Autism Spectrum Disorder Diagnosis using Optimal Machine Learning Methods

Maitha Rashid Alteneiji[1], Layla Mohammed Alqaydi[2]
Abu Dhabi School of Management
Abu Dhabi, UAE

Muhammad Usman Tariq[3]
Assistant Professor, Abu Dhabi School of Management
Abu Dhabi, UAE

***Abstract*—Autism spectrum disorder (ASD) is the disorder of communication and behavior that affects children and adults. It can be diagnosed at any stage of life. Most importantly, the first two years of life, regardless of ethnicity, race, or economic groups. There are different variations of ASD according to the severity and type of symptoms experienced by people. It is a lifelong disorder, but treatment and services can improve the symptoms. The literature focuses on one of the main methods used by physicians to diagnose ASD. Many types of research and medical reports have been reviewed; however, a few of them only give good medical results for the strong differentiation of ASD from healthy people. This paper focuses on using machine learning algorithms to predict an individual with specific ASD symptoms. The target is to predict an individual with specific ASD symptoms and finding the best machine learning model for diagnosis. Further, the paper aims to make the autism diagnosis faster to deliver the required treatment at an early stage of child development.**

Keywords*—Autism diagnosis; autism disorder; autism detection; machine learning; ASD

I. Introduction

Artificial Intelligence has increasing importance in society. The main aim of the paper is to use artificial intelligence and machine-learning models that can help in medical fields by finding the optimal model that can recognize individuals with specific Autistic Spectrum Disorder symptoms [8]. The attention to the Artificial Intelligence field has been grown in the last few years [9]. This interest is not only motivated by the trend of designing models with human thoughts or behaviors, but also for the way of their use in real life [2,6]. The development of such models represents an ambitious and competitive task among many scientists and programmers [3]. The ability to harness artificial intelligence in medical matters has been an important topic since the beginning of this century [1]. Since then, devices and ideas have developed to facilitate the prediction and detection of specific diseases by refining machine education. Based on the importance of artificial intelligence, this paper aims to identify and implement the machine learning process that produces an algorithm capable of detecting whether a person has Autism Spectrum Disorder. Since early intervention affords the best opportunity to support healthy development and deliver benefits across the lifespan, this paper helps parents to recognize if their child has an ASD at an early age [4,7]. It also has a value in the health sector since there is no valid health reason for this disease. Based on autism speaks organization in USA, ASD diagnoses by applying behavioral exams or questionnaires, which require a lot of time and effort provided from parents and clinicians [12]. Therefore, this work shows the power of machine learning algorithms in detecting individuals with specific Autistic Spectrum Disorder symptoms [5].

Further, this paper aims toward making the diagnosis of autism a faster process that enables delivery of therapy at earlier and more impactful stages of child development using machine-learning algorithms. The introduction is the first section of this paper. It mainly gives a general overview of the autism spectrum disorder. It also represents the target group and the benefits to the community. The second section is the theoretical background concerning this paper. It focuses on Autism Spectrum Disorder diagnosing methods and comparing these methods to choose the optimal diagnosing method database. It also gives an overview of machine learning approaches and the evaluation of machine learning. The next section is the methodology, which fits the knowledge discovery process to the available ASD datasets. The fourth section is the results, which represent in detail the individual results for the entire applied machine learning models, in terms of the performance, and a comparison between them to find the optimal machine learning model. The last sections, which are a discussion and conclusion, summarize the work and a discussion about the research questions. Also, it gives an outlook for possible future papers and improvements on this topic.

II. Theoretical Background

Various autism rating scales have been developed over the past 30 years to diagnose the autism spectrum disorder (ASD) at an earlier stage [1,10]. There are various tools and instruments developed by psychologists and neuroscientists to diagnose it at earlier stages [11]. Most of the tools focus on the diagnosis of users through screening methods [3].

A. Autism Behavior Checklist (ABC)

One of the methods is the autism behavior checklist (ABC) based on identifying ASD in children at early stages [4]. The benefit of the checklist utility is to evaluate the current autistic symptoms of the user with the help of parents in different situations and conditions. It provides a set of questions to evaluate the in-depth condition of the user [6]. The method combines different scales, such as language, object recognition, body, sensory, social, and daily use skills [10]. The item scores are mostly from 1-5 based on the impairment degree [2,5]. The ABC has been in use for more than 30 years as a rapid tool for diagnosing autism in early stages [7,13].

However, there is no general agreement to use the same values as a standard method [1,3, 14].

B. Child Behavior Checklist (CBCL)

The Child Behavior Checklist (CBCL) is one of the oldest screening tools and most widely used standardized measures in child psychology for evaluating unusual behavioral and emotional problems [15]. The CBCL questionnaire focuses on internalizing and externalizing behaviors such as anxiety, over-control, aggression, and hyperactivity. There are two versions of The Child Behavior Checklist: a preschool version (aged 2 to 5) and a school-age version (aged 6 to 18). The preschool version questionnaire carries 100 questions with three scale responses ranging from 0-2, where 0 represents 'Not True,' and 2 indicates 'Very True.' The school-age version questionnaire carries 118 questions with the same rating scale responses [4, 9, 16].

