Iraqi Journal of Science, 2024, Vol. 65, No. 9, pp: 5239-5253
DOI: 10.24996/ijs.2024.65.9.36

ISSN: 0067-2904

Data Mining Technique for Diagnosing Autism Spectrum Disorder

Rasha Hani Salman[1*], Manar Bashar Mortatha[2], Riydh Rahef Nuiaa[2]
[1]Presidency of the University, Wasit University, Wasit, Iraq
[2]Department of Computer Science, College of Education of Pure Sciences, Wasit University, Wasit –Iraq

Received: 16/1/2023 Accepted: 19/8/2023 Published: 30/9/2024

Abstract

Early detection of autistic symptoms can help lower overall medical expenses, which is beneficial given that autism is a developmental disease that is associated with high medical costs. To assess whether or not a kid may have autism spectrum disorder (ASD), screening for ASD involves asking the child's parents, caregivers, and other members of the child's immediate family a series of questions. The current methods for screening for autism, such as the autistic quotient (AQ), might require a significant number of questions in addition to careful question design, which can make an autism examination more time-consuming. The effectiveness and reliability of the test could be improved, for example, by employing data mining strategies. It could be possible to create a system that can foretell ASD at an early stage and give patients, caregivers, and medical professionals dependable and precise findings on the probable need for expert diagnostic services. This research aims to develop a reliable model for estimating the likelihood of an individual being diagnosed with autism spectrum disorder between the ages of 4 and 17. To identify varying degrees of autism, one such model was constructed by utilizing the stochastic gradient descent (SGD) algorithm. Mining data is typically understood to be a decision-making process that enables more effective utilization of available resources in terms of overall performance. The results showed that the suggested prediction model, which used the stochastic gradient descent (SGD) algorithm, could find ASD with an average error of 0.03% and an accuracy of up to 94.5%.

Keywords: Autism spectrum disorder (ASD), Data mining, Classification, Logistics regression, Stochastic Gradient Descent

تقنية التنقيب عن البيانات لتشخيص اضطرابات طيف التوحد

رشا هاني سلمان[1]*, منار بشار مرتضى[2], رياض رهيف نوي[2]
[1] رئاسة الجامعة، جامعة واسط، واسط، العراق
[2] قسم علوم الحاسبات، كلية التربية للعلوم الصرفة، جامعة واسط، واسط، العراق

الخلاصة :

يمكن أن يساعد الاكتشاف المبكر لأعراض التوحد في خفض النفقات الطبية الإجمالية ، وهو أمر مفيد نظرًا لأن التوحد مرض تطوري مرتبط بتكاليف طبية عالية. لتقييم ما إذا كان الطفل مصابًا باضطراب طيف التوحد (ASD) أم لا ، يتضمن فحص ASD طرح سلسلة من الأسئلة على والدي الطفل ومقدمي الرعاية وغيرهم من أفراد الأسرة المباشرين للطفل. قد تتطلب الأساليب الحالية للكشف عن التوحد ، مثل حاصل التوحد

*Email: rashahany609@gmail.com

(AQ) ، عددًا كبيرًا من الأسئلة بالإضافة إلى تصميم سؤال دقيق ، مما قد يجعل اختبار التوحد يستغرق وقتًا أطول. يمكن تحسين فعالية وموثوقية الاختبار ، على سبيل المثال ، من خلال استعمال استراتيجيات التنقيب في البيانات. قد يكون من الممكن إنشاء نظام يمكن أن يتنبأ باضطراب طيف التوحد في مرحلة مبكرة ويعطي المرضى ومقدمي الرعاية والمهنيين الطبيين نتائج موثوقة ودقيقة حول الحاجة المحتملة إلى خدمات تشخيصية متخصصة. يهدف هذا البحث إلى تطوير نموذج موثوق به لتقدير احتمالية تشخيص الفرد باضطراب طيف التوحد بين سن 4 و 17 عامًا. لتحديد درجات متفاوتة من التوحد ، تم إنشاء أحد هذه النماذج من خلال استعمال النسب المتدرج العشوائي (SGD) الخوارزمية. تُفهم بيانات التعدين عادةً على أنها عملية صنع القرار التي تتيح استعمالاً أكثر فعالية للموارد المتاحة من حيث الأداء العام. أظهرت النتائج أن نموذج التنبؤ المقترح ، والذي استعمل خوارزمية الانحدار العشوائي (SGD) ، تكتشف التقنية ASD بدقة تصل إلى 94.5٪ وتحقق متوسط خطأ بنسبة 0.03٪.

1. Introduction

The results of this research can be applied in a variety of scenarios, making data mining a potentially fruitful and rapidly growing area of knowledge analysis. Data mining is sometimes referred to as "knowledge discovery from data" (KDD). This technology automatically or manually identifies patterns that reflect existing or implicitly collected knowledge in massive databases, data warehouses, the Internet, data warehouses, and information flows. Data mining is applied on a fairly regular basis. Information technology, machine learning, statistics, pattern recognition, information retrieval, artificial neural networks, knowledge-based systems, artificial intelligence, and data visualization are some of the academic subfields that fall under the term "data mining." The field of data mining includes several different subfields [1].

