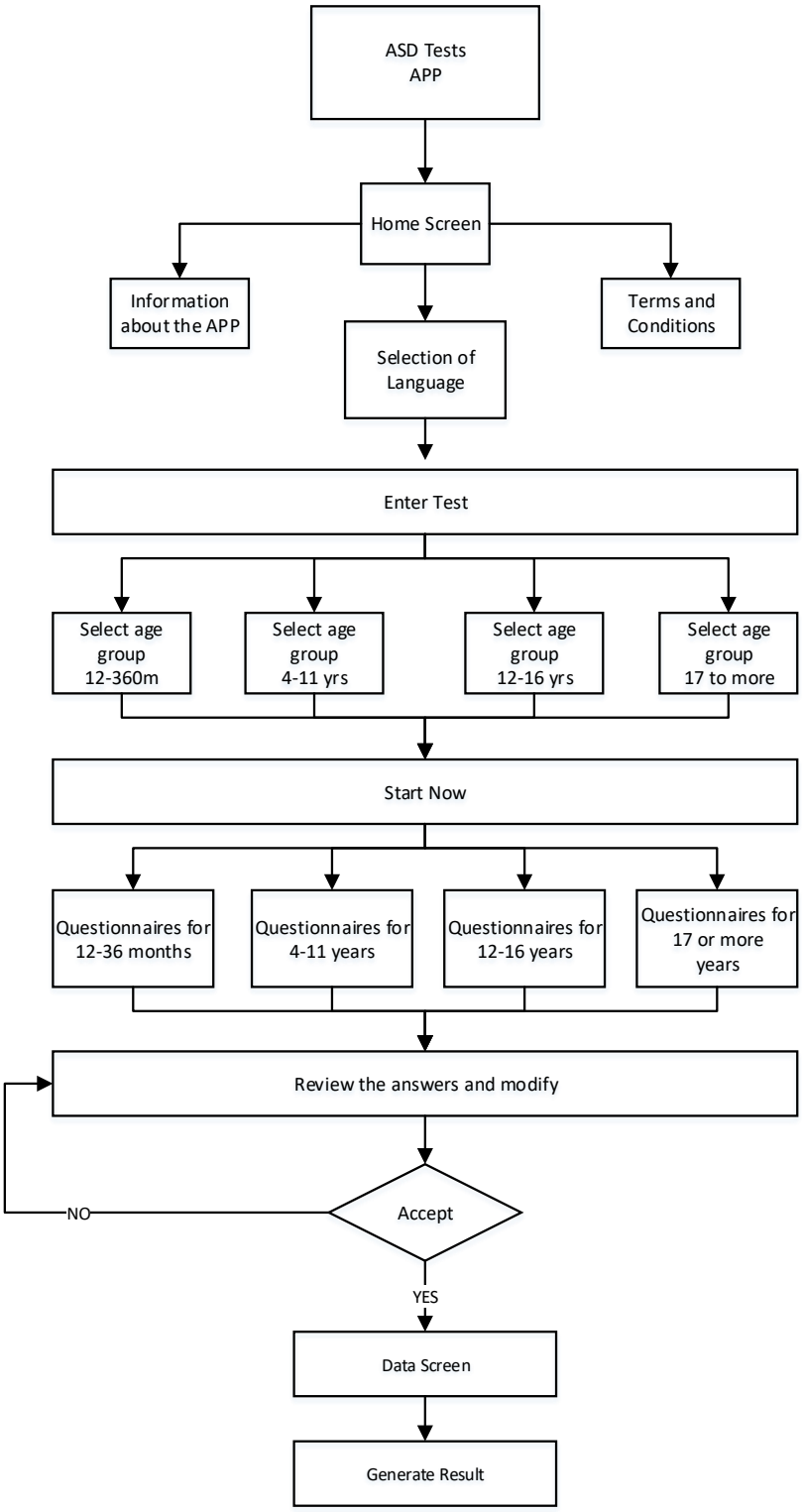


Fig.3.2 The proposed app (ASDTests) navigation diagram

The key functionality of the data input screen is to collect relevant useful data about the case undergoing the screening. In particular, the features shown in Table 3.2 are collected. These

features are stored in a MYSQL database and can be used for further data analysis later on to understand key features that may influence ASD diagnosis from a behavioural science perspective.

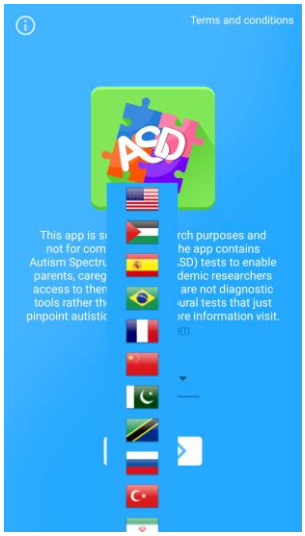


Fig. 3.3a Landing screen

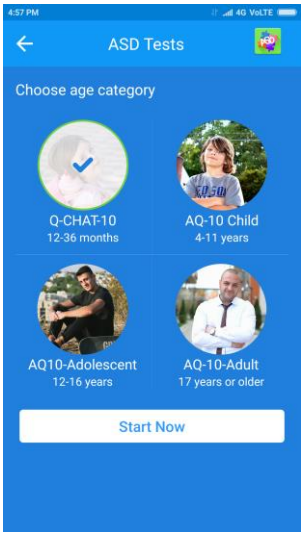


Fig. 3.3b Age selection screen

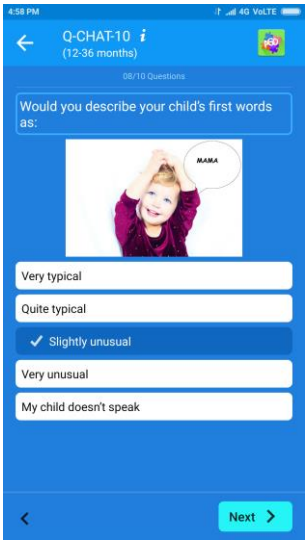


Fig 3.3c A sample question: toddler's test

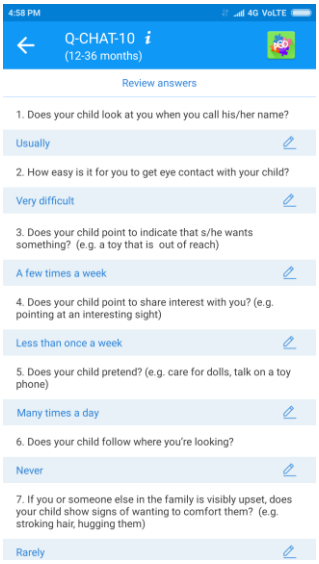


Fig .3.3d Answer review screen

Fig. 3.3e Data collection screen

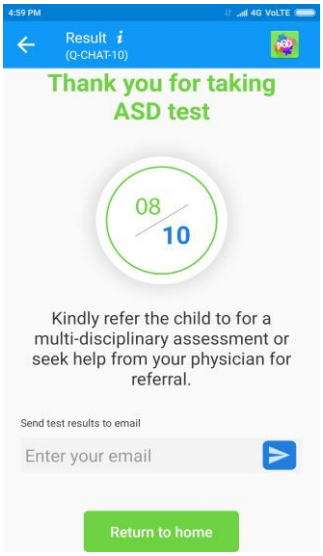


Fig. 3.3f Results screen

Once the user clicks the submit button on the data screen, they will be automatically redirected to the results. On the back end, a function is built to compute the score and then corresponds each score with the appropriate text that the users will see in the results screen. For instance, say a parent

Table 3.2: Features in the Child, Adolescent, and Adult Datasets

No	Independent Variable	Data Type	Comments
1.	A1	Binary	See Table 3.1
2.	A2	Binary	See Table 3.1
3.	A3	Binary	See Table 3.1
4.	A4	Binary	See Table 3.1
5.	A5	Binary	See Table 3.1
6.	A6	Binary	See Table 3.1
7.	A7	Binary	See Table 3.1
8.	A8	Binary	See Table 3.1
9.	A9	Binary	See Table 3.1
10.	A10	Binary	See Table 3.1
11.	Age	Continuous	Age of participant
12.	Gender	Binary	Male or Female
13.	Ethnicity	Categorical Data	Chosen from a list of predefined values
14.	Jaundice	Binary	Yes or No
15.	Family History	Binary	Whether any family members diagnosed with autism
16.	User type	Categorical	Person taking the test (parent, self, relative, caregiver, etc)
17.	Language	Categorical	used for taken the test such as English, French, Spanish, Portuguese, Farsi, Arabic, Turkish, Russian, etc.
18.	Taken the test before	Binary	If the individual had taken the test before
19.	Score	Continuous	The score generated by the screening method function
20.	Test_type	Categorical	Based on the user age four types exist (toddler, child, adolescent, adult)
21.	Country_Residence	Categorical	The country of residence of the individual (drop list)
22.	why_are_you_taken_the_screening	Categorical	Users insert the reason(s) of why are they taking the screening
23.	Target class	Binary	0 (No ASD traits) or 1 (ASD traits)

is taking the ASD test for a child and the score obtain was 8. This indicates the child should be referred to a health professional for clinical assessment of ASD. The computations of the scores for the four modules (tests) in the app have been coded to enable users to obtain the score as well as its interpretation in a facilitative way. Moreover, the user can also email the results using the

email functionality where a professional grade PDF file that contains the necessary information about the case under consideration (case features, answers given, screening results, disclaimer, etc) is constructed and emailed to the user. Finally, at any time during the ASDtests app the user can terminate the test and return to the main screen. The ASDtests app is available for free to download on iTunes and Google Play stores.

3.2.2 Data Transformation and Pre-processing

Besides the main features that are related to the screening of ASD and the case under consideration (Table 3.2), a target class variable has been created with a Boolean value to determine whether the individual undergoing the test has ASD or not. The class variable value is assigned automatically based on the final score obtained by the individual taking the ASD test. For example, if the individual has selected an age category of 12-16 on the ASDTests app the scoring will be based on the AQ-Adolescent method. In this case, if the final score was between 6 and 10, the class value for this case will be assigned "Yes," otherwise it would be assigned "No." A class value with "Yes" relates that the case requires further assessment by an expert while a class value with "No" indicates that the individual has no autistic traits.

Table 3.1 displayed earlier depicts the mapping between A1-10 features shown in Table 3.2 and their corresponding questions in the screening methods to reveal the differences and similarities of the features. The values in the A1-A10 variables in each 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 AQ-10-Child method, "1" was given for questions 1, 5, 7, and 10 if the participants answered any of them with "Definitely" or "Slightly Agree"; and "0" otherwise. For the rest of the questions "1" was considered if the answer was "Definitely" or "Slightly Disagree"; otherwise "0" was assigned. For the AQ-10-Adolescent method, "1" was allocated to questions 1, 5, 8 and 10 if the given answer was "Definitely" or "Slightly Agree" for each of those questions whereas "1" was allocated to "Definitely" or "Slightly Disagree" on the remaining questions. Lastly, for the AQ-10-Adult method, "1" was given for "Definitely" or "Slightly Agree" answers for questions 1, 7, 8, and 10. For the rest of the questions in this method "1" was allocated when "Definitely" or "Slightly Disagree" was chosen for questions 2, 3, 4, 5, 6, or 9. This representation of "1" or "0" per feature in the screening method can ease the process of

data processing by the machine learning algorithms during the building of the classification systems.