C. Social Communication Questionnaire (SCQ)

The Social Communication Questionnaire (SCQ) is an ASD-screening tool used for age four and above. The SCQ consists of 40-items based on parent-report screening measures after a semistructured parent interview with a trained clinician or researcher that can be used for diagnostic ASD symptomatology. There are two different versions of the SCQ. The SCQ' Current' asks a respondent to indicate whether behaviors have been present during the past three months. The other version is the SCQ 'Lifetime' references complete developmental history and asks respondents to indicate whether behaviors have ever been present [11, 17].

D. Autism Spectrum Quotient (AQ)

The Autism-Spectrum Quotient (AQ) is among the most widely used scales assessing autistic traits in the general population. The AQ is a self-administered questionnaire for measuring how adults with normal intelligence show autistic traits. It consists of 50 questions, with ten questions assessing five different domains relevant for autistic traits: social skill, attention switching, attention to detail, communication, and imagination. People with a clinical diagnosis tend to score above 32 out of 50 on the AQ [10].

E. Short Autism Quantitative (AQ-10)

In 2012, Allison et al. create a new version of Autism-Spectrum Quotient (AQ) for adults consisted of 10-items only to make it simpler and more timesaving. AQ-10 has a predictive power similar to the origin AQ version [5,18]. Later on, Allison created shorter versions for adolescents and children with ten items too. Score calculations for the adolescent and child short versions are different from the AQ adult short version. The diagnosis depends on the final questionnaire score in behaving with some genetic information [10, 20].

F. Comparison of ASD Diagnosing Methods

The paper evaluated different ASD diagnosing methods. Table I provides a comparison between the ASD screening methods discussed in the first section of this section. As noted, most ASD screening tools focus on infants, toddlers, and children. Almost all diagnosing methods used questionnaires to diagnose ASD behaviors [19]. CBCL has the maximum number of items in the questionnaire with 118 items, while the AQ-10 has the minimum number of questionnaire items with only ten items. Screening is valid if it detects most cases with the target disorder, which gives a high sensitivity rate and excludes most cases without the disorder, which gives a high specificity rate [21, 30].

TABLE I. COMPARISON OF ASD DIAGNOSING METHODS

Diagnosing method	# Q	Target	Specificity	Sensitivity
CBCL	118	Children & Adolescents	82%	75%
ASSQ	27	Children & Adolescents	86%	91%
ABC	57	Children	91%	77%
AQ	50	Children	95%	95%
AQ	50	Adolescents	NA	NA
AQ	50	Adult	52%	93%
AQ-10	10	Children	74%	77%
AQ-10	10	Adolescents	NA	NA
AQ-10	10	Adult	NA	NA
SCQ	40	Children & Adolescents	93-100%	58-62%

In ASD screening methods, sensitivity refers to the true positive rate, which is the ability of the screening tool to identify a person with autism [29]. Specificity refers to the true negative rate, which is the power of the screening tool to identify a person who is control of autism. In terms of validity, almost all the screening methods have acceptable sensitivity rates, ranging from 70%- 100% and specificity between 80% and 100% [6, 22]. As shown in Table I, AQ screening is the most efficient method with only ten questions, which require less time to complete than other methods. Further, AQ deals with many age segments, and each of them has a specific questionnaire, which will be discussed in detail in the methodology section [15, 23].

G. Knowledge Discovery in Database

Knowledge Discovery in Databases, KDD, is an exploratory analysis and modeling of big data. KDD is the organized process of identifying useful, valid, and meaningful patterns from big databases [29]. The core of the KDD process is data mining, which explores the unknown patterns of the algorithms to develop the models. The model uses for predicting new unknown instances [30,31]. There are nine iterative processes in KDD listed below:

1) Understanding the application domain: This process defines the goals of the end-user and the environment in which the KDD process will take place with a full understanding of what should do.

2) Creating the data set: This process checks the available data and then integrates it with additional obtained data into one data set for the knowledge discovery.

3) Preprocessing: The available data goes to the preprocessing step, which includes handling missing values

and removing noise or outliers or duplicated data to enhance the reliability of the data.

4) Data transformation: This step prepares and develops better data to create the best possible model. It includes customizing the data dimension by feature selection and record sampling.

5) Choosing the appropriate data mining task: The main goal of this process is to decide on which type of data Mining to use. Data mining types include clustering, classification, and regression, depending on the DM goal, either prediction or description.

6) Choosing the data mining algorithm: This step includes selecting the appropriate searching patterns to use. Each algorithm has parameters and tactics of machine learning.

7) Applying the Data Mining algorithm: To get a satisfying result in this process, it might require applying it several times.

8) Evaluation: The model with the found patterns is evaluated and interpreted concerning the goals mentioned in the first process. This step focuses on the usefulness and comprehensibility of the induced model.