The primary goal of health services nowadays is the automated evaluation of medical care. In general, the health system is an important and essential part of people's lives. Medical professionals and patients rely on automated analysis to standardize data and give researchers a simple and adaptable data collection model. One of the biggest hurdles is obtaining an early diagnosis of autism, as counselors can learn the basic characteristics of autism over a period of up to six months. Data mining techniques have been used to speed up the process of diagnosing the characteristics of people with autism in order to prevent this lengthy process, which in turn has led to tremendous progress in this sector to improve the condition of the person with autism [2].

As mentioned [3], an important aspect of autism spectrum disorder (ASD) is that it is a neurological disease characterized by emotion through reciprocal actions, communication, and learning abilities. Autism can be diagnosed at any age, although its mental and physical symptoms are noticeable in children at an early age and get worse over time [4]. It is claimed that autism is characterized by some challenges, most notably an inability to focus, an inability to learn, mental issues including anxiety, depression, etc., motor skills difficulties, etc. The recent increase in the prevalence of autism worldwide has been significant, and its prevalence is accelerating. According to the reports of the World Health Organization, autism spectrum disorder affects 1 in 160 people. Only a small percentage of them can live independently, while the rest have ongoing needs for care and assistance [5]. Note that a significant investment of time and financial resources is required for an autism diagnosis. An accurate diagnosis made at the right time can be beneficial, as it will enable the patient to receive appropriate treatment as soon as possible. He will not be able to grow as a result, and the amount of money that will have to be paid will be reduced as a result of his late discovery. As a direct result of this, there is a great need for a scanning tool that is not only fast but also

accurate and easy to use. This tool should be able to predict an individual's characteristics and decide whether or not that individual deserves comprehensive autism screening.

The advantage of the iterative gradient descent method is that it minimizes a cost function. The slope or gradient function's partial derivative should be able to be calculated. In order to reach the local minima after a few iterations, the coefficients are generated at each iteration by calculating the derivative's inverse and lowering the coefficients at each step by a learning rate (step size) multiplied by the derivative. As a result, the iterations finally come to an end when the cost function reaches its minimum value, at which point the cost function stops decreasing [6]. The SGD technique has certain drawbacks, such as the ease with which local optimums can form and the requirement to tackle vanishing gradient problems. It also converges to the least value of the cost function, after which there is no further reduction in the cost function. [7].

The goal of this research is to create a model for the early diagnosis of autism spectrum disorder (EDASD), which can then be used to make accurate predictions about autism. This program, which utilizes data mining techniques, can generate accurate predictions of autistic features in individuals at an early age. In a different sense, the endeavor is focused on finding ways to improve a program that can diagnose autism and forecast the signs of autism spectrum disorder in people between the ages of four and seventeen years old.

They employ a mixture of arithmetic and searching methods that they learned from studying computers. As a result, data mining techniques offer an automatically generated classification scheme that is both innovative and useful in the context of autism spectrum disease. Researchers are currently employing a wide range of approaches that are interrelated to combat the issue of autism spectrum disorder (ASD). It is possible to see the diagnosis of autism spectrum disorder as an example of a perfect classification issue in machine learning. This is because a prototype may be constructed utilizing controls and cases that have been previously classified. As a direct consequence of this, the paradigm is utilized to arrive at educated assumptions regarding new case diagnoses (ASD, no-ASD). The remaining parts of the research can be broken down into the following sections: In the second section, shed light on the previous studies that are related to the topic. In this section, the many methods and approaches to the subject are discussed. In Section IV, specific information about the suggested model was investigated, just as in Section I, the methodology for validating the model and the precision of the model were investigated. The last part of the report summarizes the findings of the research, paying particular attention to the study's significance, the limits of the research, and suggestions for how the work should be expanded.

2. Review of Literature

According to what was stated in J. A. Cruz et al. [8], each action connected to the prediction process in autism spectrum disorder was covered in this article for a limited amount of time. It is remarkable how accurate machine learning can be when applied to the task of forecasting various diseases based on syndromes. With the use of machine learning, it can be determined whether or not a person has diabetes. (D. P. Wall et al. [9]) add that with a support vector machine, we were able to use machine learning with the objective itself, achieving a specificity of 59% and a sensitivity of 89%. Their research included a total of 1264 people who were diagnosed with autism spectrum disorder (ASD), as well as 462 people who did not have ASD. In any event, their research has not been approved for use as a screening method for people of varying ages since the age group they studied was too young (4–55). Allison et al. made use of the term "red flags." (C. Allison et al. [10]) screening both

children and adults using the Autism Spectrum Quotient (ASQ) to determine whether or not they have ASD, after which they developed a short list of candidates to be assessed further. On the AQ-10, it is more accurate than 90% of the time.

