INFORMATICS FOR HEALTH & SOCIAL CARE
2017, VOL. 00, NO. 00, 1–20
https://doi.org/10.1080/17538157.2017.1399132

Taylor & Francis
Taylor & Francis Group

Machine learning in autistic spectrum disorder behavioral research: A review and ways forward

Fadi Thabtah

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

ABSTRACT

Autistic Spectrum Disorder (ASD) is a mental disorder that retards acquisition of linguistic, communication, cognitive, and social skills and abilities. Despite being diagnosed with ASD, some individuals exhibit outstanding scholastic, non-academic, and artistic capabilities, in such cases posing a challenging task for scientists to provide answers. In the last few years, ASD has been investigated by social and computational intelligence scientists utilizing advanced technologies such as machine learning to improve diagnostic timing, precision, and quality. Machine learning is a multidisciplinary research topic that employs intelligent techniques to discover useful concealed patterns, which are utilized in prediction to improve decision making. Machine learning techniques such as support vector machines, decision trees, logistic regressions, and others, have been applied to datasets related to autism in order to construct predictive models. These models claim to enhance the ability of clinicians to provide robust diagnoses and prognoses of ASD. However, studies concerning the use of machine learning in ASD diagnosis and treatment suffers from conceptual, implementation, and data issues such as the way diagnostic codes are used, the type of feature selection employed, the evaluation measures chosen, and class imbalances in data among others. A more serious claim in recent studies is their development of a new method for ASD diagnoses based on machine learning. This article critically analyses these recent investigative studies on autism, not only articulating the aforementioned issues in these studies but also recommending paths forward that enhance machine learning use in ASD with respect to conceptualization, implementation, and data. Future studies concerning machine learning in autism research are greatly benefitted by such proposals.

KEYWORDS

Autism spectrum disorder; artificial intelligence; classification; data analysis; predictive models; feature selection; machine learning

Introduction

Autism Spectrum Disorder (ASD) is a brain development disorder that limits certain communication and social behaviors from natural growth.[1] In fact, its causes have been linked to genetic and neurological factors. Notwithstanding its genetic roots, ASD is mainly diagnosed utilizing behavioral indicators such as social interaction, imaginative ability, repetitive behaviors, and communication among others.[2] Children with ASD encounter more serious early developmental difficulties compared to other infantile groups. Those behavioral challenges vary and include difficulties in responding to sensory information (hearing, smelling, tasting, etc.), lagging language acquisition and difficulties in communicating, impacting early learning, and causing a difficult time in interacting with others.[3] The study conducted by Wiggin et al.[3] found that 33% of children with difficulties other than ASD have some ASD symptoms while not meeting the full classification criteria.

CONTACT Fadi Thabtah f.thabtah2@hud.ac.uk

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/imif.

ASD is diagnosed clinically by evaluating three behavioral domains (American Psychiatric Association, 2000):

- Communication and language
- Reciprocal social interaction
- Restricted activities

There have been a number of clinical and non-clinical diagnosis methods for ASD. Examples of clinical diagnosis methods are Autism Diagnostic Observation Schedule-Revised (ADOS-R) and Autism Diagnostic Interview (ADI) among many others.[4,5] Further utilized were self-administrated or parent-based non-clinical methods such as the Autism Quotient Trait (AQ)[6] and Social Communication Questionnaire (SCQ).[7] Most existing clinical ASD diagnostic instruments have shown competitive performance, such as ADOS-R and ADI-R which have derived acceptable sensitivity, specificity, and validity results in several experimental research studies.[8–10] Yet the vast majority of those studies rely on handcrafted rules that employ mathematical summation formulas of scores to come up with the appropriate diagnosis. Therefore, they require extremely careful use and necessitate the availability of expert clinicians in addition to accuracy. More importantly, the majority of existing ASD diagnostic tools require substantial time to produce a complete diagnosis.

To improve the diagnosis process of ASD, researchers have recently started to adopt machine learning intelligent methods[11–16] 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.

Machine learning is a research field that integrates mathematics, artificial intelligence, search methods, and other sciences to derive accurate predictive models from datasets.[17] Machine learning methods such as neural network, support vector machine, decision trees, and rule based classifiers are automated tools which normally require minimal human involvement during data processing. Examples of software packages that have embedded machine learning methods are R,[18] Scikit-learn,[19] Statistics and machine learning MATLAB toolbox,[20] and WEKA[21] among others. Since the diagnosis process of a case involves coming up with the right class (ASD, No-ASD) or so called "Best estimate clinical" (BEC), based on input case features, then this process can be seen as a predictive task in machine learning. In other words, ASD diagnosis process is a typical classification problem where the clinician is trying to build an automated model (classifier) using machine learning to guess whether a case is ASD or not. This classifier is usually constructed from an input dataset (former cases who were with and without ASD classified by a typical diagnostic tool), and then evaluated on independent test instances (new cases) to measure its effectiveness in predicting the type of diagnosis. Overall, the diagnosis process in autism research is a classification task (See Figure 1 in Section 2) for further details.

This research article investigates recent studies on machine learning in clinical and non-clinical ASD. More importantly, we critically evaluate improvements in these studies related to

- Development of new machine learning methods for ASD classification
- ASD diagnostic time reduction and minimizing the number of features
- Improving classification, accuracy, specificity, and sensitivity
- Differentiating ASD from Attention Deficit Hyperactivity Disorder (ADHD) cases

Moreover, we highlight noticeable conceptual, implementation, and evaluation downsides along with result discrepancies within the recent studies. Findings of this article can be utilized by future studies in intelligent data analysis and the behavior sciences to understand the sensitivity of employing machine learning in diagnosing autism. More importantly, we present the necessary steps that must be taken into account when machine learning is adopted for constructing predictive ASD