9) Using the discovered knowledge: The final step is to try the knowledge into other systems for further action and make changes to the system and measure the effects.

H. Recent Machine Learning Research on ASD Screening and Diagnosis

With the fast growth in the big data field, Autism Spectrum Disorder research must benefit from this area as other searching fields. By looking at the available research about using machine learning in diagnosing ASD, it has been clear that there is a positive effect of using machine learning in diagnosing ASD with a database containing an ASD screening method with some genetic information. Machine learning can be sorted as unsupervised and supervised learning. ASD diagnosing is a supervised machine learning that has a dependent variable to be predicted, which is the result of the diagnosis [5, 24]. A successful supervised machine-learning model is the one that can predict the target correctly and generalize new instances predicts. Usually, model validation can be measured by accuracy, which has two subtypes, sensitivity and specificity. Another measurement for the model's accuracy is the area under the receiver operating character curve, AUC [25]. The AUC shows how well a method makes positive and negative categorical distinctions between sensitivity and specificity. According to Kayleigh's research, most of ASD diagnosing models used ADTree and SVM algorithms. ADTree is a classification machine-learning algorithm that consists of an alternation of decision nodes, which specify a predicate condition, and prediction nodes, which contain a single number. An ADTree classifies an instance by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed. [12, 26] The objective of the support vector machine algorithm, SVM, is to find a hyperplane in N-dimensional space (N: the number of features) that distinctly classify the data points. [2, 5, 27] data, data mining, uses complex mathematical machine learning algorithms [28].

III. Methodology

This section starts by describing the paper's technical framework and then identifies the data-mining goal of this paper. The following parts include describing the data collection processes, the specification of a DM approach and algorithm, the optimal machine learning models to be used with supervised cases, and how it will be evaluated. The predefined goal was to develop a model that can recognize an individual with specific ASD symptoms using specific machine-learning methods. The next practical working steps will take place in this paper are:

- Find the right datasets for paper use and apply data mining steps to find the appropriate machine learning models that can expect ASD symptoms.
- Apply several machine-learning algorithms to critically review, evaluate, and compare one another to choose the optimized prediction model.
- Provide a machine-learning algorithm with optimal prediction quality for identifying an individual with specific ASD symptoms without access to the class label.
- Introduce a proper framework for ranking the quality of the diagnosing models.

A significant limitation is that this work focuses on evaluating some ASD characteristics that may affect the diagnostic result, while the simple reasons and factors for having an Autistic Spectrum Disorder are still not clear all over the world.

A. Technical KDD Framework

The proposed Knowledge Discovery in Database framework of ASD detection traits is shown in Fig. 1. The Knowledge Discovery process (KDD) process starts with understanding the autism spectrum disorder symptoms and factors, which are discussed throw the second section. The next three processes are related to data collection, data pre-processing, and data transformation.

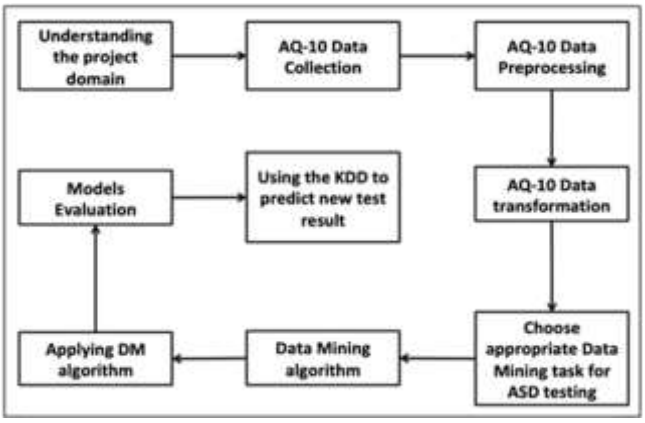

Fig. 1. KDD Process in Database.

B. Data Collection

The data used in this paper is secondary data; the AQ-10 ASD data. Data was collected through a mobile application using the ASD AQ-10 diagnostic method, and healthcare professionals relied on the final diagnosis. The primary purpose of the app is to collect useful data about ASD cases. This data is stored in a MySQL database to understand the main features that may affect ASD diagnostics using data analysis. Also, the data was divided into three databases, since each age category has a specific diagnosing questionnaire.

C. Data Preprocessing

The collected data usually cannot be used directly in performing the analysis process. Therefore, the raw data needs to be cleaned and performed in a usable format. Cleaning the data includes replacing or removing missing values and discretization for certain continuous variables such as the age of individuals; this step is called the data pre-processing. Before using the available data, all the redundant or unnecessary variables must be removed from the Database. The Null values, the columns that do not have unique values, must be removed too. This process will improve the model's prediction and raise the accuracy rate. The variables that have been removed in the ASD databases are as listed:

- Case number: the case number was added as a counter of all the cases that used the ASD screening.
- ASD Screening type: this variable was added to split the data into spirit databases based on age categories.
- Reasons for taking the screening: this variable contains texts, and it did not add any value to the data analysis. Also, it will negatively affect the prediction model results.
- Language: since the app is available for people worldwide, the diagnostic test is also available with the most common languages.
- User: this variable represents the person who answers the screening instead of the child.
- Used app before: the users ware asked this question to avoid attribute duplication.