To the prior data processing phase, we discarded a number of variables from the datasets as they either redundant or do not have added value. To be exact, below are the variables that have been discarded along with rationale:

- Screening_type: This variable is redundant since we hold the user age and therefore we can obtain the type of test. We only use this variable to split the SQL table into different types of data based on the screening type.
- Final score: This variable is redundant since we have replaced it with a binary class based on the score level obtained after the screening. If this variable is kept then the models learnt will be overfitted.
- why_are_you_taken_the_screening: This variable gives us the reason(s) on why individuals are participating in the screening process and thus it has nothing to do with the data processing.
- Language: This variable is automatically generated based on the language used to perform the screening process. This variable has no impact on the behaviour indicators of ASD.

The age variable has been also discretised using Entropy discretization method prior data processing. In discretization, the continuous values are mapped to a set of categorical values to ease the search of items during data processing.

3.2.3 Feature Evaluation and Selection

Chi Square (CHI) and Information Gain (IG) are two different methods normally used in classification research to evaluate the worthiness of variables in an input dataset (Shannon 1948; Quinlan, 1986; Liu & Setiono, 1995). The aim of these methods is to reduce the dimensionality of the training dataset by producing a smaller number of variables so that efficiency or the quality of predictive models can be enhanced. The worthiness of the variable is typically measured statistically, using specific metrics by computing the correlation between each variable and the target variable (class). In the case of ASD screening, the variable will be an item/question in the screening method, and the class is whether an individual has ASD traits. For instance, CHI-SQ

computes the correlation between variable-class (*v*,*l*) using their expected and observed probabilities in the training dataset (T) based on Equation (3.1),

$$CHI - SQ(v,l) = \frac{S \times (AD - BC)}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (3.1)$$

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

On the other hand, IG calculates the worthiness of a variable, using the entropy of the class with and without the presence of a variable according to Equation 3.2,

$$G(l,v) = E(l) - E(l|v) \quad (3.2)$$

where l is the class, v is the variable, E () is the Entropy and $G()$ is Gained information. So, the IG for a training set $T = \{(v, v, \dots, v,\ l)\}_i$ where v_a is the value of the vth variable, and l is the corresponding class, can be calculated for vth by Equation 3.3:

$$G(l,v) = -\sum_{l \in L} p(l) \log p(l) + p(v) \sum_{l \in L} p(l|v) \log p(l|v) + p(\bar{v}) \sum_{l \in L} p(l|\bar{v}) \log p(l|\bar{v}) \quad (3.3)$$

One of the notable problems with statistical data reduction methods is the discrepancies in their variables' scores, especially when applied to different datasets. For example, when the CHI-SQ and IG methods are applied to the "arrhythmia" dataset (Lichman, 2013), they select 110 (37 variables respectively, using 15 and 0.15 minimum thresholds for CHI-SQ and IG). If minimum thresholds values are tuned by the end-user, the results derived by the filtering methods may change significantly. The results variation is attributed to the different formulas adopted by the filtering methods when they pre-process the features in the datasets. The "arrhythmia" dataset example, if limited, shows that results produced by filtering methods are highly diverge because of the different metrics used in computing the scores. Therefore, there is a need to unify the scores of variables in

order to stabilize them especially in applications such as autism screening where not having an important feature can give a wrong screening result. Unification can be achieved by utilising multiple scores per variable that can be initially normalised. By normalising the scores of methods similarities and deviations will be minimised and a unified global weight can then be assigned to each variable which may result in less redundant features. This is the basic idea of the feature selection method (Va), which is part of the ML architecture.

Va aims at unifying score volatility in IG and CHI-SQ by a) amalgamating their scores per variable, 2) Normalising their scores to one score per variable, and 3) utilising the new score as a new metric for ranking the variables. By accomplishing this idea, a new truer rank will be given to each variable which reduces variable-to-variable correlations and maintains good variable-to-class correlation. This metric could substantially reduce the number of variables selected, especially for medical and behaviour science applications such as autism classification.

Va creates a new vector based on the CHI-SQ and IG scores and computes the magnitude of the vector as M_score. In Va, M-score is utilised as a robust measure to differentiate among the goodness of variables (features that are selected from the ones that get discarded). The mathematical formulations of Va are given in Equations 3.4-3.7. Initially, the scores derived by CHI-SQ and IG are different since each method employs a specific metric to evaluate the worthiness of variables, therefore these scores must be normalised in order to make them comparable. To accomplish this task, Va normalises the scores by defining CHI_SQ_{max} *and* IG_{max} to be the maximum score for a variable obtained by the CHI-SQ and IG methods on a training dataset, T (Equations 3.4-3.5). The normalised scores of the vth variable in CHI-SQ and IG results can then be defined.

$$\overline{CHI\_SQ_v} = \frac{CHI\_SQ_v}{CHI\_SQ_{max}} \tag{3.4}$$

$$\overline{IG_v} = \frac{IG_v}{IG_{max}} \tag{3.5}$$

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

$$sv_a = \begin{pmatrix} IG_v \\ CHI\_SQ_v \end{pmatrix} \tag{3.6}$$

The new score vector is fundamental since it holds important information related to the scores of CHI-SQ and IG. Magnitude of the score vector, sv_v, can be utilised as a scalar measure of the

vector as shown in Equation 3.7. This is since the magnitude of a vector can be computed by taking the square root of the sum of the squares of its coordinates. This new measure (sv_v) can easily be employed to compare between variables in which those with a larger sv_vcan assigned to a higher order and thus have an increased chance of being selected.

$$|sv_v| = \sqrt{(IG_v)^2 + (CHI_v)^2} \qquad (3.7)$$

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

Variables in the proposed Va method with low scores are eliminated and only variables with high scores are kept . To determine the line between influential and non-influential variables, any variable with a score below 50% of the largest score obtained on a dataset is discarded. Nevertheless, the user is allowed to determine the cut-off points (selection criterion) since application data varies based on the criteria of determining the acceptable score per variable. For instance, medical diagnostic applications, such as ASD screening, may require a small set of variables if compared with text categorisation or email classification and thus a cut-off score of 50% is ideal. The cut-off values have been suggested based on extensive warming up experimental analysis on different datasets in which 5%, 10%, 25%, 50%, and 75% values have been tested. The results showed stability on the number of remaining variables when the cut-off is set to 50%.

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

their decision making. In the context of ASD, patients, their families, and medical staff will only have to exploit a lesser number of autistic traits when seeking to understand the causes and effects of ASD, at least at the screening level.

3.2.4 Machine Learning Methods for the Detection of Autistic Traits

A new learning mechanism based on Covering classification, which is an enhancement of recently developed algorithm by the author, is proposed. The learning method pseudocode is shown in Figure 3.4. The learning algorithm utilises two thresholds named the Frequency (F) and Rule Strength (R_S) to find and extract the rules (Definitions 2 & 3 respectively). The F 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 or above the F 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 (R_S), which denotes the rule's items plus class frequency divided by the items frequency (see definition 5).

A rule is represented as $(A_1, v_1) \wedge (A_2, v_2) \wedge ... \wedge (A_k, v_k) \rightarrow C_n$ 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 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.

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. $([(A_1, v_1), (A_2, v_2), ..., (A_k, v_k)], C_n)$.

Definition 2: *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 $\frac{|[(A_1, v_1), C_n]|}{|T|} \geq Freq$

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

The learning algorithm (RML) initially scans the training dataset and discards 1-item that have failed the F threshold test. All remaining items with computed frequencies above the F 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 has no error rate. When this occurs, the rule is then saved into the classifier and all training instances linked with R are erased from the original training dataset. 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 R's data removal and 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' frequency are continuously updated since the training dataset is shrinking whenever a rule is generated. As a result of this dynamic 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.

RML 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 the algorithm 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, the proposed algorithm allows the generation of such rules as long as 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.

RML assumes that the variables in the training dataset are categorical (they are associated with 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 RML, 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 order 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 method (RML):

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 require 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 smaller number of 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 Chapter 4 for further details on the results).

1. $E\_S\_R \leftarrow \{\}$ //Empty set of rules storage
2. $r_1 \leftarrow \{\}$ // Empty rule
3. $Temp \leftarrow Train$ // temporary data storage
4. Do { // for each variable items over *Temp*
5. If [(p(Ai, vi) | c_i = I) /| $Temp$ |] >= *min_freq* { // Computing item actual frequency with the class
6. If [(p(Ai, vi) | c_i = I) /|(p(Ai, vi)] >= *R_S* { // Computing potential rule strength
7. $r_i \leftarrow (Ai, vi)$ //Append the max (Ai, vi) to r_i
8. Repeat steps 5-7 until r_i accuracy cannot improve
   }} // if statements
9. $E\_S\_R \leftarrow r_i$ // Insert the rule into the rules set
10. $Temp \leftarrow (Temp - Locate\ (r_i, Temp))$ // Locate is a function that returns the row numbers of r_i in *Temp* and then erase these instances
11. $Train \leftarrow (Train - Temp)$ // update the original dataset after erasing r_i
12. Repeat steps 2-11
13. Exit when *Train* has no more instances OR all p(Ai, vi) have been tested
14. } // Do
15. Generate E_S_R
16. Classify Test (Test, E_S_R) // A function to classify cases and control based on the rules

Fig. 3.4 RML algorithm

3.3 Summary

A possible way to improve the classification accuracy and efficiency of the current screening tools is to adopt new intelligent methods based on machine learning and computational intelligence. This chapter has proposed a new ML architecture for detecting autistic traits that consists of multiple phases including data collection, feature selection analysis and classification. For the data collection phase, a new mobile application was designed and developed that captures important features related to behaviour characteristics of individuals. This has resulted in new datasets being collected. In the feature selection, a new method called Va was proposed in which it significantly reduces the number of features needed for ASD screening methods while maintaining sensitivity, specificity, and predictive accuracy rates. For the classification phase, we adopt a Covering classification method that learns useful and easy to navigate rules in the form of "if then" rules. Later in chapter 4, we show extensive experimental analysis on the upsides and the downsides of the ML architecture. In particular, the focus will be on the feature analysis and the classification of cases and controls.

Chapter Four[4]

Testing and Performance Evaluation

4.1 Introduction

In this chapter, the performance analysis of the proposed machine learning autistic trait model is presented based on the three subsets of data (Child, Adolescent, and Adult). Since the RMC has been implemented in the Java programming language using WEKA platform then all experiments have been conducted in the same platform for fair comparison. WEKA is a machine learning tool that contains large collections of learning, visualisation, filtering, and dimensionality reduction techniques. RMC predictive model have been integrated in WEKA's package called "Classifier" under "Rules". In addition, Va feature selection method has also been implemented in Java within WEKA under "Attribute Selection" package.

Ten-fold cross validation testing method has been adopted (described in Section 4.2) to build the classification systems from the distinctive feature sets derived by the filtering methods. Lastly, all error rates generated are averaged to come up with one global error rate for the classifier. All experiments have been performed on a computing machine with 2.0 GHz processor and 8 RAM memory.