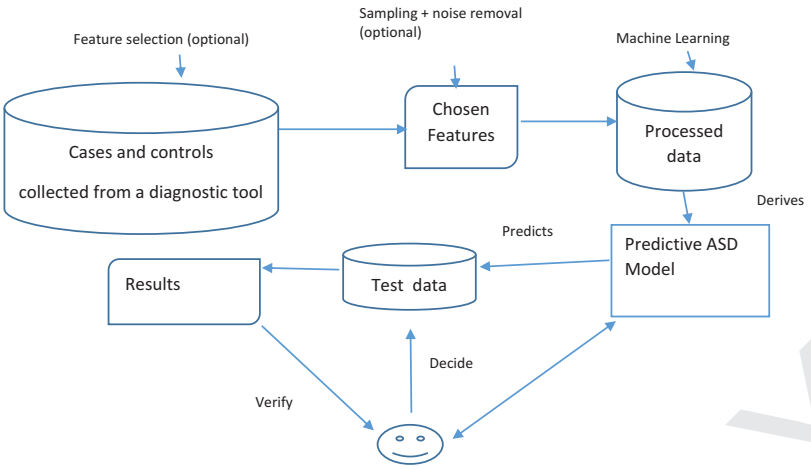

Figure 1. ASD diagnosis as a classification problem.

models. In particular, these include data pre-processing, algorithms integration within exiting ASD diagnostic tools, building and accessing a benchmarked data repository, examining ASD as a multi-level classification, establishing classifiers as rules set, balancing classes and sampling, and evaluating measures among others. Overall, the adoption of machine learning in autism research is promising yet the claim of proposing a new classification algorithm must be justified with appropriate steps that lead to on the fly diagnosis of test cases. This is unlike most studies which isolate the problem to just training from a dataset and coming up with a standalone predictive model that has yet to be integrated in an autism screening tool.

There have been a limited number of reviews on the utilization of machine learning in the study of autism.[9,13,14,22–25] Nevertheless, most reviews covered one or limited aspects of ASD (the role of classification during the diagnosis process, application of a certain machine learning class of algorithms, differentiating between features of autism and ADHD, identifying small sets of features to shorten diagnosis time among others). For instance, Pratap, et al.,[24], Wall et al.,[13] and Bone et al.,[16] and Lopez Marcano,[23] reviewed the applicability of different algorithms such as Neural Network and Tree models (decision tree, random forest) to shorten time of ASD diagnostic process. Bone et al.,[14] and Bone et al.,[22] synthesized and re-examined the results and the methodology of Wall et al.,[16] using the same algorithms. The authors highlighted a number of pitfalls including discarding hard to classify cases resulting in biased accuracy, absence of real clinical environment during the evaluation process, the use of incomplete ASD codes to build the classification system. Duda et al.,[13] and Chu et al.,[26] conducted an experimental review comparing six machine learning algorithms in order to determine the best fitting model for ADHD and ASD. Wolfers et al.,[25] reviewed common problems related to psychiatric disorders including small sample sizes, external validity and machine learning algorithmic challenges without a clear focus on autism. This research however, goes beyond most earlier studies by considering new issues related to application of machine learning to autism such as the importance of interpreting the classifiers derived by the machine learning algorithm, the availability of datasets, the characteristics of feature selection process, and detailed description of data selection, processing, analysis and interpretation. More importantly, we show what it takes to embed an intelligent classification algorithm within an existing diagnostic tool.

This article is structured such that section 2 critically analyses related works on machine learning that are linked with ASD, while section 3 presents the major issues related to machine learning, ASD studies, and ways forward to resolving them. Lastly, conclusions and further research are presented in section 4.

Machine learning studies on autistic spectrum disorder

ASD diagnosis as a classification problem

Figure 1 exhibits the ASD classification problem in machine learning, the input utilizing a training dataset of cases and controls that have already been diagnosed. Usually, the cases and controls have been generated using a screening tool such as ADOS-R, ADI-R, etc., in a clinic and administered by a behaviorist, clinical psychologist, or a licensed clinician specialized in that tool. Once the training dataset is identified, then an optional step to reduce data dimensionality by selecting a smaller set of features can be implemented. The aim of this step can vary and is not limited to simplifying the problem, identifying the best ASD features, reducing computing resources used during the data processing, etc. Noise removal is the next step, is also optional, and relies on factors such as the machine learning methods' tolerance to noise. Noise can be missing values, duplicate records, data balancing, etc. For instance, most of the current studies have employed sampling to improve data issues such as the diagnosis type (class) imbalance in the dataset.

Two common types of sampling can be performed in binary classification (two class problem): Over-sampling or under-sampling.[27] Over-sampling samples instances from the larger frequency class equal to those of the other class. While under-sampling samples instances from the minority class until both classes numbers of instances are somehow balanced.

Once pre-processing the data is completed a machine learning algorithm can be applied. Currently, researchers normally employ ready software packages such as R[18] and WEKA[21] to accomplish this task by loading the processed dataset and then choosing the machine learning algorithm. The outcome of this phase is different measures that the end user can use to evaluate the effectiveness of the chosen machine learning method on guessing the type of diagnosis. Examples of the evaluation measures are accuracy, processing time, false positive rates, false negative rates, true negative rate, etc. Often these evaluation measures are embedded within the machine learning software package.

The process of ASD diagnosis described in Figure 1 elaborates the necessary steps taken to decide the type of diagnosis used in machine learning. However, this process must be integrated into an existing screening tool in order to be utilized by the appropriate domain expert. This scenario is still under researched and has not yet been implemented. In other words, none of the existing ASD screening tools actually contain a machine learning diagnosis algorithm that experts are using to come up with the appropriate ASD diagnosis type. The classification process in machine learning should be automated rather than static, and without appending the machine learning algorithm in an existing screening tool the automation process is incomplete.[28]

Overall, below are the necessary requirements to perform diagnosis using machine learning:

(1) Input: A dataset with controls and cases
(2) Process: A machine learning algorithm embedded in a diagnostic tool to help build a predictive model for ASD classification
(3) Output: A predictive model that forecasts the diagnosis type for test data.
(4) A licensed expert to administer the process and verify the outcome of the predictive model and derive a final decision

The optional steps that may help in the outcome quality or improve process efficiency:

(1) Feature selection
(2) Noise minimization in the input dataset as well as other processes such as sampling for data balancing, missing values treatment, etc.

Table 1. Confusion matrix for ASD diagnosis problem.

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

Evaluation measures used in ASD problem with machine learning

Normally, the predictive model derived by machine learning is assessed with a number of evaluation measures. For a binary classification problem (ASD, No-ASD) as the basic form of the classification problem, Table 1 shows the possible answer for a test case prediction. Classification accuracy (Equation 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 3) identifies the ratio of the test cases that have truly ASD (true positive rate) whereas specificity (Equation 4) is the ratio of the test cases who do not have ASD (true negative rate).