D. Data Transformation

On the other hand, data transformation includes customizing the data dimension by feature selection according to the model needs [31]. These two operations are mandatory to achieve better performance and accuracy in the Machine Learning prediction models. Data mining processes are the next step in the KDD processes that consists of applying data analysis and discovery algorithms that produce a particular enumeration of models over the data. In order to find the optimal machine-learning algorithm that gives the best ASD result prediction and discover stricter rules, the data mining type for the user database must be specified. The rule phase will be discovered by a classification system that will be used to predict the value of the unseen cases. All the possible machine-learning algorithms will be evaluated using the confusion metrics to find the accuracy, sensitivity, and specificity. The last process is sharing the discovered knowledge with the health professional so it can be used in a professional way to serve society and help parents to discover their kids' status from an earlier age.

E. Data Description

In this paper, the used databases relating to specific age groups, which are infants, children, and adolescents. The datasets can be divided into ten behavioral questions for each age group, and several variables that influence the final evaluation of the condition are used in the diagnostic database. The influencing variables include age, gender, ethnicity, jaundice, and family history. Table II shows these variables with the data type and the description for each variable.

The next three Tables III, IV, and V show the ten variables details in the toddler, adolescent, and children screening methods. The three databases carry out ten symptoms to be answered with either yes or no. These ten questions are the most noticed symptoms in diagnosing individuals with ASD.

TABLE II. VARIABLES USED FROM ASD DATABASE FOR DIAGNOSIS

Variable	Data type	Description
A1 – A10	Binary	YES/NO
Age	Continuous	Age of Individual
Gender	Binary	Male/Female
Ethnicity	Categorical	List: White Middle Eastern White European Asian Black Latino Mixed Others
Jaundice	Binary	YES/NO
Family history	Binary	YES/NO
Nationality	Categorical	List (All the worlds' counties)
Target class	Binary	YES/NO

TABLE III. AQ-10 CHILDREN SCREEN FEATURES

Variable	Child screening features
A1	S/he often notices small sounds when others do not
A2	S/he often notices small sounds when others do not
A3	In a social group, s/he can easily keep track of several different people's conversations
A4	S/he finds it easy to go back and forth between different activities
A5	S/he doesn't know how to keep a conversation going with his/her peers
A6	S/he is good at social chit-chat
A7	When s/he is read a story, s/he finds it difficult to work out the character's intentions or feelings
A8	When s/he was in preschool, s/he used to enjoy playing pretending games with other children
A9	S/he finds it easy to work out what someone is thinking or feeling just by looking at their face
A10	S/he finds it hard to make new friends

TABLE IV. AQ-10 ADOLESCENT SCREENING FEATURES

Variable	Adolescent screening features
A1	S/he notices patterns in things all the time
A2	S/he usually concentrates more on the whole picture rather than the small details
A3	In a social group, s/he can easily keep track of several different people's conversations
A4	If there is an interruption, s/he can switch back to what s/he was doing very quickly
A5	S/he frequently finds that s/he doesn't know how to keep a conversation going
A6	S/he is good at social chit-chat
A7	When s/he was younger, s/he used to enjoy playing games involving pretending with other children
A8	S/he finds it difficult to imagine what it would be like to be someone else
A9	S/he finds social situations easy
A10	S/he finds it hard to make new friends

TABLE V. AQ-10 TODDLER SCREENING FEATURES

Variable	Toddler screening features
A1	S/he often looks at you when you call his/her name
A2	S/he often can easily get eye contact with you
A3	S/he can easily point to indicate that s/he wants something
A4	S/he can easily point to share interest with you
A5	S/he can easily pretend
A6	S/he is good at social chit-chat
A7	S/he can easily follow where you are looking
A8	When you or someone in the family is upset, s/he can show sings of wanting to comfort them
A9	S/he first words was typical ones
A10	S/he finds it easy to use simple gestures

Based on the available database, the total number of attributes in this paper is equal to 1811 attributes after removing the NULL and redundant attributes. The attributes are divided into three spirit databases since each age group must deal with specific symptoms. The toddler age group carries more than half of the aggregate data, with around 70% diagnosed as having autism spectrum disorder. Child and Toddler age groups distribution divided by half for each gender, which also divided by half for the diagnosing result.

F. Specification of DM Approach and Algorithm

This part deals with the KDD-steps 5, 6 and 7, including choosing the appropriate Data Mining task and algorithm then applying it. Evaluating the results of all appropriate algorithms leads to choose the optimal machine-learning model. The data used in this paper mainly focuses on diagnosing individuals with ASD symptoms, based on AQ-10 with several variables that usually affect the diagnosing result. Hence, the prediction model is considered a classification problem that results from either having ASD or not. Therefore, many suitable supervised models were applied for the given task, the results were analyzed and evaluated. Before applying these machine-learning models, the feature selection process must be applied to support the evaluation results and the models' accuracy by removing the weak variables from the databases. The following feature selection techniques and supervised classification models were considered suitable for the ASD diagnosing task.