We initially evaluate the features collected (the autistic trait feature sets) using the proposed feature selection method (discussed in Chapter 3). The feature analysis section in this chapter (Section 4.4) contains independent experiments from the ML model experiments to show whether the feature selection method (Va) has reduced feature to feature correlations and maintained features to class correlations. Once features have been selected then we evaluate the proposed ML model using these features sets on the three datasets (adult, adolescent, child) (See Section 4.5 for

[4] Parts of this chapter have been published in the Journal of Medical Informatics 117, 112-124, and other parts have been published in the Journal of Health Informatics (1460458218796636).

further details). The basis of the comparison for all experiments is different evaluation metrics; including sensitivity, specificity, and predictive accuracy among others (see Equations 4.1-4.6).

4.2 Evaluation Measures

Normally, the predictive model derived by machine learning is assessed with a number of evaluation measures such as predictive accuracy, precision, recall and harmonic mean among others (Power, 2011). For a binary classification problem (ASD, No-ASD) as the basic form of the classification problem, Table 4.1A shows the possible answer for a test case prediction. Classification accuracy (Equation 4.2) is one of the most common evaluation measures. Using this measure, we can identify the number of test cases that have been correctly classified from the total number of test cases. Sensitivity (Equation 4.3) identifies the ratio of the test cases that have ASD (true positive rate) whereas specificity (Equation 4.4) is the ratio of the test cases who do not have ASD (true negative rate). Positive Predictive Value (PPV) and Negative Predictive Value (NPV) have also been included as shown in equations 4.5 and 4.6 respectively. NPV and PPV represent the percentages of negative and positive diagnostic test instances that are true negative and true positive outcomes respectively.

Table 4.1A: 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)

$$Error\_Rate = 1 - Accuracy \quad (4.1)$$

$$Accuracy = \frac{|TP+TN|}{|TP+TN+FP+FN|} \quad (4.2)$$

$$Specificity = \frac{|TP|}{|TP+FN|} \quad (4.3)$$

$$Specificity = \frac{|TN|}{|TN+FP|} \quad (4.4)$$

$$Precision\ Or\ PPV = \frac{|TP|}{|TP+FP|} \quad (4.5)$$

$$NPV = \frac{|TN|}{|TN+FN|} \quad (4.6)$$

$$F1 = 2 \times \frac{|\mathrm{Pr}\,ecisoin \times \mathrm{Re}\,call|}{|\mathrm{Pr}\,ecisoin + \mathrm{Re}\,call|} \quad (4.7)$$

Cross-validation is a testing method used to examine predictive models' performance (Kohavi, 1995). This method starts by splitting the training dataset into N partitions, i.e. often N is set to 10. The model is then trained on N-1 partitions and tested on the holdout partition. This procedure is repeated N times by randomly partitioning the training dataset. Finally, the average accuracy of the model is derived from the N runs. When the data is split, random shuffling with class representation is done to make instances for each class to exist in each partition. This process is called stratification, hence the testing method being known as stratified cross validation. In this thesis, we have used ten-fold cross validation in the testing of the machine learning algorithms and this testing method was implemented within RML model.

4.3 Datasets

We have used three different datasets in the experiments that have been collected using the proposed ASDTests screening app. To be exact, there is one dataset for children (Child dataset), one dataset for adolescents (Adolescent dataset) and one dataset for adults (Adult dataset). The child, adolescent, and adult datasets have been collected based on the AQ-10 Child, AQ-10 Adolescent and AQ-10 Adult screening methods as described earlier in Chapter three. Table 4.1B

shows 20 sample data instances that have been collected based on the AQ-10 Adult assessment. A total of 1,452 instances that belong to toddlers, children, adolescents, and adults were collected over a period of 4 months using the ASDTests app and based on Q-CHAT-10, AQ-10 Child, AQ-10 Adolescent, and the AQ-10 Adult screening methods respectively.

After an initial investigation on the collected instances it was clear that the vast majority of the instances that belong to toddlers have been associated with a "no ASD" class label, making such a group of data completely imbalanced. To be exact, 96% of the cases who participated in the screening test for the Q-CHAT-10 (toddlers) have not been associated with ASD, and therefore the toddlers instances are separated from other instances as well as omitted from further analysis, at least at the mean time. We intend to perform data analyses on the toddler dataset in near future when we have more instances with ASD to minimise any statistical insignificance. Though, the under sampling or over sampling methods can be utilised to balance the dataset it can also cause

Table 4.1B: Sample sixteen instances from the Adult dataset

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age	Sex	Ethnicity	jaundice	Family with austim	Residence	used_ app_ before	Who is taken the test	ASD
1	1	1	1	0	0	1	1	0	0	26	f	White	no	no	USA	no	Self	NO
1	1	0	1	0	0	0	1	0	1	24	m	Latino	no	yes	Brazil	no	Self	NO
1	1	0	1	1	0	1	1	1	1	27	m	Latino	yes	yes	Spain	no	Parent	YES
1	1	0	1	0	0	1	1	0	1	35	f	White	no	yes	USA	no	Self	NO
1	0	0	0	0	0	0	1	0	0	40	f	White	no	no	Egypt	no	Self	NO
1	1	1	1	1	0	1	1	1	1	36	m	Others	yes	no	USA	no	Self	YES
0	1	0	0	0	0	0	1	0	0	17	f	Black	no	no	USA	no	Self	NO
1	1	1	1	0	0	0	0	1	0	64	m	White	no	no	UK	no	Parent	NO
1	1	0	0	1	0	0	1	1	1	29	m	White	no	no	USA	no	Self	NO
1	1	1	1	0	1	1	1	1	0	17	m	Asian	yes	yes	UK	no	Health care professional	YES
1	1	1	1	1	1	1	1	1	1	33	m	White	no	no	USA	no	Relative	YES
0	1	0	1	1	1	1	0	0	1	18	f	Arab	no	no	USA	no	Parent	NO
0	1	1	1	1	1	0	0	1	0	17	f	White	no	no	USA	no	Self	NO
1	0	0	0	0	0	1	1	0	1	17	m	Arab	no	no	Austria	no	Self	NO
1	0	0	0	0	0	1	1	0	1	17	f	White	no	no	UK	no	Self	NO
1	1	0	1	1	0	0	1	0	1	18	m	Arab	no	yes	UK	no	Parent	NO

biased results especially when the data is completely unbalanced such as the case of the toddler dataset. Table 4.1C show basic statistics about the considered datasets.

Table 4.1C: Summary of the instances in the considered datasets

Dataset	Average Age (years)	Age Range (years)	Gender Distribution		Instances with Screening Decisions	
			Female	Male	ASD	No-ASD
Adult	29.20	17-64	337	367	189	515
Adolescent	14.13	12-16	50	54	63	41
Child	6.35	4-11	84	208	141	151

This left 1,100 instances that belonged to three target audiences (children, adolescents, and adults). In addition, since the "scoring result" was conventionally used by the AQ-10 version to classify cases and controls, this variable has been discarded prior to data processing.

Figure 4.1A shows the instances of distribution with respect to age, and Figure 4.1B shows the class distribution per age category. It is clear from Figures 4.1A and 4.B that there are more adult instances than adolescent and children, as well as more instances associated with the "No ASD" class label. A basic explanation for more none ASD cases is that the population normally contains a much higher individuals with no ASD traits than those with ASD markers. Moreover, Figure 4.1B reveals that child instances are somewhat balanced with respect to class labels when compared with adult and adolescent instances respectively.

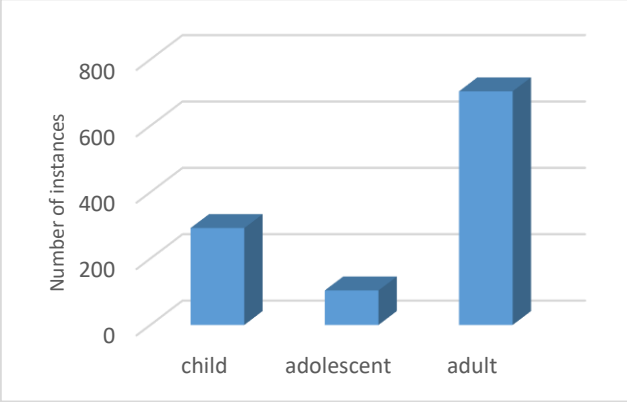


Fig. 4.1A the distribution of instances (with and without ASD traits) per age group

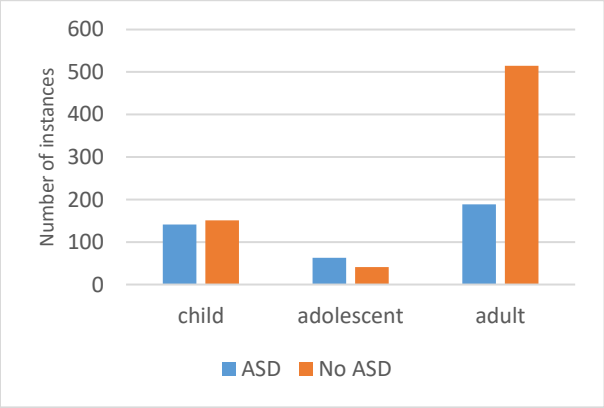


Fig. 4.1B the distribution of age instances (with and without ASD traits) per class label

The average age in years for children, adolescents, and adults in the three datasets are 6.3, 14.1, and 29.7 respectively. More male instances have occurred than female, as the number of instances for male and female in the three subsets (1100 total instances) are 625 and 475 respectively. The top ethnicities which participated in the data collection were Caucasian-European, Asian, Middle Eastern, South Asian, and African/African-American with 381, 185, 128 and 65 respectively. In the three subsets of data, there were 707 and 393 instances linked with the "ASD trait" and "No ASD trait" class labels. More of the tests for adults have expectedly been taken by the individuals themselves while many tests for the child category have been taken by parents, teachers, or caregivers. Among the gathered instances, there were 194 cases with family members diagnosed with ASD. Lastly, some values were missing in variables such as ethnicity and who_is_taken_the_test. Any missing value was denoted by the symbol "?" and was treated as any other value in the datasets. For data transformation and discretisation processes refer to Chapter 3 for further details.

4.4 Feature Selection Results Analysis

4.4.1 Experimental Setting

In this section, we evaluate the autistic trait features in the three datasets using the proposed feature selection method and the obtained performance is then compared with sets chosen by four other common feature selection methods, namely IG, Correlation Attribute Evaluation, Correlation Features Set (CFS), and CHI (Quinlan, 1986; Witten & Frank, 2005; Hall, 1999; Liu & Setiono, 1995). The key is to determine a small set of effective features that can assess the different stakeholders and understand symptoms that red flag autism detection. The case of no feature selection is also considered. Reasons behind choosing these filtering methods are twofold:

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

Two machine learning algorithms have been employed in this set of experiments beside the feature selection methods, named Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and C4.5 (Decision Tree) (Cohen 1995; Quinlan, 1993). These predictive models have

been adopted in order to produce ASD classification systems from the different subsets of features chosen by Va, IG, Correlation, CFS, and CHI. These classification systems will show the true performance (upsides and downsides) of the Va when contrasted with different filtering methods, particularly predictive accuracy, sensitivity, and specificity among others. The reason for employing two different predictive algorithms is to generalise the results obtained, especially goodness of the distinctive features. RIPPER is a rule-based classifier similar to RML that normally employs excessive rule pruning to generate rules. Meanwhile, C4.5 is a decision tree algorithm that constructs classifiers using Entropy in the format of trees. Both algorithms are well studied in the machine learning and data mining communities, and produce high quality results with respect to classification accuracy according to different experimental studies by Abdelhamid, et al. (2017) and Mohamed, et al. (2014) among others.