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

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

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

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

ROC, an acronym for Receiver Operating Characteristic,[29] can be utilized as an evaluation measure for the machine learning models. ROC is a function that contrasts FPs (x-Axis) and TPs (Y-Axis) of the models derived graphically. Poorly performing models can be represented on ROC as equal points of FPs and TPs, respectively; whereas an FP of 0 and a TP of 1 is ideal performance on the ROC graph.

Another recently developed measure to overcome the class imbalance problem in data that was used in Duda et al.,[13] is called Unweighted Average Recall (UAR).[30] This evaluation measure considers the mean of sensitivity for one class and the mean specificity for the other class together. This method is normally exploited since it signifies the class labels regardless of the number of instances that they possess.

Cross-validation[17] is a testing method existed in machine learning to examine the predictive models and to produce its effectiveness. It 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.

Literature review

A well-known clinical diagnosis method in ASD research that has been widely used is ADOS-R.[5] This method relies on four different modules embedded within a computerized 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 behavioral 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.[5] For instance, module 1 is designed for cases that do not regularly utilize phrase speech such as young children. Wall et al.[15] and Wall et al.[16] 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[21] 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.[31] (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. 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.,[8] 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 utilize at most ten features/questions as a form for quick clinical referrals of potential ASD cases. The authors then analyzed 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 utilized 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 be 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 this study, mainly most cases being recruited during data collection and therefore are aware of ASD potentially creating biased results. More importantly, we believe that despite the promising results achieved by Allison, et al.,[8] 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 on 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.[12] 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.[12] 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.,[13] 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 minimize 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[32] based on the Social Responsiveness Scale (SRS) which is a parent administrated questionnaire that is often utilized 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,[11] 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, Behavior) 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 utilized 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 article of Mythili and Shanavas,[11] has insufficient data and no clear methodology or analysis, and therefore its results cannot be generalized. 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.[15]. Nevertheless, shortcomings related to result analyses have been highlighted with rationale in Bone et al.[14] 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[15]) have been identified by Bone et al.[13]

(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.,[33] 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. Behavioral attributes from ASD diagnostic tools were not utilized 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.,[24] on a dataset based on the Childhood Autism Rating Scale (CARS) diagnostic tool.[34] CARS is a behavioral rating scale usually employed for testing symptoms related to ASD. The authors aim was to measure the effectiveness of the probabilistic Naïve Bayes technique[35] as well as Artificial Neural Network based on a Self Organizing Map (SOM-ANN)[36] 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[31] the results of detecting ASD cases improve at least on the dataset used. A later study, Pratap and Kanimozhiselvi,[38] 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 utilized 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 centers.

Machine learning in ASD classification

Conceptual issues and suggestions

Is it really a diagnosis algorithm?

Most current studies claim derivation of a machine learning method for automatic ASD diagnosis whereas in most cases the researchers are merely adopting existing machine learning algorithms and applying them separately on autism datasets. All machine learning articles for autism have adopted existing software packages for predictive methods, like the ones in WEKA,[21] R,[18] and LIBSVM,[39] to use on an input dataset with ASD, ADHD and non-ASD cases and controls. Common algorithms employed in the training phase are Support Vector Machines such as in the studies of Bone, et al.,[22] Duda, et al.,[13] Kosmicki, et al.,[40] Logistic Regression and Decision Trees such as in Wall et al.[15] and Wall et al.[16] and SOM and Naïve Bayes such as in Pratap and Kanimozhiselvi,[38] and Pratap et al.,[24] The main aims were to enhance sensitivity, specificity, and classification accuracy besides differentiating between ASD and ADHD. Therefore, different versions of the input dataset (different features) are trained to maximize the aforementioned measures and the version that yields the best performance results is claimed to be the ASD predictive system. This means if a different dataset with little variation in features is used, a new possible system will be the outcome and thus the systems derived earlier are no longer valid. Thus, the ASD classification systems' predictive powers in all the current studies rely heavily on the input features besides sampling and feature selection methods.

The classification process within machine learning is automated and on the fly, not a standalone problem with a static training data that gets trained using machine learning algorithms. Rather, it is a complex dynamic process. We expect the process of medical diagnosis of ASD to be automated, and with the presence of appropriate medical staff as a decision maker inside a clinic. This necessitates:

(a) The classifier to be embedded within the medical screening tool
(b) The case under examination to be the test data case
(c) A knowledge base (classifier) that can be amended periodically based on the classified test cases and a training dataset that grows exponentially.
(d) The diagnosis to be taken within a valid environment

Unfortunately, none of the above conditions hold in these studies. The aforementioned studies deal with the problem of ASD classification from a static manner whereas existing machine learning algorithms are merely applied on an historical ASD dataset. Then, a classifier is built without measuring its impact inside the true diagnostic tool. In other words, the hand crafted rules inside the diagnostic tool must be replaced with the automated classifier to show the effectiveness of machine learning with respect to time, accuracy, sensitivity, and specificity. Hence these studies can be seen as promising research, but not complete classification systems for ASD or even diagnostic methods.

We believe that very limited examination of the machine learning perspective in autism research literature has been conducted, in particular, on the validation process of the "adopted" machine learning algorithms for ASD. Hereunder are the steps needed to fulfil a claim of developing an automated predictive ASD classification model:

(1) Input: A historical case and controls, a machine learning algorithm, a diagnostic tool (ADOS-R, ADI-R, etc), specialized licensed clinician.
(2) Process:
(3) Pre-processing the data (optional)
(4) Integrating the machine learning algorithm within the diagnostic tool
(5) Training on the data using the machine learning method
(6) Building the classifier which will be the predictive ASD model
(7) Replacing the original handcrafted rules of the diagnostic tool with the classifier
(8) Reading a new test case using the diagnostic tool
(9) Evaluating the new case using the classifier of the diagnostic tool
    (a) Generating the evaluation results and providing them to the licensed clinician
    (b) Diagnose the test case by the licensed clinician

Overall, current research studies that have been conducted have not yet integrated their machine learning models within standardized tools like ADOS-R, ADI-R, etc., so we cannot pragmatically rely on the classification power of these machine learning models. This is since the quality of the outcome and the decision made depend on the specialized licensed clinical staff that at this moment cannot be replaced with just an automated model.