IV. RESULTS AND DISCUSSION

This section gives an overview of the achieved results, the experiment process to solve the research question, and visualization of those results. It focuses on feature selection techniques and results, the machine-learning model's evaluation measures based on feature selection results, and the standard rules related to ASD detection that has been extracted by the best machine-learning model.

A. Simulation and Implementation

The simulations and operations evaluate the strength of the statistical procedure and identify the machine-learning model's strengths and weaknesses using the confusion matrix that leads to the simulation result. In RStudio, a machine-learning model can give a confusion matrix as a model evaluation tool. Applying the model and extracting the matrix can be used to define a set of mathematical values to determine the efficiency of the model and choose the best model to anticipate the results of the database set. There are sets of values that are taken into account, and they are error rate, accuracy evaluation, sensitivity, specificity, and FN-value. The visualizations associated with these values were also constructed using the Tableau software.

B. Feature Selection Results and Analysis

As mentioned earlier in Section 3, the ASD databases' prediction model is considered a classification-predicting problem that carries out a categorical label variable, with categorical input variables. The feature selecting algorithms mainly evaluate the relationship between ASD test results independently with each other variable in the database using two filter-based techniques: Chi-Squared and mutual information. The lines of code related to the mutual information techniques, which is the information gain method, and the Chi-Squared were applied to evaluate the autistic trait features in all the available datasets then compared the performance for each technique. All the databases have eight selected features. It shows the highest eight variables in performance for each database based on the result of the mutual information and the Chi-Squared techniques. Both feature selection techniques give the same result with a few differences in the order of the variables based on the efficiency in the toddler database. All databases relied on the AQ questions and showed a high correlation with the ASD diagnosing results. The variables in the feature selection techniques were the only variables used in the machine learning models to improve the model's performance.

Both feature-selecting algorithms show that the A4 variable in the Child database has the highest correlation with the target class, resulting from the ASD test. Also, both show that the A6 variable in the Adolescent database has the highest

correlation, which carries out the "S/he is good at social chit-chat" questionnaire sentence. For the Toddler database, both techniques show that the A9 has the highest correlation with the label class. The A9 in the Toddler database carry out the "S/he first words were typical ones" questionnaire sentence.

C. Machine Learning Model Evaluation Measures based on Feature Selection

The evaluation techniques used in this paper are based on the result of the confusion matrix of each machine-learning model. The models' performance can be evaluated by calculating the error rate, accuracy, sensitivity, and specificity.

D. Error Rate and Accuracy Evaluation

After applying all the machine learning models that fit the ASD classification problem in RStudio, the accuracy rate measurement is described in Table VI.

The above comparison shows that the Neural Network model has the highest accuracy rate measurement in each database compared with the other machine-learning models. The toddler database has the best accuracy results compared to the child and adolescent databases. The number of attributes in the toddler database is much higher than the other two databases, affecting the accuracy rate result. This result indicates that the toddler age group is the best age to diagnose if they have an ASD. Neural networks can learn complex and non-linear relationships. It can infer unseen relationships on unseen data and give the model the ability to generalize and predict unseen data. Fig. 2 depicts each database's trends based on the error rate measurement in percentage and the applied machine-learning models. The color indicates the three databases: the Adolescent database, Child database, and Toddler database. Since the figure deals with the error rate, the lower the value, the batter in results performance. It shows that the toddler database controls their error rates better than the other databases, with rates between 0.96% and 5.75%. The adolescent database shows the highest error rates, which may be due to the small size of the database compared to the rest. Also, the child's psychological changes during the adolescent period may significantly affect the validity of expectations. In comparing the machine learning models used in this paper, the Neural Network model gives the lowest error rate measurements comparing with the rest models in all databases. This shows that the Neural Networks model does not perform well only on datasets with significant data attributes, such as the Toddler dataset, but also with datasets with a limited number of attributes, such as the Adolescent dataset.

E. Sensitivity and Specificity

Fig. 3 and 4 display the sensitivity rates and specificity rates derived by the SVM, Naïve Bayes, Neural Network, Random Forest, GBM, XgBoost, AdaBoost, and CV Boosting algorithms on the Child, Adolescent, and Toddler datasets. Both the sensitivity rates and specificity rates results generated by the considered algorithms on all the datasets have shown acceptable levels of performance.

Neural Networks Model has higher sensitivity and specificity rates than most of the remaining algorithms on all the available datasets. For the Child database, Neural Network Model achieved a 96.3% sensitivity rate and 97.2% specificity rates, while the adolescent database achieved 100% and 90.9% sensitivity rate and specificity rates. For the Toddler database, the sensitivity and specificity rates achieved 98.6% and 100%. As shown in the previous dashboards, some of the machine-learning models cannot perform well in small datasets, while the Neural Network Model shows outstanding sensitivity and specificity rates.