Accordingly, D. P. Wall et al. [11] made an effort to categorize information by making use of a short screen in addition to validating it and found the following: Both the AD tree and the functional tree performed well, exhibiting high levels of specificity, sensitivity, and accuracy, respectively. The screening method was streamlined with the help of the Alternating Decision Tree, which sped up the identification of ASD characteristics. They used the Autism Diagnostic Interview, Revised (ADI-R) method, and they were able to attain a high level of accuracy by making use of data collected from 891 individuals by D. Bone et al. [12]. A significant amount of brain imaging data was obtained through the Autism Imaging Data Exchange, which was used by him and his colleagues to carry out the implementation of a neural network and a deep learning method for the diagnosis of sick individuals who have ASD. In addition to this, they found that the accuracy of their classifications ranged from 66% to 71%, with an average accuracy of 70%. The accuracy of the stochastic gradient descent classifier was calculated to be 65%, whereas the accuracy of the random forest classifier was calculated to be 63%. A. S. Heinsfeld et al. [13] attempted to perform a diagnosis for cancer utilizing ML, just as N. S. Khan et al. [14] utilized ML to predict whether an individual has diabetes or not.

B. van den Bekerom [15] noted that researchers used a variety of machine learning mechanisms, such as the random forest algorithm, naive Bayes, and SVM, as well as the random forest algorithm, to detect the symptoms that can occur in children, such as developmental delays, being overweight, and not getting enough physical activity. These symptoms can be caused by a lack of adequate physical activity or insufficient physical activity. Accordingly, W. Liu et al. [16] used an ML algorithm for the analysis of a dataset of eye movement to classify and investigate if the patterns of face scanning can be useful for detecting children who have autism spectrum disorder (ASD). The findings of the investigation showed a sensitivity of 93.10%, an accuracy of 88.51%, an area under the curve of 89.63%, and a specificity of 86.21%.

On the other hand, F. Thabtah [17] analyzed the results of all previous studies and algorithms, including ML, in an attempt to forecast the symptoms. F. Hauck et al. [18] made an effort to discover much more critical screening questions for autism. Autism and the diagnostic observation schedule and an assessment interview re-examined screening approaches and found that improving their performance by merging them would be possible after making some adjustments. D. Bone et al. [19] looked at the earlier works of D. Bone et al. [12] and J. A. Kosmicki et al. [20] from a theoretical point of view to figure out the complexity of a mental-concept-based problem forming, implementation related to methodology, and interpretation, and made money again by using their ML approach. In general, for identifying problems in the creation of a mental-concept-based difficulty, implementing a technique associated with interpretation, and generating money with their ML approach. According to the literature review, some research was carried out on this topic; however, specialists were unable to reach a consensus regarding the application of the ML approach to the expansion of the use of an autism screening tool across a range of ages. In the past, autism screening has involved the modification of a variety of instruments and procedures; however, none of these have been converted into app-based solutions that are appropriate for use across age ranges.

It turns out from the above that it appears that many of the published studies do not have enough validation or testing. The researchers' ability in D. P. Wall et al. [9] to measure specificity was limited by the small numbers of non-spectral states in the research data used. More study is required in research (C. Allison et al. [10]) to examine the relationship patterns shared by people with anorexia and autism and to create friendship-focused therapies for anorexia patients. The development of mobile tools for initial assessment and clinical prioritization, particularly those focused on brief, child-friendly home assessment videos, will improve classification (D. P. Wall et al. [11]). This will speed up the initial assessment and increase access to a much higher proportion of children at risk of autism. In D. Bone et al. [12], researchers have failed to produce findings utilizing larger, more balanced data that are equivalent to those reported by Wall and colleagues.

In A. S. Heinsfeld et al. [13], three algorithms are used (SVM, RF, and DNN). The SVM algorithm achieved the highest accuracy of 0.65. The 1-way approach significantly increased the accuracy from 54.1% to 90.2%. (B. van den Bekerom [15]) Further investigation is necessary in this area because of this and the fact that the severity was determined only by the opinions of the children's caregivers. The results shown by W. Liu et al. [16] are encouraging for the use of a machine learning system based on facial scan patterns to identify children with autism, with a maximum classification accuracy of 88.51 when using the SVM algorithm. The proposed system, on the other hand, uses a different method. The proposed system achieved the best accuracy ratio, as we can see in Section 6.

3. Data Set

The dataset used in this research was obtained from the UCI machine learning repository at https://archive.ics.uci.edu/ml/. In addition, there is the dataset known as PHP, which is open source. We used two different types of datasets: The Adolescent Dataset for Autism Examination consisted of a total of 104 observations, each displaying 21 features. And the Children's Dataset for Autism Examination had a total of 292 observations organized into 21 categories. The final score consists of both attributes and the attribute score integer, which is generated by the scoring algorithm for the sorting approach that was used. Data can also be used by dividing the data into two groups, educating users, and performing mining model estimation testing. Our data collection can be divided into two different categories. There is a practice test and a test exam that you can choose to take. Before using it on recent data, the model first considers the result of a well-known training set and takes it into account. After training with the data from the training set, the model is validated by making predictions based on the data from the test set. As a result, rotation estimation is a method that can be used to assess how well the classifier is doing as it sorts new cases related to the task at hand. When the turnover estimation is performed again, the data sample is first divided into two complementary subsets. Then the classifier is trained on one of the subsets (referred to as the training set), and its effectiveness is evaluated using the other (the test set). The attributes and their descriptions displayed in Table 1 are below.