Simplifying the problem to binary classification

In the majority of the research studies that employed machine learning in ASD diagnosis, the datasets used in the experiments contained only two class labels, e.g. ASD and Non-ASD. This makes the problem under consideration a binary classification problem, and ignores the fact that ASD can be decomposed into further classes like "Light Autism", "Severe Autism", etc. Researchers further simplify the autism classification problem by only considering clear cases of ASD and non-ASD in the datasets, ignoring cases that may have common features with other Pervasive Development Disorder (PDD) categories such as ADHD and Asperger Syndrome. This surely does not reflect the complexity of the problem which contains overlapping data cases among the class labels (diagnosis type), and the challenge is discovering the fine line that separates these cases. Existing studies such as Wall et al.[15] and Wall et al.[16] removed during the pre-processing phase any cases that are not clear ASD or non-ASD. These cases are hard to be determined because they may belong to multiple categories which can confuse the machine learning algorithm during detection, thereby increasing false positive results. This may also lead to generating biased classifiers that have unreliable performance in terms of sensitivity, specificity, and accuracy. This exercise is common at least among some of the existing studies involving the utilization of machine learning.

We recommend an in-depth analysis of the data before applying machine learning in which the dataset can be clustered using hierarchical clustering, which allows data that belongs to ASD to be decomposed into multiple hierarchies using divisive or agglomerative approaches. This enhances data presentation prior to the learning phase and could produce more advantageous results to the different stakeholders. These stakeholders are interested in further deconstructing the cases into sub classes rather than just two in order that more specific knowledge can be revealed. This may also improve the diagnosis process as well as the outcome of the classification which may associate the ASD case to the right medical assistant level.

Another possible recommendation is the use of predictive models that generate multi-label classifiers such as MMAC[41] or lazy classifiers.[42] These approaches allow generating extra knowledge, thereby minimizing the amount of tossed knowledge during the training phase on the dataset. Now,

since cases and controls in ASD may share common characteristics and there is overlapping among the different PDD types in the data samples the suggested multi-label approaches may bring up the additional knowledge representing the overlapping data samples in the diagnosis types as rules with multiple labels. However, applying a multi-label algorithm requires careful data transformation.

Reduction of diagnostic time

Another serious problem is that some of the current machine learning studies such as Duda et al.,[13] claimed that machine learning models they evaluated reduced the time of ASD pre-diagnosis or even in some cases like Wall et al.[16], the actual clinical diagnosis. The reduction of time strategy in these studies is carefully selecting the features of the input dataset before applying a machine learning algorithm or looking into the features that are present in the predictive models. These features are usually based on existing diagnostic tools like ADOS-R, ADI-R and others. Since these features can be large, researchers employed feature selection based on either fileting or wrapping methods[43] so they can only keep effective features that have direct influence on ASD classes. Hence the hypothesis put forward by scholars in the above studies is to reduce the input data dimensionality and minimize the time taken to build the model, thus diagnosis efficiency with respect to time improves.

The approach toward selecting effective features is promising when the aim of the study is to identify smaller sets of influential features. However, if the aim is to reduce time of diagnosis of an ASD tool, all features of the current tools must be collected first by a licensed clinician and therefore no reduction of time is possible during the diagnosis process. The claims of Duda et al.,[13] and Hall, et al.,[44] can only be valid if the actual diagnostic tool is adjusted to accept only 6, 8 or 10 features, which means in practice proposing a completely new ASD diagnostic tool. However, this proposal requires major comprehensive research, testing and evaluation from behavioral scientists and clinical psychologists before it can be deemed a success.

A direction toward research surrounding the ranking and grouping of features related to diagnostic tools is promising. However, the claim of reducing the time needed for the diagnoses is still yet to be achieved since the reliability and validity of using a reduced set of features in existing diagnostic tool has not yet been tested or even implemented. The current practice of using feature selection methods to reduce dimensionality of a standalone ASD dataset and then applying machine learning methods on the selected features does not necessarily mean a reduction in the ASD diagnoses time. A more serious approach will be investigating the influence of the data reduction during actual diagnoses and by a fully licensed clinician compared with the results of those derived when the complete features are utilized.

Understandable machine learning classifiers is imminent

Most of the current studies on ASD using machine learning have adopted existing algorithms to enhance certain evaluation measures such as AUC, classification accuracy, time needed to construct the model, and others. Examples of these studies are

- Duda, et al.,[13] which adopted six machine learning algorithms and applied them on a dataset of 2925 cases from SSC to distinguish ASD cases from ADHD cases. The algorithms employed are Support Vector Machine, two Logistic Regression methods, Categorical Lasso, C4.5, and Random Forest.
- Wall et al.[15] and Wall et al.[16] applied fifteen algorithms, mainly decision tree based, against a dataset based on the ADI-R tool from AGRE repository.
- Pratap and Kanimozhiselvi,[38] and Pratap et al.,[24] applied a number of supervised and unsupervised machine learning methods (Naïve Bayes, SOM, K-Means, etc.) on a 100 instances collected using CARS diagnostic tool.

The aims of the above mentioned studies were either faster screening by minimizing features used by the machine learning algorithm or improving classification accuracy, specificity, or sensitivity. None of the above approaches paid high attention to thoroughly analyze classifiers derived to find sharp lines, for instance between ADHD and ASD cases as in Duda et al.[13] In fact, researchers tried to select the algorithm that maximizes accuracy and using the smallest needed features. Hence the main criteria used is to maximize certain statistical measures without paying high attention to the knowledge derived and whether such knowledge is advantageous for the decision maker and other stakeholders in understanding autism and its diagnostic process.

Covering and rule based classifications can be seen more advantageous to ASD research for the below reasons:

(1) Their classifier contains easy If-Then rules that different stakeholders, including the clinicians, parents, care givers, etc., can easily interpret and exploit. In fact, this rule set can serve as a decision making tool in clinics. This is unlike the complex outcomes of current machine learning methods used in ASD such as Support Vector Machine, Logistic Regression, and decision trees.
(2) The classifiers derived are smaller in size with respect to the number of rules and hence it will be easy to control and maintain by the end users. End-user can amend the classifier by adding human knowledge without drastically changing the content of the classifier. This is unlike other machine learning methods such as decision trees or random forest that require reshaping the entire classifier in case of any addition or removal of knowledge.
(3) Each rule represents a correlation between a number of features values and class value, so a better understanding of the features and their impact on the ASD class is obtained.
(4) Rule based classifiers have shown promising results in other domain applications such as medical diagnoses, web security, stock market, and credit card scoring among others.