TABLE VI. ACCURACY RESULTS

ML model	Accuracy in %		
	Child-database	Adolescent-database	Toddler-database
SVM	95.41284	91.89189	96.83544
XgBoost	92.07921	75.51020	97.14286
AdaBoost	90.13158	93.24324	94.60317
CV Boosting	92.73084	82.37096	94.68691
Neural Network	96.73203	96.10390	99.03537
Random Forest	87.33333	88.60759	94.24920
Naïve Bayes	92.53438	94.75806	94.59203
Random Forest- GBM	93.33333	93.67089	95.84665

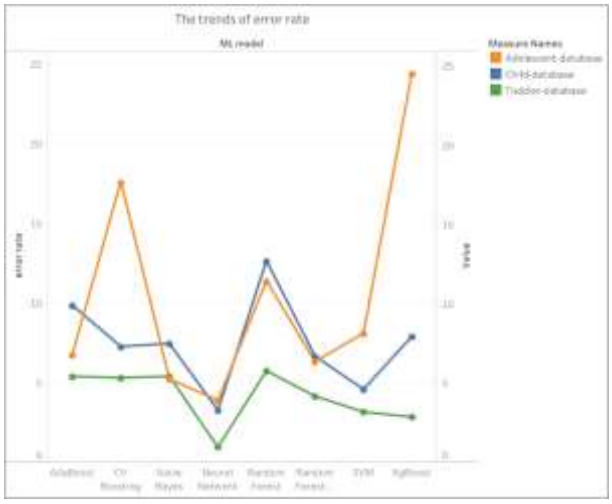

Fig. 2. Error Rate Results.

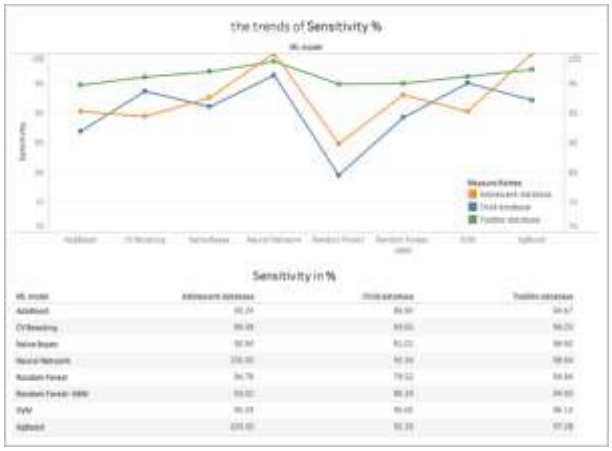

Fig. 3. Sensitivity Rate Results.

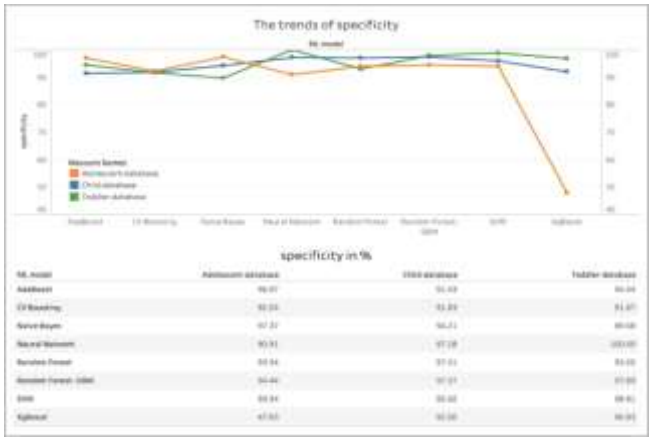

Fig. 4. Specificity Rate Results.

Overall, the results reported higher performance for the Neural Network model on the datasets when compared with the considered machine-learning algorithms, and these results are consistent with the error rate produced earlier and can be attributed to the non-redundant rules sets generated by the Neural Network model that will be discussed in the next section.

F. False Negative Rate

In a binary classification medical test, false negative is an error in which the test result incorrectly indicates that there is no case, when present in reality. Overall, many models show good results in the FN rate in the three databases, as shown in Table VII. Neural Networks Model has the best FN rate in two databases and a good result in the third one, the adolescent database.

TABLE VII. FN RATE

ML model	Child-database	Adolescent-database	Toddler-database
SVM	1.8%	2.7%	0.3%
XgBoost	3.96%	24.5%	0.95%
AdaBoost	2.6%	1.4%	2.1%
CV Boosting	4.1%	3.6%	2.6%
Neural Network	1.3%	3.9%	0%
Random Forest	1.3%	2.5%	2.2%
Naïve Bayes	2.8%	1.2%	3.3%
Random Forest- GBM	1.3%	2.5%	0.6%

G. Association Rules Result based on Feature Selection

The Association Rule is a rule-based machine-learning technique that discovers interesting relations between variables in large databases. Since the number of founded rules using the apriori function can be controlled by the values of Support and Confidence parameters, fixing the generated parameters will give balanced rules results. Many variables have frequently appeared within the rules that cover specific ASD characteristics within the databases. Variables A4, A1, and A6 in the Child database strongly influence the class labels. Additionally, variables A3, A4, and A6 appeared in multiple rules in the Adolescent database, while items A2, A4, and A9 appeared in many rules in the Toddler database. The association rule shows that combining the results of two or three variables gives the best correlation between the variables and the diagnosing result, which is the class variable.