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

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

4.4.2 Number of Features Selected

Figure 4.2 shows the number of features chosen by the filtering methods. It is clear from the derived figures that Va consistently selects a lesser number of features for all datasets considered in comparison with the other methods. Specifically, Va selected 8, 8, and 6 items from the AQ-Child, AQ-Adolescent, and AQ-Adult datasets respectively. This indicates that the proposed method not only seeks for feature-class correlations but also eliminates feature-feature correlations, leading to less redundant features in the final set. It is notable from the figures that the remaining number of features in Va from the larger datasets is smaller than those from the smaller datasets. In particular there were 6 features remaining after applying Va against the Adult dataset, which is an indication that Va works well in situations where instances are more represented among class labels. This may boost the performance of classifiers generated against the features selected by Va during assessment of large datasets when compared to those selected

from the limited size datasets. To investigate this case the features selected by Va from the adult datasets are displayed in Table 4.2A.

Table 4.2A shows that there are six common features (items 3, 4, 5, 6, 9, and 10 in the AQ-10 Adult screening method) detected by Va from the adult dataset. When these items were checked in the AQ-10 Adult screening method, it was discovered that most of these items correspond to social interaction and social communication (Category A) in the DSM-5 manual (American Psychiatric Association, 2013). However, these items do not fully fulfil the new ASD diagnostic criteria under DSM-5. For instance, items 4-6 and item 10 are covering conditions under categories A and A.1 in DSM-5 (social communication and social interaction). This means that despite these items fulfilling multiple criteria in Category A, they still do not cover any condition in Category B (Restricted and Reparative Behaviour) whiles the DSM-5 requires at least two criteria in category B to be met before an individual can be diagnosed with ASD. Therefore, these sets of items do not comprehensively cover the required minimum conditions for meeting an ASD classification. Nevertheless, screening for ASD does not necessarily require fully meeting the diagnostic conditions of ASD as its ultimate aim is merely to reveal potential autistic traits rather than diagnosing individuals since to do so necessitates the involvement of expert clinicians and clinical setup. Identifying the most influential features in AQ short versions to their minimum number is, therefore, a definite advantage.

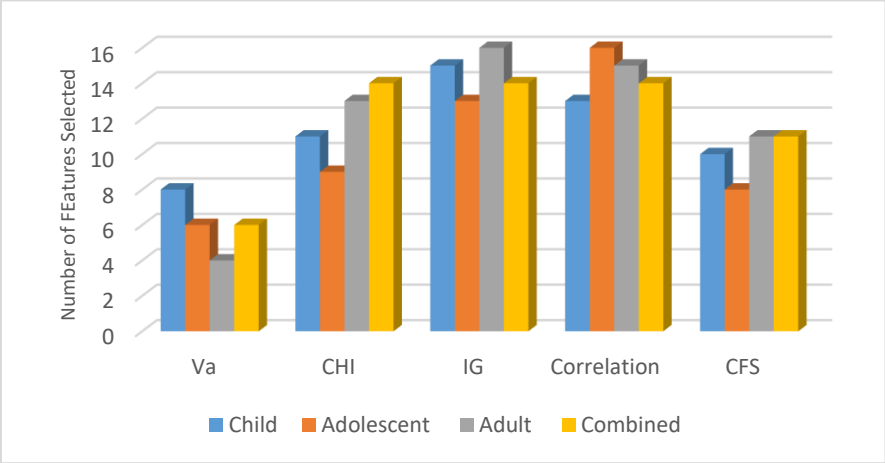


Fig. 4.2 Numbers of variables selected by the considered filtering methods from the child, adolescent, adult and combined datasets respectively

Table 4.2A Features remained along with their scores on the AQ-Adult dataset after applying Va

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

Table 4.2B Features remained along with their scores on the AQ-Adolescent dataset after applying Va

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

Table 4.2C Features remained along with their scores on the AQ-Child dataset after applying Va

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

Tables 4.2B and 4.2C show the reduced sets of items detected by the Va method for the AQ-10 Adolescent and AQ-10 Child assessments. The results show a reduction in the items chosen by Va from these two screening methods, detecting 8 items from the adolescent and child datasets respectively. An interesting finding based on the figures within these tables is that there are 5 items in common between the AQ-10 Adolescent and AQ-10 Child screening methods as detected by Va (highlighted in yellow within Tables 4.2B and 4.2C). In addition, item_4 (adolescent) can also be matched with item_4 (child), making the common items between AQ-Adolescent and AQ-Child significantly high (six out of eight). These results demonstrate that there are high levels of overlapping between the AQ-10 Child and AQ-10 Adolescent screening methods, at least at the autistic traits level. Less overlapping occurs between the AQ-10 Adult and AQ-10 Adolescent methods as detected by Va. Overall, the results clearly show that the top three items detected by Va from the AQ-10 adolescent and AQ10 -Child screening methods are related to communication and social traits. The top three items selected by Va from the AQ-10 Adult screening method, however, are related to behaviour and social traits.

Table 4.3 depicts the percentages of relative difference (Equation 4.7) of features chosen from the datasets between Va and the considered methods. The rates show that Va reduced the number of remaining features significantly when compared with the considered filtering methods. For example, Va minimised features in the Child dataset by 27.3%, 46.7%, 38.5%, and 20% when compared with results obtained by the CHI, IG, Correlation, and CFS filtering methods. The reduction is also clear in the adult dataset, where Va reduced the number of features by 53.8%, 62.5%, 60.0%, and 45.5% respectively when compared to results obtained by CHI, IG, Correlation, and CFS. The results of Table 4.3 clearly reveal that in the screening methods the Va method was also able to cut down the number of specific items as it takes into account multiple scores per feature and then normalises these scores into a new, single, global score that is then assigned to the feature. This gives the accurate rank per feature as Va reduces the deviations of the feature scores and ensures stability and the true weight per feature.

$$\frac{(\#\text{ of features chosen by Method i} - \#\text{ of features chosen by Va})}{\#\text{ of features chosen by Method i}} \quad (4.7)$$

Table 4.3 relative reduction of the variables selected in % from the ASD datasets based on the considered filtering methods versus Va

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

4.4.3 Accuracy, Sensitivity Specificity, PPVs, NPVs Results based on Feature Selection

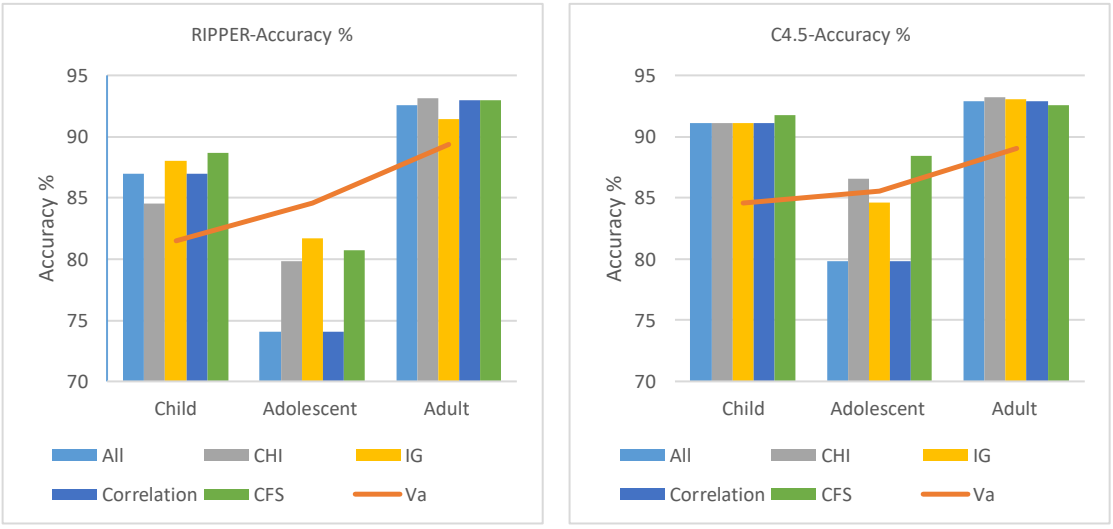
To reveal the performance of Va when compared with other filtering methods dealing with the ASD classification problem, a number of experiments using two machine learning algorithms, RIPPER and C4.5, have been conducted on the sets of feature data obtained in Section 4.4.2. Answers to the questions "will Va maintain the performance despite reducing the number of features" and "what will be the differences in sensitivity, specificity, PPVs, NPVs and accuracy between Va and other known filtering methods assessing the different ASD datasets" were sought. To answer these questions, instances that belong to each feature set, obtained earlier using the RIPPER and C4.5 algorithms, were processed.

Typically, the classification of the features obtained by the Va method in this section are performed using rule-based classifiers that produce rules. These rules have been derived by RIPPER and decision tree algorithms for the feature selection experiments. Using each training dataset (i.e. Child, Adolescent, Adult) as well as each target class, RIPPER starts with an empty rule (If empty then ASD) and then appends features into the rule until it cannot grow any further. At this stage, RIPPER evaluates the rule against a pruning set in order to improve its predictive accuracy. Once the rule gets evaluated, RIPPER generates the rule and removes its corresponding training instances and repeats the process on the remaining uncovered instances until no more data is left for class ASD. On the other hand, C4.5 uses Entropy to build decision trees that in turn are converted into rules sets. In classifying a test instance, both RIPPER and C4.5 assign the class of the first rule that matches the test instance's items (features values) to the test instance. There was no involvement of the "scoring result" variable in the classification process of the machine learning algorithms. Only features that have been chosen by Va have been used in building the classifiers. This offers a new way of reducing the amount of features by offering limited influential variables

for autism screening to the machine learning algorithms in order to enhance ASD screening performance.

Figures 4.3A and 4.3B show the predictive accuracies (calculated based on Equation 4.2) for the different subsets of data chosen by the filtering methods and using the C4.5 and RIPPER algorithms. In these figures the accuracies derived by C4.5 and RIPPER against CHI-features, IG-features, Correlation-features, Va-feature, and "no feature selection" (the original features) were considered. The accuracy results revealed that Va scales well with IG, CHI, Correlation, and CFS, especially for the Adolescent data based on the derived classifiers. Despite the slight drop in accuracy for Va derived features for the Adult and Child datasets, Va maintained an acceptable accuracy rate and, more importantly, was able to significantly reduce the number of features. Specifically, for the largest dataset (Adult) the C4.5 algorithm derived classifiers from Va feature sets with just under 2.8% less accuracy than those of IG, CHI, Correlation, and CFS. If the Va only selecting 6 features from the Adult dataset is considered, this 2.8% drop in accuracy can be tolerated in light of the more than double in number of features remaining in these datasets by the other filtering methods considered (features that relate to both screening items and individuals characteristics).