Lastly, results represented as rules can help specialized and computationally literate clinical staff understand the nature of the autism classification problem in order to come up with potential domain specific machine learning algorithms in near future.

Data and features issues and suggestions

The unavailability of benchmarked proceeded data repository

One noticeable challenge associated with autism behavioral data analysis research is the unavailability of processed benchmarked datasets that researchers have "agreed" on for the use of machine learning. Currently, there are multiple sources of autism classification datasets, particularly Boston Autism Consortium (AC),[32] AGRE,[31] National Database of Autism Research (NDAR) [38] and Simon Varian in Individuals Project (SVIP) (Simons VIP, 2012). Regrettably, the majority of the current studies conducted on machine learning using the aforementioned datasets vary in the ways researchers have processed the initial dataset under investigation. In particular, researchers such as Bone et al.,[22] Duda, et al.,[13] Kosmicki, et al.,[40] Bone et al.,[14] Wall et al.[15], and Wall et al.[16] have all employed different "versions" of the original datasets that may vary in the following characteristics:

- Size (number of cases)
- Features selection
- The actual diagnostic tool that produced the codes
- Features in the dataset
- Sampling and bootstrapping
- Cross validation (shown as CV in Table 2)

- Aim of the test: simple binary classification (ASD, non-ASD)[15], multi-class classification (Autism, Autism Spectrum, No autism)[16], differentiating ASD from ADHD [13], etc.
- Pre-processing settings of the original data. Some researchers, i.e. Wall, et al.[15], Wall, et al.[16], decided to merely remove the "Autism Spectrum" class. Others like Duda, et al.[13], have decided to only keep instances that have less than five missing answers. All other instances have been discarded before training on the predictive model.

By not agreeing on a universal autism behavioral data for machine learning, discrepancies among researcher results may exist even for the same dataset, which complicates other researchers' ability to validate them. This case was clear in particular between two research groups, Duda et al. (2015), Wall et al.[15], and Wall et al.[16], in which Duda et al. reproduced results that are totally different than Wall, et al. using the same dataset. Certain steps taken by Wall, et al.[15] in preparing the data and analysing the outcome after applying the machine learning algorithms are delegitimized by this fact. There are urgent needs for:

(a) Developing a behavioral data repository for ASD classification for the use of machine learning researchers. These datasets can be accessed by scholars who are interested in conducting research on ASD classification and hence quicker progress can be achieved. All relevant anonymity and ethical consideration procedures can be set prior to accessing the dataset.
(b) Sharing datasets among scholars working on ASD from both machine learning and behavioral science fields. We were unable to obtain processed data from scholars who published in ASD classification so we can reproduce results for validation and verification. Ethics and confidentiality agreements prevent the accessing of data, even if just for the purpose of replicating results.

Overall, in ASD behavioral research, there is no single dataset that researchers can exploit for machine learning data analysis. Extensive pre-processing is required since making the data ready for mining requires multiple necessary steps such as feature selection, ranking, noise elimination, cross validation, and sampling. Therefore, most of the current studies have processed different versions of the same dataset, which will make it hard for others to reproduce the results for validation and reliability reasons. One may claim that ASD data is sensitive and granting access should go through a screening process or data ownership. We are not objecting to either case, though we believe that for the validity of any published results on machine learning for behavioral science the authors should make their data available by publishing its revised version(s) in data prospective source webpages so other researchers can use, confirm, and pursue new research directions.

Imbalanced datasets

Most of the datasets employed in deriving evaluation measures of ASD (sensitivity, specificity, error rate, AUC, etc.) are imbalanced with respect to class labels. According to Table 2 almost all studies on machine learning for ASD classification have used imbalanced datasets. This limits any statistical power and may demand extensive pre-processing to ensure the reliability and validity factors of the resulting figures later on. The imbalanced issue is attributed again to the limited data resources and costs associated with recruitment of ASD cases in this application.

There are attempts to minimize this issue by either replicating the original minority class instances (non-ASD data) or to removing data instances that belong to the majority class label (ASD). The former strategy is called over-sampling while the latter strategy is referred to as under-sampling. Over-sampling methods were employed in Wall et al.,[16] and Wall et al.,[16] to balance the class in the dataset. This balancing approach is normally criticized by scholars of being time consuming, and demanding high computing resources besides potentially overfitting the training

Table 2. Sample of studies on the use of machine learning for ASD classification in behavior science.

Machine Learning Automated Methods													
Year	Diagnostic Tool	CV	Feature selection	Machine Learning Mehtods used	Software	Data size and source	Balancing	# of features used to derive best results	best algorithm	Specificity %	Sensitivity %	Accuracy%	Publication Details
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%		[7]
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	ADTree	94%	100%	99.70%	[15]
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	ADTree	93.8%-99%	100%	100%	[16]
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	SOM and Naïve Bayes			100%	[21]
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)	[34]
2016	SRS	Yes	Yes	6 (SVM, LR, DT)	Scikit-learn	65 attributes, instances: 2775 ASD, 150 ADHD, sources: Boston AC, SSC	backward and undersampling	5	SVM				[12]
2016	ADI-R and SRS	Yes	Yes	SVM	LibSVM	65 attributes, instances: 1264 ASD, 462 non-ASD sources: Boston BAC	stepwise forward	5	SVM	56.20%	87.95%		[19]

dataset. The authors of Wall, et al.[16] and Wall, et al.[16] replicated instances of non-ASD minority classes using under-sampling techniques to balance out the original input dataset.

Recently, a study with 1:3 ratios between ASD and non-ASD instances was conducted by Bone, et al.[22], and was one of the serious attempts at balancing class labels in the input dataset. Moreover, Duda et al.,[13] tried to avoid class imbalance in the dataset by using stratified cross validation and repeated under sampling. This allowed the model to learn from 90% of the training dataset, and test the remaining 10% repeatedly, over ten different times. In under-sampling the authors performed arbitrary ten samples on the majority class in the training and test datasets so they can end up with 1.5:1 ASD and ADHD in each sample set. Often under-sampling methods remove data instances from the majority class (ASD) to trade off the frequency distribution of the input dataset with respect to classes. However, as this sampling strategy discards "real" data instances for the majority class, the resulting predictive models may provide less useful knowledge that is vital for prediction as well as decision making.