Table1: Characteristics of the ASD illness

Attribute	Kind	describing
Age	Numbers	Years
Sex	String	Boy or Girls
Social Group	String	Listing for widespread Ethnicities in txt
Jaundice since birth	Boolean, Yes or No	If patient was born with With jaundice
The member of a family with PDD	Boolean, Yes or No	If anyone belongs directly to the family
Who finishes the testing	String	Father or mother self or Self, paid helper doctor And medicine –related People, etc.
Place of residency	String	A listing of every country in txt
Utilize the application designed to screen Previously	Boolean, Yes or No	If the using person utilized an application to screen
The way of screening kind	A whole number	Every kind of screen Selected built on the group age (zero=infant,
		One=kind, two= teenager, three=grown-up
Question one answer	Binary	Question answer code Built on the way utilize to screen
Question two answer	Binary	Question answer code built on the way utilize to screen
Question three answer	Binary	Question answer code built on the way utilize to screen
Question four answer	Binary	Question answer code built on the way utilize to screen
Question five answer	Binary	Question answer code built on the way utilize to screen
Question six answer	Binary	Question answer code built on the way utilize to screen
Question seven answer	Binary	Question answer code built on the way utilize to screen
Question eight answer	Binary	Question answer code built on the way utilize to screen
Question nine answer	Binary	Question answer code built on the way utilize to screen
Question ten answer	Binary	Question answer code built on the way utilize to screen
Score of screening	Whole number	The final score gained built on the algorithm of Scoring of the screening way utilize. it has been calculated in an automatic manner

4. **Data mining classification**

Initially, some training data are used to construct a model, and then, in the second step, that model is put to use to assign a class label to an unknown tuple [21]. The first step involves the construction of the model based on the training data. To generate detection

models from the dataset, two data mining algorithms called logistics regression (LR) and stochastic gradient descent (SGD) were applied.

4.1 Logistics Regression (LR)

Applications of machine learning make extensive use of the technique known as logistic regression (LR). This model evaluates a set of inputs, calculates coefficients or weights for each variable, and then predicts the classification of the tweet in the form of a word vector. In mathematical terms, the logistics regression function is responsible for providing an estimate of the number of linear functions. Assumptions made for the logistics regression:

• The results of a logistic regression do not demonstrate a linear connection between the variables being studied (dependent and independent).
• The dependent variable needs to be binary.
• It is necessary for the independent variable to be linearly correlated and not to have a normal distribution with the same variance for each group.
• Mutual exclusion between the groups is required [22].

To identify the actual class label, the logistic regression model calculates p for a linear combination of independent factors.

As seen in Eq. 1, the computed regression model can be represented [23].

$$p^{\wedge} = \frac{e^{\beta 0+\beta 1x1}}{1+e^{\beta 0+\beta 1x1}} \ldots \quad (1)$$

Algorithm (1): Logistics regression [23]
Input: training and testing data
Algorithm (1): Logistics regression
Input: training and testing data Output: ASD prediction
Begin 1. For j= 1 to k 2. For every occurrence of training data Ei 3. Set the regression's goal value to 4. Ci ← $\frac{yj - P(1\|Ei)}{[P1\|Ej).(1-P(1\|dEj)]}$ 4. initialize the instance weight Ej to P(1|Ej) (1-P) (1|Ej) 5. Finalize a s(j) to the class-valued data (Ci) & weight (wj) Classification Label Decision. 6. Assign (class label :1) if P (1|Ej) > 0.5 otherwise (class label :2) End

Figure 1: Pseudocode for Logistics Regression

4.2 The Stochastic Gradient Descent Algorithm (SGD)

According to [10], gradient descent is a frequent strategy that has the potential to provide a new perspective on the process of problem-solving. An approach for minimizing functions is gradient descent. When given a function and a set of parameters, the gradient descent algorithm will iterate over a collection of parameter values until it finds the point on the function that is lower than any other point. Given the initial values for the parameters, the SGD procedure starts. To obtain an aligned line that moves closer and closer to the minima, the minimization strategy makes use of calculus derivation. When working with especially large datasets, the gradient descent algorithm may move at a reasonable pace.

When there are millions of examples, one iteration of the gradient descent approach includes making a forecast for each occurrence in the training dataset. This could take some time. The updated coefficient methodology for SGD is the same as that for gradient descent, with the exception that the cost is only calculated for one training pattern rather than the sum of all training patterns. SGD is a slightly different algorithm than gradient descent because the coefficient update is only performed during the training procedure.