V. DISCUSSION

The paper's primary aim was to provide the best machine-learning model that diagnoses individuals with specific Autistic Spectrum Disorder symptoms. Several processes were needed to select the best machine-learning model. It was necessary to choose the most efficient ASD questionnaire diagnosing method and collect a high-quality database for each age group, which was done by the help of previous research studies and surveys conducted among the health professionals. With the available datasets and the applied data mining algorithms, the most accurate model was selected as the best machine-learning model to diagnose ASD symptoms. The first part focused on Knowledge Discovery in Databases processes, which is the best way to predict an individual with specific Autistic Spectrum Disorder characteristics using machine-learning models. The Knowledge Discovery in Databases processes that had been followed in this paper is as the following:

- Understanding the application domain
- Creating the data set
- Preprocessing and Data transformation
- Choosing the appropriate Data Mining task
- Choosing and applying the Data Mining algorithm
- Evaluation
- Using the discovered knowledge

These seven processes were followed in this paper, starting by searching and reading about Autistic Spectrum Disorder symptoms and diagnosing methods to choose the most efficient ASD questionnaire diagnosing method. Collecting and processing a high-quality database for each age group are two essential success factors for any data-mining paper. The data was managing in cooperation with the health professionals to ensure that the diagnosing results in the database are correct. The feature selection process supports the evaluation results and the models' accuracy by removing the weak variables from the databases.

After preparing the dataset, the data-mining task was set to a classification task, since the class label is categorical. Many suitable supervised models were applied for the given job, such as support vector machine, naïve Bayes, neural networks, and ensemble methods. The evolution process showed the result performance for each machine-learning models. All these processes answered the first research question, asking about how to use machine learning to diagnose individuals with specific Autistic Spectrum Disorder characteristics. The second part has been answered in the evaluation process in the Knowledge Discovery in Databases. They used machine-learning models were compared by the performance, which includes the accuracy rate, the sensitivity rate, and the

specificity rate. This comparison concluded that the neural networks model gives the best performance practice for the three databases.

VI. Results Comparison

The results of this paper give a better performance comparing with the papers reviewed in Section 2. The available databases in this paper are considered one of the best available datasets those days, since it deals with each age group as a separate database, and contains a good number of attributes. The toddler database using the Neural Network model gives the best accuracy result, while adolescent and child databases using the Neural Network model also have excellent accuracy results compared with the other models.

VII. Conclusion and Future Work

This paper aimed to provide useful and accurate ASD screening models to help parents and interested parties quickly diagnose their children's condition. Unfortunately, some families and adult patients do not have sufficient knowledge of ASD symptoms, so cases of autism spectrum disorder are not dealt with early. Artificial intelligence and machine learning are used at this time in most living areas, and their use in the field of medical diagnosis contributes to a pioneering step in using the available data as a tool for development and progress. All the primary seven processes of the KDD was used and described in this paper. These processes contain data gathering and data preprocessing, choosing an appropriate data mining approach to find patterns among the data and interpret them. Finally, the results were used for further research. The empirical results on the used datasets related to children, adolescents, and toddlers show that the neural networks model yielded the highest performance results compared to the other machine learning models used in this paper concerning predictive power, sensitivity, and specificity.

The development of this paper into an application program will provide families with a quick and straightforward scan tool using the lowest set of elements related to ASD, which contributes to increased accessibility and early detection. In the future, it is possible to develop this paper for use in the health system of the Ministry of Health and Prevention, where data are available for all patients registered in all the hospitals affiliated with the Ministry in this system. It is also possible to provide the departments of schools, kindergartens, and nurseries with an easy-to-use system for this paper to be applied to children for early detection. Another potential area for further use of this study could be the application of machine education and artificial intelligence models in health systems that store patient data for a range of diseases and health symptoms to contribute to the early detection of potential diseases.

References

[1] American Psychiatric Association . (2013, 10 10). American Psychiatric Association . Retrieved from American Psychiatric Association : https://www.psychiatry.org/.

[2] Ben-David, S. S.-S. (May 2014). Understanding Machine Learning: From Theory to Algorithms. -: Cambridge University Press.

[3] Becerra-Culqui, T. A., Lynch, F. L., Owen-Smith, A. A., Spitzer, J., & Croen, L. A. (2018). Parental first concerns and timing of Autism Spectrum Disorder diagnosis. Journal of autism and developmental disorders, 48(10), 3367-3376.

[4] Bishop-Fitzpatrick, L., Movaghar, A., Greenberg, J. S., Page, D., DaWalt, L. S., Brilliant, M. H., & Mailick, M. R. (2018). Using machine learning to identify patterns of lifetime health problems in decedents with autism spectrum disorder. Autism Research, 11(8), 1120-1128.