Overall, given the sensitivity of autism application and its effect on human welfare, we believe that intelligent and domain specific sampling methods are necessary to guarantee reliable results that decrease the loss in statistical power by generalizing the models' performance. Moreover, over-sampling seems more applicable to medical datasets with imbalance problems since the original dataset will not loss any instances given the sensitivity of application. Empowering over-sampling and under-sampling methods with an initial clustering phase though seems a way forward taken the fact that the process of data replication or data removal will be intelligent rather pure random. Hence, approaches such as Zhang, et al.,[45] and Yen and Lee,[46] seem promising. For example, regarding under-sampling the input dataset is grouped into N clusters. The ratio of the majority class instances to the minority class instances is then taken to select instances that can be used for the training phase from different groups based on the computed ratio.

Furthermore, under-sampling approaches that utilize k-nearest neighbors from supervised learning are also appropriate for reducing randomization in the process of sampling, that is Chyi.[47] In these approaches, the instances chosen for the majority class are obtained from different subsets of data based on a distance function metric.

Another common issue that could arise from processing imbalanced datasets is the disagreement on the evaluation measures used beyond the point of building the ASD predictive model. For instance, Kosmicki, et al.,[40] has primarily adopted classification accuracy to measure the integrity of the classifiers derived whereas Duda, et al.,[13] noted that accuracy measure is not appropriate for measuring classifiers generated from imbalanced datasets and hence UAR which integrates ASD recall and non-ASD recall is a more appropriate measure.

Lastly, there are attempts to ingrate datasets from multiple sources in order to generalize the ASD predictive models' performance and reduce class label imbalance data. Nevertheless, removing

feature similarity before integrating cases and controls from multiple sources is a must. As we know some codes in the ASD diagnostic tools are similar, such as the modules in ADOS-R. Therefore, a better approach before data integration is to remove high similarity among features (codes) so the new integrated features will be dissimilar from each other and thus the correlation, "goodness," of these features can be explicitly measured with the class (diagnosis type) during feature selection or even the diagnosis process.

Feature selection

Feature selection plays a significant role in the success or failure of the machine learning algorithms.[48] Now, most of the existing ASD diagnostic tools are associated with a different number of features grouped on behavioral or social impairment characteristics. For instance, Autism Quotient (AQ) which is a measure for autistic traits for adults with autism is linked with fifty features, the SCQ has forty items and ADI-R contains ninety-three items. Current studies on machine learning for ASD have utilized one or more of the above mentioned diagnostic tools to derive predictive models. The ultimate aim of employing feature selection is to reduce the dimensionality of the input dataset and select the highest ranked features according to a specific threshold. Studies like Duda, et al.,[13] applied the minimal

redundancy maximum relevance (mRMR)[49] filtering method to rank features and Kosmicki, et al.,[34] adopted the stepwise back word selection method.[21]

Normally, two types of feature selection exist in the literature; one that sorts features based on a threshold and only derives features fulfilling the threshold, and another that extensively adds features and tests their impact on enhancing the accuracy of the predictive model. The former feature selection type is called filtering and the latter one wrapping.[43] A typical filtering method like Chi-Square[50] usually produces a ranked set of features quickly and hence filtering methods are time efficient. On the other hand, wrapping methods normally continuously tests each features' effect before deciding to add that feature, and thus this type of feature selection is resource demanding and computationally expensive. The majority of existing studies on machine learning and ASD utilize wrapping methods.

Unfortunately feature selection methods normally deviate in their results and a feature with a higher rank by applying one method has a lower rank in another filtering method. For instance, if we run two common filtering methods, Information Gain[51] or Chi Square,[50] on the "Labor" and "Hepatitis" datasets from the University of Irvine data collection,[52] we will end up with different chosen features. In particular, IG selects 14 and 17 variables from "Labor" and "Hepatitis" datasets respectively, using the predefined threshold of 0.01. Chi Square, however, selects three and nine features respectively from the same datasets using a predefined threshold of 10.83. Moreover, the feature selection results also may vary more significantly if the user has decided to use different thresholds other than the default. Hence, it is essential to come up with a global rank per feature. This example, although limited, illustrates high disparities in results obtained by applying feature selection methods. Therefore a comprehensive method that may reduce this discrepancy is needed. This method will stabilize the score per feature and increase confidence in its rank since the new score is obtained from multiple disparate methods.

To reduce the risks of results deviations, a filtering method that maintains each feature score and gives it a true rank is essential since it minimizes the instability in the scores without losing overall sensitivity, accuracy, and specificity of the expected predictive model. Good examples of such filtering methods are Rogati and Yang,[53] Uncu and Türkşen,[54] and Tsai and Hsiao.[55] Finally, it is worth mentioning that machine learning studies in ASD normally choose the features using extensive filtering (wrapping). This means when adding a feature to the chosen feature set, and accuracy decreases, they stop the filtering process. However, this stopping criterion varies from one filtering method to another and thus the chosen features will only perform well on the dataset employed during the experiments and may perform poorly when another independent dataset is used. Therefore, generalizing the performance of the resulting model cannot be guaranteed.

Real data collection in clinical environment

Though traditional clipboards and data sheets are still commonly used to record data in clinical environments, registering data as accurately, reliably and validly as possible in real time is essential for the use of algorithms in diagnosing psychiatric disorders such as autism. Real time data can only occur through appropriate use of technology. This allows for the easy storage, pre-processing and sharing of data for future use and live data transfers. Simultaneously, clinical psychologists or behavioral clinicians could collect ASD data more spontaneously and meticulously with technology since children, adolescents and adults with autism react to changes in their environments swiftly. The use of technological methods and portable devices such as tablets, Personal Digital Assistants (PDAs) or even smartphones could make the clinical data collection easier, in real time. For instance, a child could be undergoing an autism screening during which the clinician observes hearing or visual signs of unpredictable behavior indicative of autistic traits. In such cases, jotting down notes or ticking "yes" to a question in sheets is replaced with simple video capturing of the child in that particular moment could accommodate the clinical data collection for a particular question.[56,57] It is imperative to allow for live data collection and within the clinical environment using more recent innovative data collection and storage techniques so

A) More accurate data of several types are recorded and stored in a secure place

B) The machine learning algorithm will be able on the fly to perform data processing for classifying test cases and the test cases once classified become part of the historical dataset and can be used later during the training phase

For a conventional diagnostic instrument that utilizes machine learning approach to be effective, the aforementioned one or more of the available innovative data collection methods should be in use. However, since machine learning has not yet been embedded within a conventional screening method researcher in behavioral science and computational intelligence must consider the use of recent technological developments in data storage including both hardware, software and even cloud technologies while maintaining the specificity, security and sensitivity of the clinical setting of the autism diagnosis process.