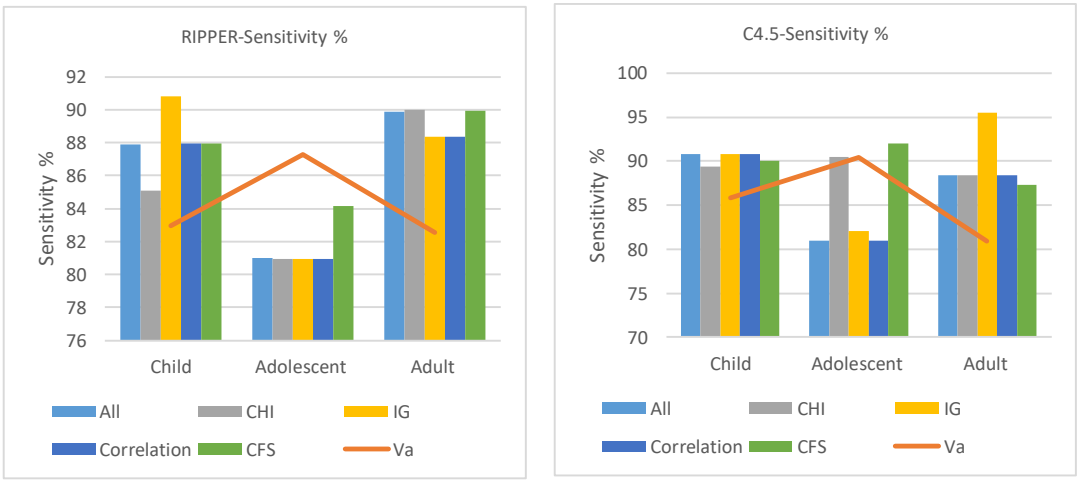
Surprisingly, when the Va chosen feature set of the Adolescent data is mined, the highest accuracy rate when using the RIPPER algorithm is derived in comparison with all remaining feature subsets. In fact, the accuracy rate of the classifier derived from the Va feature set was 10%



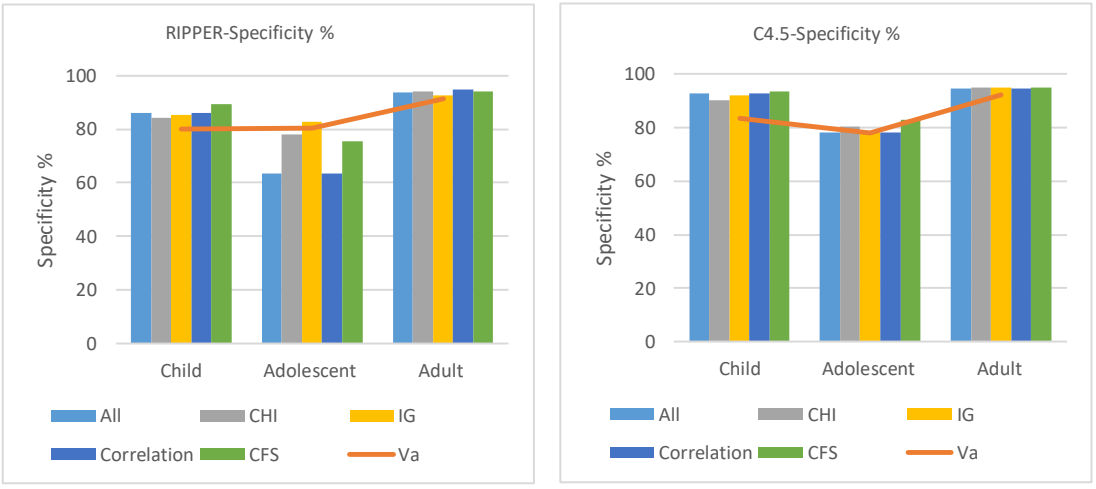
Figs. 4.3A & 4.3B Classification accuracies of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods

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

Figures 4.4A – 4.4B and 4.5A-4.5B display the sensitivity and specificity rates derived by the RIPPER and C4.5 classifiers from the different distinctive feature sets' data. Often in medical research, including autism, acceptable levels of sensitivity and specificity should be at least 80% (Towle & Patrick, 2016). The results of the sensitivity and specificity rates derived by the machine learning algorithms against the feature sets' data has shown an acceptable level, except for the Adolescent subset, for the majority of the filtering methods aside from Va. For that particular subset the specificity rates, derived by the machine learning algorithms against the Va and IG feature sets, showed adequate rates thus proving that these two filtering methods perform well even when a limited number of instances are present. When processed, the Va feature set showed slightly lower sensitivity and specificity rates on the Child and Adult datasets but maintained



Figs 4.4A & 4.4B sensitivity rates of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods



Figs 4.5A & 4.5B Specificity rates of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods

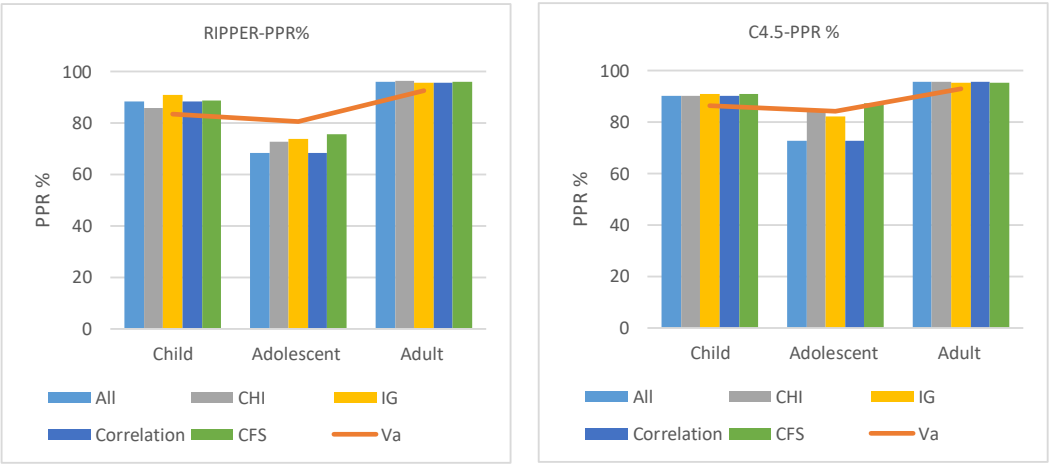
adequate rates. In particular, the specificity rates derived from the Va features of the adult dataset were 2.8%, 1.9%, 1.4%, 3.6%, and 3.0% less than those of the "no feature selection," IG, CHI, Correlation, and CFS feature sets. Va has also only used 6 features while original data, IG, CHI, Correlation, and CFS are associated with 21, 14, 14, 14, and 11 features respectively.

On the other hand, for the Adolescent dataset, Va was superior to "no feature selection," CHI, Correlation, and CFS, having achieved higher accuracy by 17.1%, 2.5%, 17.10%, and 4.9% respectively, thereby proving that Va can handle noisy data better than the other filtering methods. The sensitivity rates derived by RIPPER from the features of Va, CHI, IG, Correlation, CFS, and "no feature selection" on the AQ-Adolescent dataset were 87.30%, 80.95%, 80.95%, 84.13%, and 80.00%. These rates clearly show a good level of sensitivity by the considered filtering methods and the superiority of Va on this particular dataset.

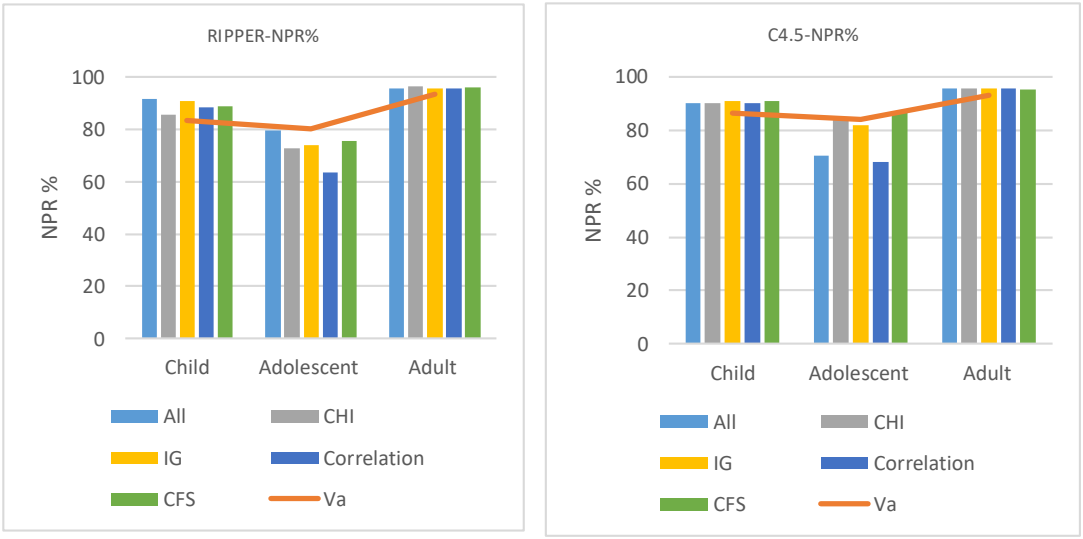
For the Adult dataset, the sensitivity rates derived by the C4.5 and RIPPER algorithms from Va's features were 80.95% and 82.54% respectively, and were lower than the results obtained by the remaining filtering methods. For this data subset, out of 189 positive instances that were supposed to have "ASD" classification the C4.5 algorithm misclassified 36 instances to ASD (False negatives). In addition, 14 instances were incorrectly classified by C4.5 as having ASD, having been screened as not on the spectrum (No ASD) (False Positives). These misclassifications have caused lower sensitivity rates for the features chosen by Va, at least for the Adult dataset. It is believed that this can be attributed to not having enough selected features that cover the needed criteria of ASD. In addition, some of these "No ASD" instances have overlapping features with

ASD cases while not fulfilling the entire criteria of ASD. These instances are the hardest to predict since they confuse the learning algorithm during the classification process, resulting in a slight increase in the false positive and false negative rates. Overall, Va has performed well in terms of sensitivity and specificity rates for the Adolescent dataset and comparably adequate to most filtering methods on the other two datasets (Adult, Child).

Figures 4.6A-4.6B and 4.7A-4.7B demonstrate the PPV and NPV rates derived by the RIPPER and C4.5 algorithms from the different datasets considered. These PPV and PNV results show acceptable levels, except for the Adolescent dataset. For this dataset the classifiers extracted by the RIPPER algorithm on Va's selected subset were superior in terms of PPVs and NPVs in comparison to those extracted from the remaining features sets. For example, the NPVs produced by RIPPER from the Va features set are 0.64%, 7.76%, 6.57%, 17.07%, and 4.88% larger than those derived from "no feature selection," IG, CHI, Correlation, and CFS feature sets respectively. On the other hand, the classifiers derived from the Va data subset showed slightly lower PPV and NPV rates on the Child and Adult datasets but maintained adequate rates. The PPV and NPV results were consistent with the predictive accuracy results derived previously. In the next section, we show how the proposed ML model boost the predictive performance in detecting autistic traits on the considered datasets.



Figs 4.6A & 4.6B PPV rates of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods



Figs 4.7A & 4.7B NPV rates of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods

4.5 RML Results Analysis

4.5.1 Experimental Settings

This section presents the experimental settings of the proposed ASD predictive model and other common ML algorithms based on rule induction, Bagging, Boosting, and decision tree approaches on the Adult, Adolescent and Child datasets. We used six different algorithms in addition to RML to reveal the performance of the proposed model. In particular, RIPPER, RIDOR, Nnge, Bagging, CART, C4.5, and PRISM algorithms have been adopted in the experimental results (Cohen, 1995; Gaines , 1995; Brent 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.