The stochastic gradient algorithm process is to pick to minimize J (a search technique is used to make an initial prediction for and then alter the value of to keep the output from J as little as possible). The following is the formula for the updating process that is repeated in SGD [6]:

$$\Theta j := \theta j - a \frac{\partial}{\partial \theta j} j(\theta) \quad (2)$$

The learning rate is the pace at which every j number between 0 and n is updated at the same time. A partial derivative is used on the right side. In the event of a single training example, the entirety of J's definitions may be ignored (x,y). It may be obtained using the power rule or chain rule.

$$\frac{\partial}{\partial \theta j} j(\theta) = \frac{\partial}{\partial \theta j} \frac{1}{2} ((h_\theta(x) - y))^2 \quad (3)$$
$$= 2 \frac{1}{2} (h_\theta(x) - y) \frac{\partial}{\partial \theta j} (h_\theta(x) - y)$$
$$= (h_\theta(x) - y) \frac{\partial}{\partial \theta j} i \sum_{i=0}^{n} \theta_i x_i - y)$$
$$= (h_\theta(x) - y) xi$$

The learning rate can be defined as the speed at which all j numbers in the range from 0 to n are simultaneously updated. On the right-hand side, only partial derivatives are employed in the case of a single demonstration. It is possible that not all of J's definitions will be taken into consideration (x, y). Both the power rule and the chain rule can be utilized to get this result

$$\theta j := \theta j + (y^{(i)} - (h_\theta(x^{(i)})) x_j{}^{(i)} \quad (4)$$

Algorithm (2): Stochastic Gradient Descent(SGD)
Input: training and testing data
Output: ASD prediction
Begin: 1. Initialize s: =0^{n-1} r: =0. 2. for iteration h∈ [1, ..., H]h∈ [1, ..., H]: • draw random example with replacement: ⟨x[i], y[i]⟩∈ D. • compute loss L[i]: = L(y^[i], y[i]). • compute gradients Δ s: =−∇L[i]s, Δr: = −∂L[i] /∂r. • update parameters s: =s+Δs,r:=+Δr End.

Figure 2: Pseudocode for Stochastic Gradient Descent (SGD)

4. Classification Model Evaluation

Estimation using data mining algorithms, the unique compromise level between genuine positive and true negative ratios, as well as remembering and precision, recalling, and F-Measure are common methods for retrieving data and doing measurements. important to

achieve it most efficiently. In the following subsections, the whole system is described by viewing its structure and modules' tasks [24] [25].
Calculating such metrics necessitates constructing the confusion matrix, which is a matrix that allows for the division of test instances into two kinds. as shown in Table 1.

Table 2: Matrix of Confusion

	Positive	**Negative**
Positive	TP	FN
Negative	FP	TN

True Positives (TP): refers to the number of data rows inside a set of tests that either already had a positive target or were anticipated to already have a positive target.
True Negatives (TN): refers to the number of data rows in a set of tests that had a negative aim or were predicted to have a negative target. This can be a positive or negative figure.
False Positives (FP): The number of data rows in a set of tests that started with a negative aim but were expected to end up with a positive target.
False Negative (FN): The number of data rows in a set of tests that had a positive goal but were projected to have a negative target is called.
[26] noted that recall, accuracy, and the F-measure are among the most commonly used measurements. They can be determined with equations (5, 6, and 7). This statistic can supplement the ratio-based misclassification measurement.

$$\text{Recall} = TP/\ TP+FN... \quad (5)$$

$$\text{Precision} = TP/\ TP+FP... \quad (6)$$

$$\text{F1-measure} = 2^* \ precision * recall \ / \ precision + recall... \quad (7)$$

MAE and RMSE are also important parameters for defining system errors. (MAE) is a statistical metric for quantifying model mistakes that is used as a benchmark. Root Mean Square Error (RMSE) is a useful statistic for measuring the mistakes of a model's evaluation [26]. More to add: MAE and RMSE are also important points to consider when defining errors in the system, as shown in the equation below:

$$\text{MAE} = \frac{1}{n}\sum_{1=1}^{n} /ei/ \quad \ldots\ldots (8)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i}^{n} ei^2} \quad \ldots\ldots (9)$$

4.1 Cross-Validation

Cross-validation is a technique used to evaluate a classifier's performance as it categorizes new examples of the job at hand. A repeated cross-validation involves splitting a data sample into two complementary subsets, training the classifier on one of them (referred to as the training set), and evaluating how well it performs on the other (referred to as the testing set). In k-fold cross-validation, the actual data set can first be partitioned into k folds or segments of about equal size. As a result, each repeat now includes k repetitions of exercising and validating. The residual k-1 folds have been used to train, while a different data fold has been maintained for validating. 10-fold cross-validation in data mining (k = 10), which serves as a standard procedure for doing estimation and model selection, can be regarded as the most popular method [27] [28].

4.2. The Definition of EDASD

Based on a series of questions posed to people in charge of the patient, the primary architecture of the proposed system focuses on offering quick and accurate guidance for diagnosing autism, thereby saving time in diagnosing this disorder. This disease can be diagnosed early using a method that makes treatment as simple and convenient as possible**.**

4.3. The Proposed System's temple

The organizational structure of the suggested system (EDASD) is explained in Figure 1.