Conclusions

One of the vital issues in autism research today is improvement of diagnosis type performance in existing diagnostic tools so that individuals can have more specific, improved, and faster service as early as possible. This can be accomplished in many ways, such as efficiently reducing the diagnosis time or increasing predictive accuracy of the diagnosis without compromising the validity or sensitivity of the test. In the last few years, scholars in behavior science started to adopt automated intelligent methods based around machine learning to accomplish the aforementioned goals.[13,14] The application of machine learning in the discovery and diagnosis of ASD and other Syndromes such as ADHD and Asperger among others is nascent.

Machine learning methods have showed promising results in varying applications aside from ASD research. In particular, recent research studies in machine learning claimed one or more of the following

- New ways to diagnosis cases related to ASD
- Shortening time associated with the diagnosis process of ASD, or at least self-administered autism trait tests
- Reducing the features associated with existing ASD tools without hindering sensitivity, specificity, or the accuracy of the test
- Identifying the best ranked features that influence ASD
- Determining overlapped features among different types of PDD

However, current studies that applied machine learning in ASD research have not considered conceptual, implementation, evaluation, and data related issues. These issues are not limited to ways diagnostic tool features/codes are used, but include data cases that overlap in features, processing of noise, feature extraction and selection, measures employed during evaluation, and training dataset imbalances with respect to diagnosis type among others. More importantly, recent studies have not yet integrated the machine learning algorithm inside an existing ASD screening tool which makes machine learning adaptability to ASD incomplete. This article has examined and critically analyzed these recent machine learning studies on ASD and the aforementioned conceptualization, implementation, and data related aspects. For each investigated aspect, we recommended possible ways to enhance and validate machine learning use in ASD so that future research studies can take these conceptualizations, implementations, and data issues into consideration. This not only pinpoints these challenges but also serves as potential research directions concerning machine learning applicability as a predictive model in ASD research. Based on the diagnostic algorithms we believe that SVM is the most frequently utilized algorithm to derive ASD classification models due to its high predictive power during the learning phase. When integrated into an existing diagnostic tool, these predictive models particularly SVM can greatly serve experts and researchers as effective decision making tools.

References

1. Bolton P, Macdonald H, Pickles A, Rios P, Goode S, Crowson M, Bailey A, Rutter M. A case-control family history study of autism. J Psychol Psychiatry. 1994;35(35):877–900. doi:10.1111/jcpp.1994.35.issue-5.
3. Wiggins LD, Reynolds A, Rice CE, Moody EJ, Bernal P, Blaskey L, Rosenberg SA, Lee LC, Levy SE. Using standardized diagnostic instruments to classify children with autism in the study to explore early development. J Autism Dev Disord. 2014a;45(5):1271–80. doi:10.1007/s10803-014-2287-3.
2. Chakrabarti, B, Dudbridge, F, Kent, L, Wheelwright S, Hill-Cawthorne G, Allison C, Banerjee-Basu S, Baron-Cohen S, et al. Genes related to sex steroids, neural growth, and social emotional behavior are associated with autistic traits, empathy, and Asperger syndrome. Autism Res. 2009;2(3):157–77. doi:10.1002/aur.80.
4. Lord C, Rutter M, Le Couteur A. Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord. 1994;24(24):659–85. doi:10.1007/BF02172145.
5. Lord C, Risi S, Lambrecht L, Cook EH Jr, Lambrecht BL, DiLavore PC, Pickles A, Rutter M. et al. the autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord. 2000;30(30):205–23. doi:10.1023/A:1005592401947.
6. Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. J Autism Dev Disord. 2001;31(31):5–17. doi:10.1023/A:1005653411471.
7. Berument SK, Rutter M, Lord C, Pickles A, Bailey A, MRC Child Psychiatry Unit, & Social, Genetic and Developmental Psychiatry Research Centre. Autism screening questionnaire: diagnostic validity. Br J Psychiatry. 2000;(175):444–51.
8. Allison C, Auyeung B, Baron-Cohen S. Toward brief "red flags" for autism screening: the short Autism Spectrum Quotient and the short quantitative checklist for autism in toddlers in 1,000 cases and 3,000 controls. J Am Acad Adolesc Psychiatry. 2012;51(2):202–12Y. doi:10.1016/j.jaac.2011.11.003.
9. Ruzich E, Allison C, Smith P, Watson P, Auyeung B, Ring H, Baron-Cohen S. Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. Mol Autism. 2015;6(2):1–12.
10. Mayer JL, Heaton PF. Age and sensory processing abnormalities predict declines in encoding and recall of temporally manipulated speech in high-functioning adults with ASD. Autism Res. 2014;7(7):40–49. doi:10.1002/aur.2014.7.issue-1.
11. Mythili M, Shanavas Mohamed R. A study on Autism spectrum disorders using classification techniques. Ijcsit. 2014;5(6):7288–91.
12. Pancers K, Derkacz A (2015). Consistency-based pre-processing for classification of data coming from evaluation sheets of subjects with ASDS. *Federated conference on Computer Science and Information Systems*, 63–67.
13. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiatry. 2016;9(6):732. doi:10.1038/tp.2015.221.
14. Bone D, Goodwin MS, Black MP, Lee -C-C, Audhkhasi K, Narayanan S. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. J Autism Dev Disord. 2014;45(5):1–16.
15. Wall DP, Kosmiscki J, Deluca TF, Harstad L, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of Autism. Transl Psychiatry. 2012;2(2):e100. doi:10.1038/tp.2012.10.
16. Wall DP, Dally R, Luyster R, Jung JY, Deluca TF. Use of artificial intelligence to shorten the behavioural diagnosis of autism. PlOs One. 2012;7:e43855. doi:10.1371/journal.pone.0043855.
17. Abdelhamid N, Thabtah F. Associative classification approaches: review and comparison. J Inf Knowledge Manag (JIKM). 2014;13(3).
18. R Development Core Team. *R*: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2008.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
20. MATLAB and statistics toolbox release. Natick, Massachusetts, United States: The MathWorks, Inc; 2012.
21. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The WEKA data mining software: an update. SIGKDD Explorations. 2009;11(1):10. doi:10.1145/1656274.
22. Bone D, Bishop S, Black MP, Goodwin MS, Lord C, Narayanan SS. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. J Psychol Psychiatry. 2016;57(57):927–37. doi:10.1111/jcpp.2016.57.issue-8.
23. Lopez Marcano JL (2016). Classification of ADHD and non-ADHD Using AR Models and Machine Learning Algorithms . (Doctoral dissertation), Virginia Tech.
24. Pratap A, Kanimozhiselvi CS, Vijayakumar R, Pramod KV. Predictive assessment of autism using unsupervised machine learning models. Int J Advan Intel Para. 2014;6(2):113–21. June 2014. doi:10.1504/IJAIP.2014.062174.