For example, C4.5 and CART construct decision tree classifiers that get converted into rules sets, and PRISM is a greedy algorithm that seeks rules that have 100% expected accuracy. On the other hand, RIPPER and RIDOR implement optimisation and pruning procedures to test rules. Lastly, Bagging and Boosting employ weak classifiers that in turn get merged to derive global rules sets. 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. Since the ASD screening process is a binary classification problem in which individuals are classified to either ASD traits or No ASD traits using characterised quantifiable variables. Therefore, performance evaluation methods that align with the binary classification problem in ML has been used. To be exact, evaluation metrics such as classification accuracy, specificity and sensitivity among others (see Section 4.2) were used.

4.5.2 Error Rate, Sensitivity, Specificity and Harmonic Mean Results of RML

Multiple experimental runs have been performed using different ML algorithms on the Child, Adolescent and Adult datasets to reveal the true performance of the proposed model. Figure 4.8 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. its error rate less than 5.6%.

In particular, for the Adult dataset, RML derived a classifier that has a lower error rate by 4.41%, 2.7%, 0.15%, 2.14%, 3.7%, 3.27%. 1.57% and 1.83% respectively than PRISM, CART,

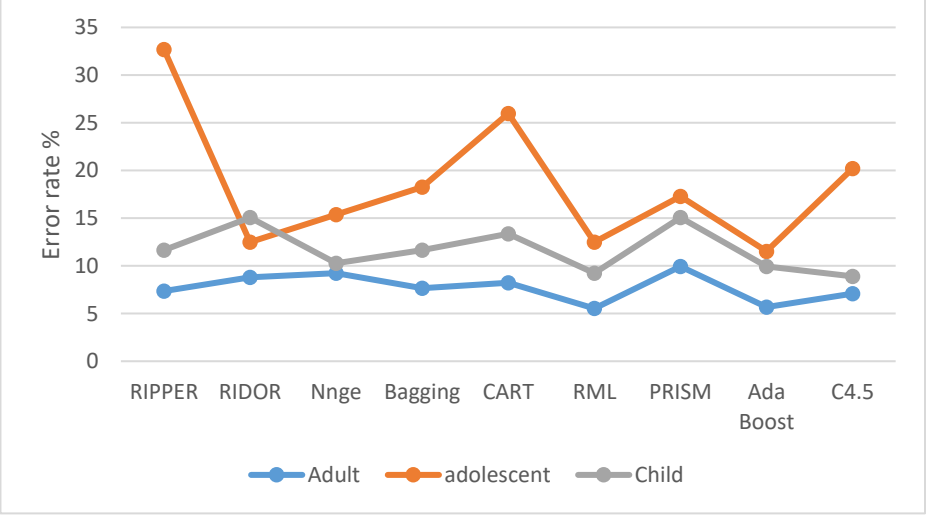


Fig. 4.8 Error rates derived by the considered ML algorithms on the adult autism dataset

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% less error than RML on this dataset. Nevertheless, RML, outperformed C4.5 on the Adolescent dataset by 7.69%. This is 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 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 in which most considered ML algorithms suffered from larger error rate due of not having enough instances in the training phase. 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% higher error rates respectively than that of RML model.

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 accurate classifier. In building the classification systems for detecting ASD, RML algorithm ensures that whenever a rule is generated all of its data instances are removed before learning the next rule from the training dataset. In addition, it amends 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 preliminary phase which reduces the search space and improves the efficiency of the training phase.

Figures 4.9A & 4.9B display 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. 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 the majority 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 rate 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 rate 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 rate 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

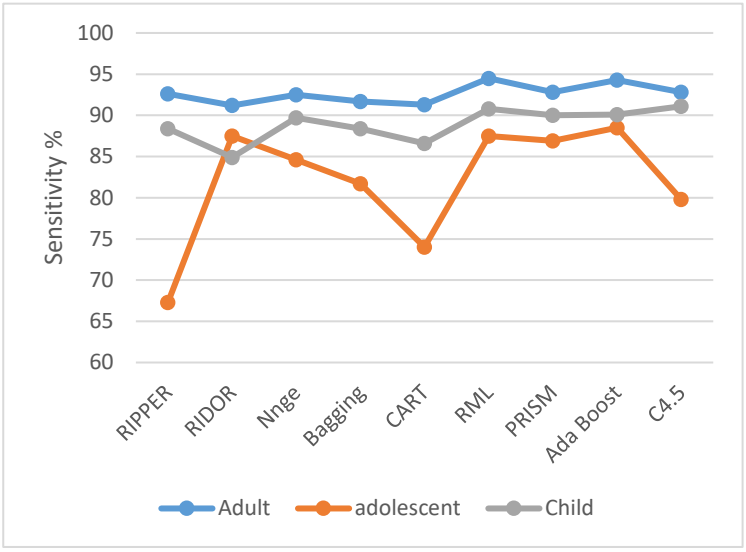


Fig. 4.9A Sensitivity rates of the ML algorithms on the Adult, Adolescent and Child datasets

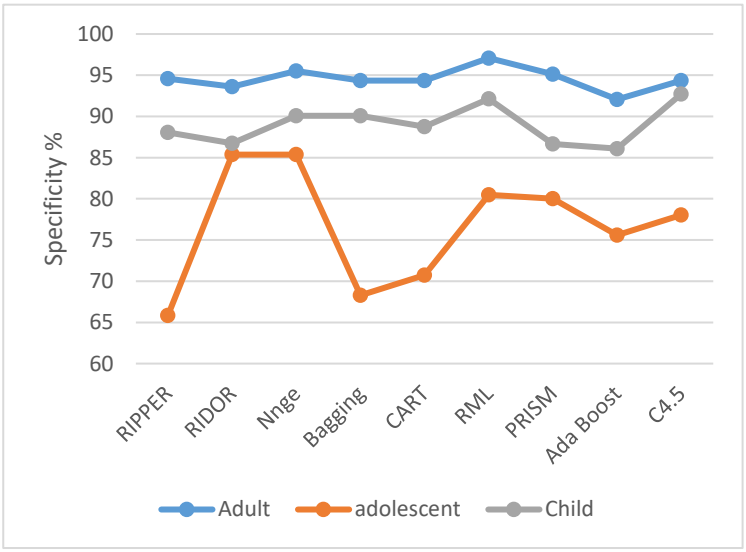


Fig. 4.9B Specificity rates of the ML algorithms on the Adult, Adolescent and Child datasets

unable to perform well in 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 outperform RML with respect to sensitivity rate on the Child dataset and by just 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 datasets when compared with the considered ML algorithms and these results are consistent with the error rate 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 has the largest number of false negatives, followed by CART and Bagging algorithms. Specifically, PRISM predicted 38 instances of individuals with No ASD traits that should have been classified with ASD traits. As a result, the sensitivity rate for class "ASD" for this algorithm was low at least on this dataset. On the other hand, PRISM has high specificity rate having only 12 false positives. In other words, PRISM only predicted 38 adults without ASD traits that potentially supposed to be on the spectrum, and 12 individuals with ASD that are supposed to be classified as No ASD. The "No ASD" class has higher data representation in the training dataset than ASD class, which means that the PRISM algorithm is sensitive to the number data linked to class labels.

Continuing, the RIDOR algorithm has the largest number of false positives, wrongly predicting 33 instances to be with ASD traits who are supposed to be without ASD traits. Overall, there was higher classification rates for class "No ASD" than "ASD" most of the all 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 supposed to be "No ASD" will have less misclassifications.

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 4.7). The F1 rates produced by the classifiers and shown in Figure 4.10 are high for Covering (RML) and Boosting algorithms. This indicates that RML and AdaBoost perform well in datasets with imbalanced class labels and higher than 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 works well in ASD detection. 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 RML. 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.

The researchers investigated the classifier content generated by the Covering, decision tree, Bagging and rule induction algorithms to seek important knowledge that can help in detecting ASD. Figure 4.11 shows the number of rules generated by the considered algorithms on the Adult

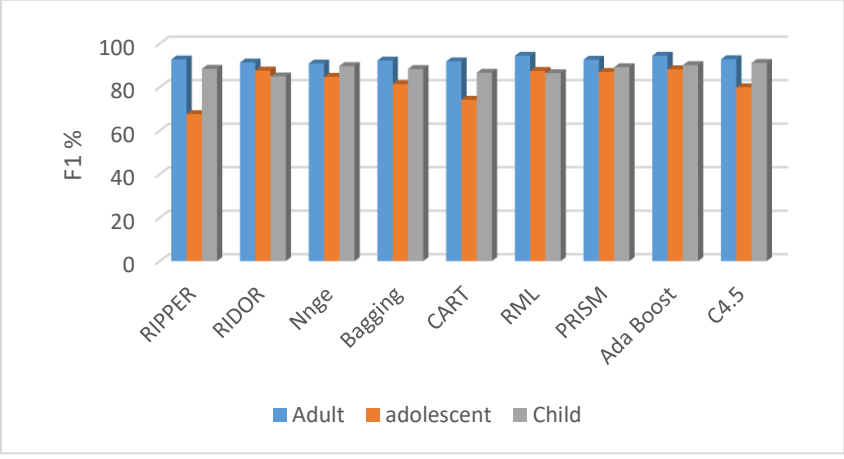


Fig.. 4.10 F1 rates in % derived by the considered ML algorithms on the Adult, Adolescent and Child datasets

dataset. The figures clearly show that PRISM and Nnge algorithms generate the highest number of rules. The reason for extracting too many rules by PRISM is that fact that this algorithm has no rule pruning strategy, so it keeps building up rules whereas Nnge is an algorithm that adopts nearest neighbour 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 rule induction and Covering approaches. The rule induction approach represented by RIPPER and RIDOR generate slightly smaller classifiers than Covering approach. This can be attributed to the rigorous pruning procedures adopted by RIPPER and RIDOR in evaluating the rules.

Table 4.4 contains the common rules related to ASD detection that have been extracted by 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 in particular items 5, 9, 8, and 4. Additionally, items 7 & 2 appeared in multiple rules in RIPPER and RML classifiers. It seems that the items that have been frequently appeared within the rules cover certain autistic behaviours within the DSM-5 manual. For instance, item 8 covers repetitive behaviour whereas item 4 is aligned with communication and lastly items 5 & 9 are aligned with social behaviour (Table 4.5).