[5] Bravo Oro A., N.-C. M. (2014). Autistic Behavior Checklist (ABC) and Its Applications. New York: Springer.

[6] Carla A. Mazefsky, R. A. (2011, March -). PubMed Central. Retrieved May 30, 2012, from ncbi.nlm.nih.gov: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3362998/.

[7] Duvekot, J., van der Ende, J., Verhulst, F. C., Slappendel, G., van Daalen, E., Maras, A., & Greaves-Lord, K. (2017). Factors influencing the probability of a diagnosis of autism spectrum disorder in girls versus boys. Autism, 21(6), 646-658.

[8] Gandhi, R. (2018, june 7). towardsdatascience. Retrieved june 7, 2018, from towardsdatascience.com: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[9] Greenhalgh, T. (1997). How to read a paper. Papers that report diagnostic or screening tests. London: University College London Medical School.

[10] Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., ... & Awashti, S. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nature genetics, 51(3), 431-444.

[11] Hershy, A. (2019). Gini Index vs Information Entropy. Retrieved from medium.com: https://towardsdatascience.com/gini-index-vs-information-entropy-7a7e4fed3fcb.

[12] Hyde, K. K., Novack, M. N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D. R., & Linstead, E. (2019). Applications of supervised machine learning in autism spectrum disorder research: a review. Review Journal of Autism and Developmental Disorders, 6(2), 128-146.

[13] Ibrahim, S., Djemal, R., & Alsuwailem, A. (2018). Electroencephalography (EEG) signal processing for epilepsy and autism spectrum disorder diagnosis. Biocybernetics and Biomedical Engineering, 38(1), 16-26.

[14] Jung, H. (2018). medium.com. Retrieved from https://towardsdatascience.com/adaboost-for-dummies-breaking-down-the-math-and-its-equations-into-simple-terms-87f439757dcf.

[15] LeBarton, E. S., & Landa, R. J. (2019). Infant motor skill predicts later expressive language and autism spectrum disorder diagnosis. Infant Behavior and Development, 54, 37-47.

[16] Lindner, L.-O. L. (2017). Is the Autism-Spectrum Quotient a Valid Measure of Traits Associated with the Autism Spectrum? A Rasch Validation in Adults with and Without Autism Spectrum Disorders. Journal of Autism and Developmental Disorders , 47.

[17] Liu, W., Li, M., & Yi, L. (2016). Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. Autism Research, 9(8), 888-898.

[18] Lord, C., Elsabbagh, M., Baird, G., & Veenstra-Vanderweele, J. (2018). Autism spectrum disorder. The Lancet, 392(10146), 508-520.

[19] Marvin, A. R. (2017). Analysis of Social Communication Questionnaire (SCQ) Screening for Children Less Than Age 4. US: springe.com.

[20] Mason, Y. F. (1999). The Alternating Decision Tree Learning Algorithm. ICML '99 Proceedings of the Sixteenth International Conference on Machine Learning (pp. 124-133). San Francisco: Morgan Kaufmann Publishers Inc.

[21] Mohammad Moshirpour, B. H. Applications of Data Management and Analysis:. Calgary, Canada: springer.

[22] NIH. (2018, March 1). Retrieved from The National Institute of Mental Health Information Resource Center : https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml.

[23] Omar, K. S., Mondal, P., Khan, N. S., Rizvi, M. R. K., & Islam, M. N. (2019, February). A machine learning approach to predict autism spectrum disorder. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE.

[24] Pagnozzi, A. M., Conti, E., Calderoni, S., Fripp, J., & Rose, S. E. (2018). A systematic review of structural MRI biomarkers in autism spectrum

disorder: A machine learning perspective. International Journal of Developmental Neuroscience, 71, 68-82.

[25] Peebles, F. T. (2019). Early Autism Screening: A Comprehensive Review . International Journal of Environmental Research and Public Health — Open Access Journal , PMC6765988.

[26] Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., & Reichenberg, A. (2017). The heritability of autism spectrum disorder. Jama, 318(12), 1182-1184.

[27] Sharma, S. R., Gonda, X., & Tarazi, F. I. (2018). Autism spectrum disorder: classification, diagnosis and therapy. Pharmacology & therapeutics, 190, 91-104.

[28] Stevens, E., Dixon, D. R., Novack, M. N., Granpeesheh, D., Smith, T., & Linstead, E. (2019). Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. International journal of medical informatics, 129, 29-36.

[29] Tan, C. D. (2018). "I'm a normal autistic person, not an abnormal neurotypical": Autism Spectrum Disorder diagnosis as biographical illumination. Social Science & Medicine, 197, 161-167.

[30] Thabtah, F. F. (2017). uci. (Manukau Institute of Technology) Retrieved from uci.edu: http://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult

[31] Thabtah, F., & Peebles, D. (2020). A new machine learning model based on induction of rules for autism detection. Health informatics journal, 26(1), 264-286.