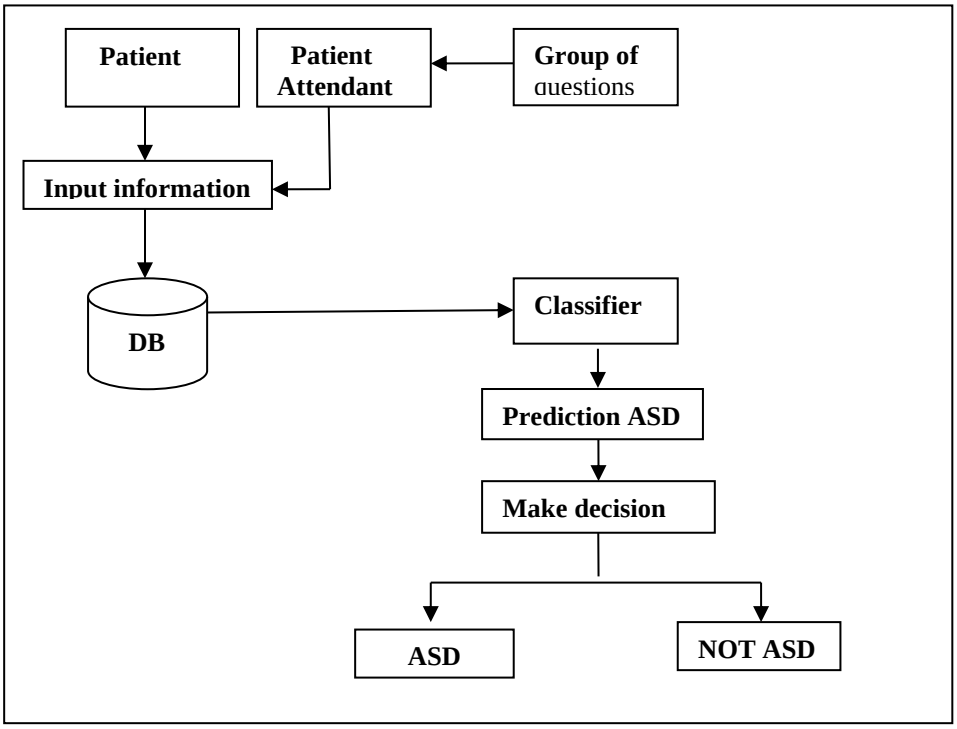

Figure 3: Block diagram of the proposed EDASD system

The process of gathering information about patients is the initial stage of the system, which consists of several distinct stages in total. This material is broken up into multiple pages, the first of which gives information about the patient. The subsequent page has 16 questions regarding the behavior of the autistic person while they are doing a screening exam. After all of the fields have been filled out, the relevant information will be saved in the databases. The second part of the system is called the classifier, and it is the location where the data is moved from the databases to the workbook. The workbook is the substantial part of the system that is responsible for classifying the information. The stochastic gradient descent classifier takes in the data and decides whether or not the patient has autism spectrum disorder (ASD) based on the results. The completion of this task triggers a message to be shown in the system, and it is also logged into a database file at the same time.

5. The Experiments Result

With two types of datasets, one for children and one for adolescents, the effectiveness of our proposed system was evaluated, and they were selected from the UCI repository. Comparing the numerical results of the data mining algorithms used in the ASD dataset is the first step. Table 3 presents the results of the children's dataset in terms of accuracy, retrieval, F-measurement, MAE, and RMSE. The data set for adolescents is presented in Table 4. Finding the best model is critical, but it is even more important to recognize which models

may not be the best choice. The results unequivocally show that the SGD algorithm outperformed the LR algorithm in detecting ASD, with the highest accuracy and lowest error rates for pediatric and adolescent datasets. Because accuracy is meaningless if accuracy and retrieval are not balanced, an F-measure was used to compare experimental results for each classification method. The value ß = 1 is used as a parameter to balance retrieval and accuracy since the goal is to increase both recall and accuracy [29] [30]. We have to use the SGD classifier as a classifier in our proposed system because of this.

Table 3: Performance results of the LR and SGD algorithms for the children's dataset

Measure	LR	SGD
Precision	0.904	0.939
Recall	0.952	0.96
F-measure	0.952	0.964
MAE	0.047	0.0308
RMSE	9.055	6.171