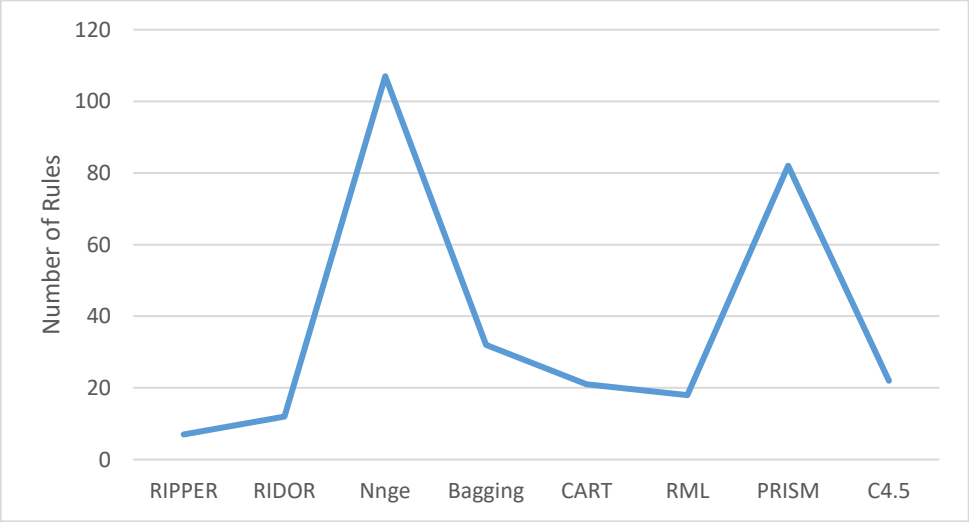


Fig. 4.11 Number of rules derived by the considered ML algorithms on the Adult dataset

Table. 4.4 Common rules derived by RML and RIPPER algorithms on the autism dataset

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_austim = 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_jundice = 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	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)

Table. 4.5 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

4.6 Chapter Summary

In this chapter, a comprehensive evaluations of the proposed ML model (RML) along with the proposed feature selection method (Va) were performed. The empirical evaluations were on three published autism datasets (Child, Adolescent, Adult) using various different ML algorithms including rule induction, Bagging, Boosting and decision trees. The result reported superiority of the proposed rule-based ML (RML) with respect to different evaluation metrics including specificity, sensitivity, harmonic mean and error rate. The results also showed that RML derived classifiers that contain useful rules for understanding the reasons behind the ASD classification and these rules can be effectively utilised for predicting autistic traits in children, adolescent and adults. Lastly, the feature selection results revealed influential items in the autism screening that

are aligned with social and communication behaviours yet not fully fulfilling DSM-5 criteria for ASD diagnosis.

In summary, this chapter clearly revealed that ML approaches especially rule-based ones showed promising results in detecting ASD cases. Next Chapter, conclusions and the study implications are highlighted.

Chapter Five[5]

Conclusions and Future Work

The accuracy and efficiency of ASD screening methods rely primarily on the experience and knowledge of the user as well as the items designed in the screening method. Unfortunately, some families and adult patients are unaware of ASD traits that may be exhibited and consequently do not seek necessary diagnostic services or contact their general practitioner (GP). Therefore, providing these families with a quick, accessible, and simple screening tool utilising the least set of items related to ASD may increase accessibility and early detection. In using ML techniques such as rule-based models, not only can users offer classification of cases and controls but also knowledge bases that can be utilised by domain experts and users in understanding the reasons behind the classification. This study investigated the applicability of rule-based classification in ML relative to ASD detection by identifying fewer, albeit influential, features in order to achieve efficient screening. Demands on evaluating the items' influences on ASD, within existing tools, is urgent.

This thesis proposed a new ML architecture for ASD classification that consists of an ASD screening mobile application, a novel feature selection method and more importantly a novel rule-based ML method for ASD traits detection. The proposed screening application was implemented and submitted to Google and Android stores and it can be seen as a time-efficient ASD screening tool to help health professionals and to inform individuals whether they should pursue formal clinical diagnosis. The new feature selection method is called Variable Analysis (Va) and it considers feature-to-class correlations and reduces feature-to-feature correlations, leading to the selection of only the highest predictive features. Rules-Machine Learning (RML) is the new classification method which uses knowledge base (rules) that are simple to interpret by novice users and clinicians. Both RML and Va methods have been implemented using Java programming language and integrated within the WEKA environment. Lastly, three new autism datasets related

[5] Parts of this chapter have been published in the Journal of Health Informatics and the Journal of Medical Informatics.

to children, adolescents and adult and which contain imperative features have been proposed using the screening ASD application. These datasets are rare and therefore enabled us and other scholars to perform thorough descriptive and predictive analyses in order to improve the efficiency, sensitivity, specificity and accuracy of the screening process for ASD.

Empirical results on the datasets related to adults, adolescent and children show that the RML offers improvement with respect to predictive power, sensitivity, and specificity than those of other ML approaches such as Boosting, Bagging, decision trees and rule induction. The results of the feature selection exhibited that Va was able to derive fewer features from adult, adolescent, and child screening methods yet maintained competitive performance. Overall, improvements related to ASD screening accessibility via a mobile environment has also been achieved. The development of a new ASD self-administered assessment tool may encourage a transition from antiquated clinical judgment tools and contribute to increased efficiency using professional diagnostic processes.

Future directions with clinical judgment tools may involve a semi-automated process, due to the need for licensed clinical specialists to verify outcomes (i.e., specific classifications). Assessment conclusions will be solely in the hands of the specialists, while ML will continue to improve predictive models and provide applicable alternatives to professionals. ML may provide assessors with potential rationales for classification decisions, improving the diagnostic process with respect to both efficiency and accuracy.

In the near future, we will investigate functions-based ML methods such as Artificial Neural Networks (ANN) for ASD screening. This is since ANN methods have shown competitive performance in other medical applications that involve classification of instances. In addition, ANN methods particularly deep learners are able to adjust the structure of the model on the fly during the training phase maximising the performance gained with respect to predictive accuracy. More importantly, deep learning method can deal with semi-structured and unstructured features such as images related to individuals' behaviours, eye movements, and audio/videos attributes among others thus these methods can positively impact the detection rate, and are more robust. We plane to propose a new deep learning method that integrates feature assessment and the training of instances in one step in order to improve the efficiency of the screening process. This method will be part of a larger architecture that consists of a mobile based interface connected to a secured

cloud database. The interface and the database communicate whenever a new user is undergoing screening, and the deep learning methods will utilise historical instances to derive models that are able to predict the outcome of the screening process.

One of the limitations of this research project is not having enough data instances related to infants for data processing. This will be vital especially that the detection of autistic traits is performed at very young ages and therefore detecting ASD in infants will increase the likelihood for them and their parents accessing the proper support services early. Moreover, extending the proposed ML architecture to consider diagnostic of autism will be advantageous as the classification models can be exploited by diagnosticians for better understanding the ASD influential traits besides empowering the human nature involved in the diagnostic decision. These models will not replace the diagnosticians rather they can be source of valuable rules to guide the formal diagnostic decision.

Another possible area of improvement of this study could be decoupling model and feature selection does not always work the best. However, that is not the claim of the thesis. The main proposition is that univariate – model independent - feature selection can be effective in certain situations. The effectiveness of feature selection is shown by our experimental results where we are able to reduce the number of features in the ML model without sacrificing accuracy. In general, the effectiveness of feature selection as a preprocessing step has been supported by various former studies. This possible area of improvement can be achieved when investigating deep learning models that integrate both feature selection and training in a single step.

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

[3] Abdelhamid N., Ayesh, A., Hadi W. (2014) MCAC: Multi-label Rules Generation via Parallel Associative Classification. Parallel Processing Letters journal. Vol 24-1 (1-24). WorldScinet.

[4] Abdelhamid N., Ayesh A., Thabtah F. (2013) Classification. Proceedings of the International conference on AI '2013, pp. 687-695. LV, USA. Associative Classification Mining for Website Phishing.

[5] Abdelhamid N., Ayesh A., Thabtah F. (2012) An Experimental Study of ThreeDifferent Rule Ranking Formulas in Associative Classification Mining. Proceedings of the *7th IEEE International Conference for Internet Technology and Secured Transactions* (ICITST-2012), pp. (795-800), UK.

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

[7] 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 [corrected]. Journal of American Academy of Child and Adolescent Psychiatry, 202-212.

[8] Allison, C., Baron-Cohen, S., Charman, T., Wheelwright, S., Richler, J., Pasco, G., & Brayne, C. (2008). The Q-CHAT (quantitative checklist for autism in toddlers): a normally distributed quantitative measure of autistic traits at 18–24 months of age: preliminary report. Journal of Autism and Developmental Disorders, 1414–1425.

[9] American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders. Washington.

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

[11] ASD Detect (2016). ASD Detect App. Retrieved from http://asdetect.org/ (accessed September 15, 2016).

[12] Asperger Test (2017) AQ Asperger Test. Retrieved from https://play.google.com/store/apps/details?id=com.kuruntham.aspergerstest&hl=en. (accessed September 12, 2017).

[13] Autism Test (2017) Autism Test App. Retrieved from https://play.google.com/store/apps/details?id=com.consurgo.autismtest&hl=en (accessed September 9, 2017).

[14] Autism and Beyond (2017) ITunes Preview. Retrieved from https://itunes.apple.com/us/app/autism-beyond/id1025327516?ls=1&mt=8# (accessed September 5, 2017).

[15] Auyeung, B. B.-C. (2008). The autism spectrum quotient: children's version (aq-child). Journal of Autism and Developmental Disorders, 38(7):1230-40.

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

[17] Baron-Cohen, S. (2001). Take the AQ test. Journal of Autism and developmental disorders, 5-17.

[18] Baron-Cohen, S., Allen, J., & Gillberg, C. (1992). Can autism be detected at 18 months? The needle, the haystack, and the CHAT. British Journal of Psychiatry, 161, 839-843.

[19] Bone, D., Goodwin, M. S., Black, M. P., Lee, C.-C., Audhkhasi, K., & Narayanan, S. (2016). Applying machine learning to facilitate autism diagnostics: pitfalls and promises. Journal of Autism and Developmental Disorders, 1121–1136.

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

[21] Bordin, I. A., Rocha, M. M., Paula, C. S., Teixeira, M. C., Achenbach, T. M., Rescorla, L. A., & Silvares, E. F. (2013). Child behaviour checklist (cbcl), youth self report (ysr) and

teacher's report form (trf): an overview of the development of the original and Brazilian versions. Cad. Saúde Pública, 13-28.

[22] Breiman L. (2001) Random forests. Mach. Learning, 45(1):5-32, 2001. 1300

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

[24] Brent Martin (1995). Instance-Based learning: Nearest Neighbor With Generalization. Hamilton, New Zealand.

[25] Bunker, R., and Thabtah, F. (2017). A machine learning framework for sport result prediction. In Applied Computing and Informatics. In press. Amsterdam: Elsevier.

[26] Campbell, H., Chambers, D., & Eaves, R. (2006). Criterion-related and construct validity of the pervasive developmental disorders rating scale and the autism behavior checklist. Psychology in the Schools, 311-321.

[27] Cendrowska, J. (1987) PRISM: An algorithm for inducing modular rules. International Journal of Man-Machine Studies, Vol.27, No.4, 349-370.

[28] Chen, Y.-W., Tanaka, S., Howlett, R. J., Howlett, J., & C. Jain, L. (2017). Innovation in Medicine and Healthcare 2017. 5th KES international Conference on innovations in Medicine and Heathcare.

[29] 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.

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

[31] Constantino, J. (2005). (SRS™) Social Responsiveness Scale. WPS.

[32] Constantino, J. (2012). (SRS™-2) Social Responsiveness Scale™, Second Edition. WPS.

[33] Constantino, J., & Gruber, C. (2014). Social Responsiveness Scale -2nd Ed (SRS-2). Journal of Psychoeducation Assessment.

[34] Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20* (3), 273 – 297.

[35] Couteur Le, A. R. (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, 659-685.

[36] Dietz, C., Swinkels, S., van Daalen, E., Engeland, H. v., & Buitelaar, J. K. (2006). Screening for autistic spectrum disorder in children aged 14–15 months. Journal of Autism and Developmental Disorders, 713–722.

[37] Duda, M., Kosmicki, J., & Wall, D. (2014). Testing the accuracy of an observation-based classifier for rapid detection of autism risk. Translational Psychiatry, 4(8):e424.

[38] 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.