25. 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;57:328–49. [PubMed]. doi:10.1016/j.neubiorev.2015.08.001.
26. Chu KC, Huang HJ, Huang YS (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. 1044–49). IEEE.
27. Chawla NV. Data mining for imbalanced datasets: an overview. In: Maimon O, Rokach L, eds. Data mining and knowledge discovery handbook. New York City, NY: Springer; 2010. p. 875–86. ISBN 978-0-387-09823-4. doi:10.1007/978-0-387-09823-4_45.
28. Thabtah F. Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, Taichung City, Taiwan, May 20–22, 2017, pp. 1–6. ACM.
29. Green DM, Swets JA. Signal detection theory and psychophysics. New York, NY: John Wiley and Sons Inc; 1966. ISBN 0-471-32420-5.
30. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2005.
31. Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P, Ducat L, Spence SJ, AGRE Steering Committee. The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. Am J Hum Genet. 2001;69(69):463–66. doi:10.1086/321292.
32. Fischbach GD, Lord C. The simons simplex collection: a resource for identification of autism genetic risk factors. Neuron. 2010;(68):192–95.
33. Tenev A, Markovska-Simoska S, Kocarev L, Pop-Jordanov J, Müller A, Candrian G. Machine learning approach for classification of ADHD adults. Int J Psychophysiology. 2014;93(1):162–66. doi:10.1016/j.ijpsycho.2013.01.008.
34. Schopler E, Reichler R, DeVellis R. Toward objective classification of childhood autism: childhood autism rating scale (CARS). J Autism Dev Disord. 1980;10:91–103. doi:10.1007/BF02408436.
35. Duda, R. O. and Hart P. E. (1973). Pattern Classification and Scene Analysis. New York: John Wiley & Sons.
36. Kohonen T (1989). A self-learning musical grammar, or 'associative memory of the second kind'. International Joint Conference On Neural Networks. Washington, DC. 1: 1–5.
37. MacQueen JB (1967). Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–97. MR 0214227. Zbl 0214.46201. Retrieved 2009- 04-07.
38. Pratap A, Kanimozhiselvi C. Soft computing models for the predictive grading of childhood Autism—a comparative study. Int J Soft Comput Eng (IJSCE). 2014;4(3):64–67. July 2014. ISSN: 2231-2307.
39. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol (TIST). 2011;2(3):27.
40. Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Transl Psychiatry. 2015;5(5):514. doi:10.1038/tp.2015.7.
41. Thabtah F, Cowling PI, 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), (pp. 217–24). Brighton, UK, November 2004. (Nominated for the Best paper award).
42. Baralis E, Chiusano S, Graza P. A lazy approach to associative classification. IEEE Trans Knowl Data Eng. 2008;20(2):156–71. ISSN: 1041-4347. doi:10.1109/TKDE.2007.190677.
43. Hall M (1999) Correlation-based Feature Selection for Machine Learning. Thesis, Department of computer science, Waikaito University, New Zealand.
44. 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.
45. Zhang YP, Zhang LN, Wang YC. Cluster-based majority under-sampling approaches for class imbalance learning 2010 2nd IEEE International Conference on Information and Financial Engineering (ICIFE), Chongqing, China, September 17–19, . 2010;400–404.
46. Yen SJ, Lee YS. Cluster-based under-sampling approaches imbalanced data distributions. Exp Sys App. 2009;36:5718–27.
47. Chyi YM. Classification analysis techniques for skewed class distribution problems. Chennai: Department of Information Management, National Sun Yat-Sen University; 2003.
48. Thabtah F. Review on associative classification mining. J Knowledge Eng Rev. 2007;22:1, 37–65. Cambridge Press.
49. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(27):1226–38. doi:10.1109/TPAMI.2005.159.
50. Liu H, 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.
51. Quinlan J. Induction of decision trees. Mach Learn. 1986;1(1):81–106. March 1986. doi:10.1007/BF00116251.
52. Lichman M. UCI machine learning repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science; 2013.

53. Rogati M, Yang Y (2002). High-performing feature selection for text classification. In *Proceedings of the eleventh international conference on Information and knowledge management* (659–61). ACM.
54. Uncu Ö, Türkşen IB. A novel feature selection approach: combining feature wrappers and filters. Inf Sci (Ny). 2007;177(2):449–66. doi:10.1016/j.ins.2006.03.022.
55. Tsai CF, Hsiao YC. Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches. Decis Support Syst. 2010;50(1):258–69. doi:10.1016/j.dss.2010.08.028.
56. Abbas H, Garberson F, Glover E, Wall DP. Machine learning approach for early detection of autism by combining questionnaire and home video screening. arXiv Preprint. 2017;arXiv:1703.06076.
57. Nakai Y, Takiguchi T, Matsui G, Yamaoka N, Takada S. Detecting abnormal voice prosody through single-word utterances in children with autism spectrum disorders: machine-learning-based voice analysis versus speech therapists. Percept Mot Skills. 2017;124(5):961–973. .