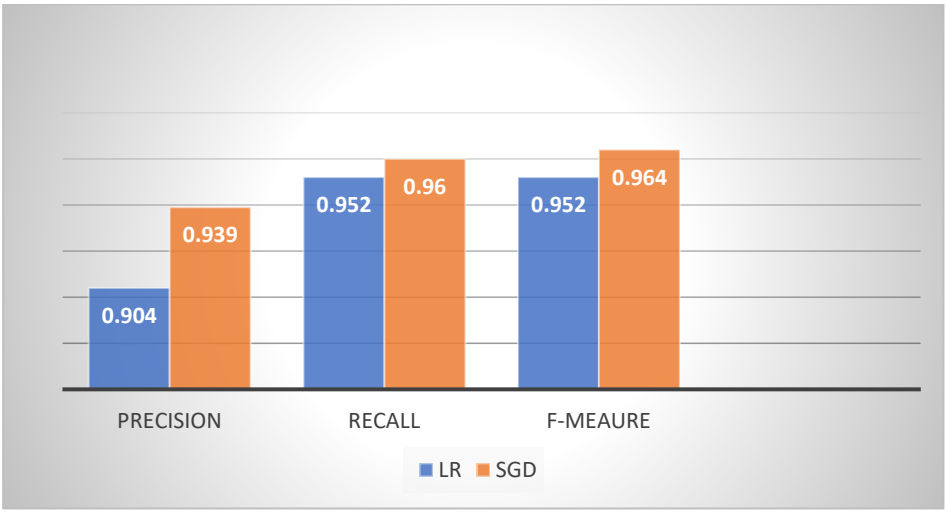

Figure 4: Algorithm Performance Measures for the Children's Dataset

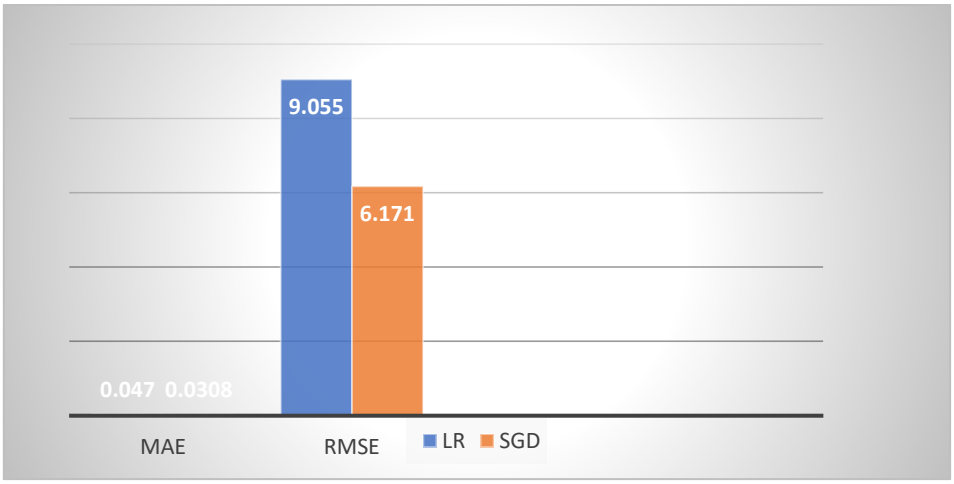

Figure 5: MAE, RMAE using (LR, SGD) for children dataset

Table 3, Figure 4, and Figure 5 present the application of two data mining techniques, namely LR (logistic regression) and SGD (random gradient descent), to the children's dataset. The proposed model was trained using a children's dataset to predict whether or not a child has autism spectrum disorder based on the provided dataset. In addition, the performance evaluation of the model is based on the previously mentioned metrics. Thus, after training, the SGD model achieved better results compared to the LR model. The precision value of the SGD model indicates that the model has a high proportion of true positive predictions compared to false positive predictions and a low portion of MAE and RMSE. In other words, the model is accurate in identifying positive cases and has a low rate of incorrectly classifying negative cases as positive. The recall value of the SGD model indicates that the model is effective in capturing positive cases, reducing false negatives, and ensuring a high level of sensitivity in detecting the positive category.

Table 4: Results of the (LR, SGD) algorithms' performance on the adolescents' dataset

Measure	LR	SGD
Precision	0.917	0.952
Recall	0.939	0.947
F-measure	0.948	0.96
MAE	0.048	0.0309
RMSE	9.057.	6.173