[39] Ehlers, S., Gillberg , C., & Wing, L. (1999). Asperger Syndrome, Autism, and Attention Disorders: A Comparative Study of the Cognitive Profiles of 120 Children. Journal of Child Psychology & Psychiatry, 207-217.

[40] Ehlers, S., Gillberg, C., & Wing, L. (1999). A screening questionnaire for Asperger syndrome and other high-functioning autism spectrum disorders in school age children. Journal of Autism and Developmental Disorders, 129-141.

[41] Einfeld, S., & Tonge, B. (2007). Developmental Behaviour Checklist (DBC). Western Psychological Services (WPS).

[42] Fischbach, G., Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron 68, 192–195.

[43] Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian Network Classifiers. Machine Learning - Special issue on learning with probabilistic representations, 29(2-3), pp.131-63.

[44] Freund, Y. and Schapire, R.E., (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1), p.119–139.

[45] Frye, V. H., & Walker, K. C. (2008). Book review: autism screening instrument for educational planning, 3rd ed. (ASIEP-3). Journal of Psychoeducational Assessments, 200-285.

[46] Gaines B. R., Compton P. (1995). Induction of Ripple-Down Rules Applied to Modeling Large Databases. J. Intell. Inf. System. 5(3):211-228.

[47] Hadi W., Thabtah F., ALHawari S., Ababneh J. (2008) Naive Bayesian and K-Nearest

Neighbour to Categorize Arabic Text Data. Proceedings of the European Simulation and Modelling Conference. Le Havre, France,(pp. 196-200), 2008.

[48] Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing heterogeneous data: the national database for autism research. Neuroinformatics. 2012;10(10):331–39. doi:10.1007/s12021-012-9151-4. 810

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

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

[51] Holte, R.C. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning, 11, pp 63-90.

[52] lsabbagh, M. et al. (2012) *Global prevalence of autism and other pervasive developmental disorders*. *Autism Research.* **5**, 160–179 (2012).

[53] Gray, K., & Tonge, B. (2005). Screening for autism in infants and preschool children with developmental delay. Australian and New Zealand Journal of Psychiatry, 378-386.

[54] Gray, K., Tonge, B., Sweeney, D., & Einfeld, S. (2007). Screening for autism in young children with developmental delay: an evaluation of the developmental behaviour checklist: early screen. Journal of Autism and Developmental Disorders, 1003-1010.

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

[56] Grzadzinski, R, Huerta, M, Lord, C (2013) DSM-5 and autism spectrum disorders (ASDs): an opportunity for identifying ASD subtypes. Molecular Autism 4: 12.

[57] Greenhalgh, T. (1997). How to read a paper: papers that report diagnostic or screening tests. British Medical Journal, 540-543.

[58] Kent, R. G., Carrington, S. J., LeCouteur, A., Gould, J., Wing, L., Maljaars, J., et al. (2013). Diagnosing Autism Spectrum Disorder: who will get a DSM-5 diagnosis? *Journal of Child Psychology and Psychiatry, 54*(11), 1242–1250

[59] Kleina, T. J., Al-Ghasanib, T., Al-Ghasani, M., Akbarc, A., Tang, E., & Al-Faria, Y. (2015). A mobile application to screen for autism in Arabic-speaking communities in Oman. The Lancet Global Health, S15.

[60] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2 (12): 1137–1143. CiteSeerX 10.1.1.48.529Freely accessible.

[61] Krug, D. A., Arick, J. R., & Almond, P. J. (1980). Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. *Journal of Child Psychology and Psychiatry, 21*, 221–229.

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

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

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

[65] Lopez Marcano, J. L. (2016). Classification of ADHD and non-ADHD Using AR Models and Machine Learning Algorithms (Doctoral dissertation, Virginia Tech).

[66] Lord, C., & Luyster, R. (2006). Early diagnosis and screening of autism spectrum disorders. Medscape.

[67] Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., & Pickles, A. (2000). The Autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. Journal of Autism and Developmental Disorders, 205-223.

[68] 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, 659–685.

[69] Maenner MJ, Yeargin-Allsopp M, Van Naarden Braun K, Christensen DL, Schieve LA (2016) Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder. PLoS ONE 11(12):

[70] Matson, J., Hattier, M., & Williams, L. (2012). How does relaxing the algorithm for autism affect DSM-V. Journal of Autism and Developmental Disorders, 1549–1556.

[71] Mazefsky CA, McPartland JC, Gastgeb HZ, Minshew NJ. Brief report: comparability of *DSM-IV* and *DSM-5* ASD research samples. *J Autism Dev Disord*. 2013;43(5):1236-1242.

[72] 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.

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

[74] Mohammad R., Thabtah F., McCluskey TL (2013) Predicting Phishing Websites using Neural Network trained with Back-Propagation. World Congress in Computer Science, Computer Engineering, and Applied Computing. , Las Vegas, Nevada, USA, pp. 682-686. ISBN 1601322461.

[75] Mythili M, Shanavas Mohamed R. A study on Autism spectrum disorders using classification techniques. Ijcsit. 2014;5(6):7288–91.

[76] Oro, A. B., Navarro-Calvillo, M. E., & Esmer, C. (2014). Autistic behaviour checklist (abc) and its applications. Comprehensive Guide to Autism, 2787-2798.

[77] Pancers, K., and 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.

[78] Pennington, M. L., Cullinan, D., & Southern, L. B. (2014). Defining autism: variability in state education agency definitions of and evaluations for Autism Spectrum Disorders. Autism Research and Treatment, 1-8.

[79] Posserud, M., Lundervold, A., & Gillberg, C. (2009). Validation of the autism spectrum screening questionnaire in a total population sample. Journal of Autism and Developmental Disorders, 126–134.

[80] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies. 2 (1): 37–63.

[81] 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* Volume 6 (2), June 2014, 113-121.

[82] 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.

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

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

[85] Quinlan, J. (1979) Discovering rules from large collections of examples: a case study. In Expert Systems in the Micro-electronic Age. Edinburgh, 1979.

[86] Risi, S., lord, C., Gotham, K., Corsello, C., Chrysler, C., Szathmari, P., . . . Pickless, A. (2006). Combining Information From Multiple Sources in the Diagnosis of Autism Spectrum Disorders. Journal of the American Academy of Child & Adolescent Psychiatry, 1094-1103.

[87] Robins, D., Fein, D., Barton, M. & Green, J. (2001). The modified checklist for autism in toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. Journal of Autism and Developmental Disorders, 31(2), pp.131-144.

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

[89] Rutter, M., LeCouteur, A., & C. L. (2003). Autism Diagnostic Interview – Revised . WPS.

[90] Sappok, T., Diefenbachera, A., Budcziesb, J., Schadea, C., Grubicha, C., Bergmanna, T., . . . Dziobek, I. (2013). Diagnosing autism in a clinical sample of adults with intellectual disabilities: How useful are the ADOS and the ADI-R? Research in Developmental Disabilities, 1642-1655.

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

[92] Schopler, E., Van Bourgondien, M. E., Wellman, J., & Love, S. R. (1980). Toward objective classification of childhood autism: childhood autism rating scale (cars). Autism Dev Disord, 91–103.

[93] Scott, F., Baron-Cohen, S., Bolton, P., & Brayne, C. (2002). The CAST (childhood Asperger syndrome test)- preliminary development of a UK screen for mainstream primary-school-age children. Sage Journal, 9-31.

[94] Shannon, C. (1948). A Mathematical Theory of `Communication. *Bell Systems Technical Journal, 27*, July `and October, pp. 379–423 and 623–656.

[95] Soleimani, F., Khakshour, A., Abasi, Z., Khayat, S., Ghaemi, S. Z., & Golchin, N. A. (2014). Review of autism screening tests. International Journal of Pediatrics, 319-329.

[96] South, M., Williams, B. J., McMahon, W. M., Owley, T., Filipek, P. A., Shernoff, E., . . . Ozonoff, S. (2012). Utility of the Gilliam autism rating scale in research and clinical populations. Journal of Autism and Developmental Disorders, 593–599.

[97] Stewart, L. A., & Lee, L. C. (2017). Screening for autism spectrum disorder in low- and middle-income countries: A systematic review. Autism: The International Journal of Research and Practice, 21, 527–539.

[98] Stone, W. L., McMahon, C. R., & Henderson, L. M. (2008). Use of the screening tool for autism in two-year-olds (STAT) for children under 24 months. Sage Journal, 557-573.

[99] Towle, P., and Patrick, P. (2016). Autism Spectrum Disorder Screening Instruments for Very Young Children: A Systematic Review. New York: Hindawi Publishing Corporation.

[100] Thabtah, F. (2017A). 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: Digital Library.

[101] Thabtah, F. (2017B). ASDTests. Available: www.asdtests..com [accessed November 30th, 2017].

[102] Thabtah, F. (2018A) Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics for Health and Social Care 43 (2), 1-20.

[103] Thabtah F. (2018B) An Accessible and Efficient Autism Screening Method for Behavioural Data and Predictive Analyses. Health Informatics Journal. Health informatics journal, pp. 1-21, 1460458218796636 . 2018.

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

[105] Thabtah F., Hadi W., Abdelhamid N., Issa A. (2011) Prediction Phase in Associative Classification. Journal of Knowledge Engineering and Software Engineering. Volume: 21, Issue: 6(2011) pp. 855-876. WorldScinet.

[106] Thabtah F., Mahmood Q., McCluskey L., Abdel-jaber H (2010). A new Classification based on Association Algorithm. Journal of Information and Knowledge Management, Vol 9, No. 1, pp. 55-64. World Scientific.

[107] Thabtah F., Cowling P., and Peng Y. (2006): Multiple Label Classification Rules Approach. Journal of Knowledge and Information System. Volume 9:109-129. Springer.

[108] Thabtah, F., Cowling, P., and Peng, Y. (2004) MMAC: A new multi-class, multi-label associative classification approach. *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM '04*), 217-224.

[109] Ventola, P., Kleinman, J., Pandey, J., Barton, M., Allen, S., Green, J., . . . Fein, D. (2006). Agreement among four diagnostic instruments for autism spectrum disorders in toddlers. Journal of Autism and Developmental Disorders, 839-47.

[110] Vllasaliu, L., Jensen, K., Hoss, S., Landenberger, M., Menze, M., Schutz, M., . . . Freitag, C. M. (2016). Diagnostic instruments for autism spectrum disorder (ASD). Cochrane Database of Systematic Reviews, 1-27.

[111] Wall, D., Dally, R., Luyster, R., Jung, J., & DeLuca, T. (2012A). Use of artificial intelligence to shorten the behaviuoral diagnosis of autism. PloS one.

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

[113] Wilkinson, L. A. (2015). Best practice review: social responsiveness scale, 2nd Ed (SRS-2). Best Practice Autism.

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

[115] Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev*. 2015;57:328–349. [PubMed]

[116] Wong , V., Hui , L., Lee , W., Leung , L., Ho , P., Lau, W., . . . Chung, B. (2004). A modified screening tool for autism (Checklist for Autism in Toddlers [CHAT-23]) for Chinese children. Pediatrics, 166-76.

[117] World Health Organisation. (1992). ICD-10 Classifications of Mental and Behavioural Disorder: Clinical Descriptions and Diagnostic Guidelines. Geneva: World Health.

[118] Zwaigenbaum L, Bauman ML, Stone WL, et al (2015) Early identification of autism spectrum disorder: recommendations for practice and research. Pediatrics. 2015;136.