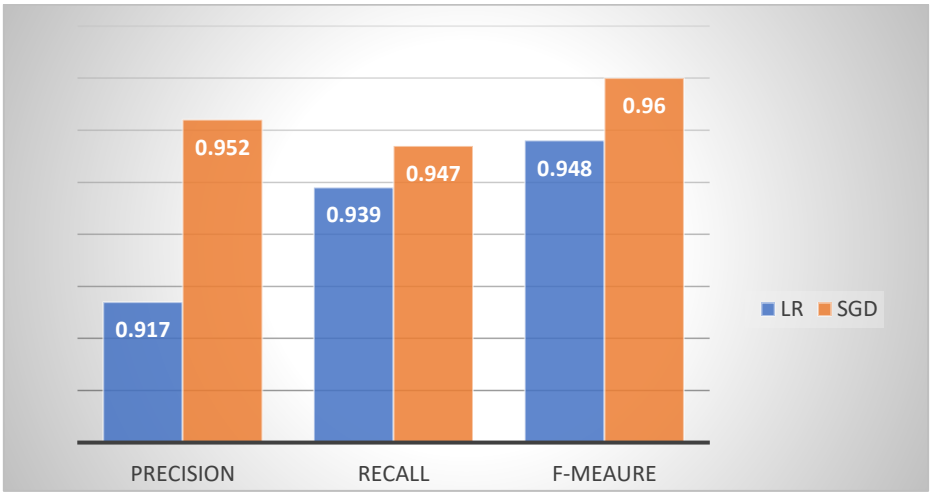

Figure 6: Algorithm Performance Measures for the Adolescent Dataset

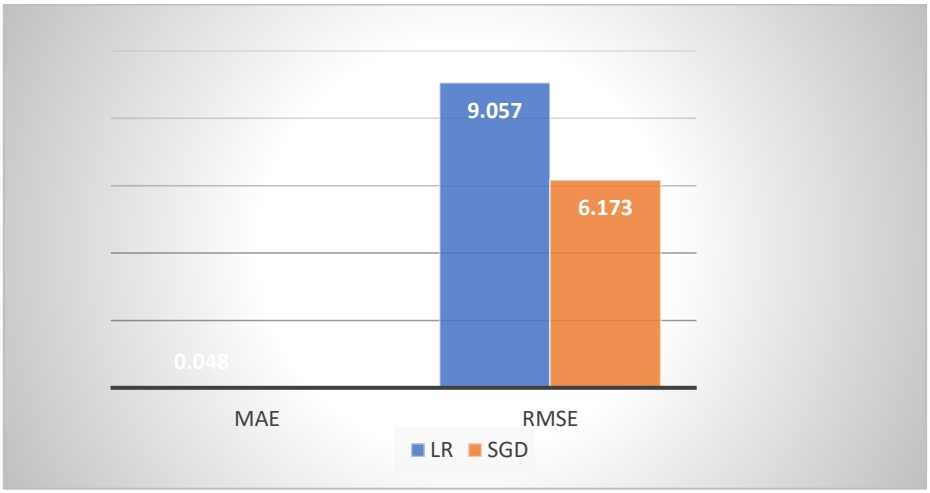

Figure 7: MAE, RMAE using (LR, SGD) for adolescents' dataset

As shown in Table 4, Figure 4, and Figure 5, a new adolescent dataset was used to validate the results obtained from the children's dataset, and two data extraction techniques, namely LR (logistic regression) and SGD (random gradient descent), were applied to the adolescent dataset.

The proposed model was trained using the adolescent dataset to predict whether an adolescent has autism spectrum disorder based on the provided dataset. In addition, the performance evaluation of the model is based on the previously mentioned metrics. Thus, after training, the SGD model achieved better results compared to the LR model. The precision value of the SGD model indicates that the model has a high proportion of true positive predictions compared to false positive predictions and a low fraction in MAE and RMSE. In other words, the model is accurate in identifying positive cases and has a low rate of incorrectly classifying negative cases as positive. The recall value of the SGD model indicates that the model is effective in capturing positive cases, reducing false negatives, and ensuring a high level of sensitivity in detecting the positive category.

8. Conclusion and Future Work

There has been a suggestion made here regarding a method for predicting autism. Using two different kinds of data mining techniques and sets of information from the UCI warehouse, an approach to verifying the suggested system has been put through its paces and found to be effective. Our method, which is based on the empirical results acquired from each piece of data, could be regarded as an alternative method of categorizing the data. This strategy enables clinicians to discover autism conditions with a reduced number of symptoms, which will prevent the case from developing further and will lower the expenditures that must be paid as a result of the delayed diagnosis. The autism spectrum disorder (ASD) condition is the most common and has a negative impact on children's social interaction, communication, and learning abilities. A person may be identified with autism at an earlier stage if an autism detection system is utilized, which stops the condition from getting worse and reduces the costs associated with a delayed diagnosis. For this article, only the pediatric and adolescent data sets were utilized; however, other studies, such as those in the future scope of work, may be organized using adult data and the utilization of algorithms that are capable of achieving better classification accuracy than the algorithms that are now being utilized by this system, such as deep learning algorithms, in conjunction with more precise standards.

Author Contributions : Conceptualization, reviewing and editing: Asha Hani Salman ; methodology, Rasha Hani Salman , Manar Bashar : investigation Manar , formal analysis: Riydh Rahef Nuiaa; All authors read and approved the final manuscript.

References

[1] M. S. Mythili and A. R. M. Shanavas, "A Study on Autism Spectrum Disorders using Classification Techniques," *Int. J. Soft Comput. Eng.*,vol.4 no. 5, pp. 2231–2307, 2014.

[2] R. A. Musa, M. E. Manaa, and G. Abdul-Majeed, "Predicting Autism Spectrum Disorder (ASD) for Toddlers and Children Using Data Mining Techniques," *J. Phys. Conf. Ser.*, vol. 1804, no. 1, pp. 0–8, 2021.

[3] J. H. Miles, "Autism spectrum disorders—a genetics review," *Genet. Med.*, vol. 13, no.4, pp. 278–294, 2011.

[4] S. Fuld, "Autism spectrum disorder: The impact of stressful and traumatic life events and implications for clinical practice," *Clin. Soc. Work J.*, vol. 46, no. 3, pp. 210–219, 2018.

[5] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *Lancet*, vol. 392, no. 10146, pp. 508–520, 2018.

[6] S. Ray, "A Quick Review of Machine Learning Algorithms," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Perspectives Prospect. Com.*, 2019.
[7] G. Lai, F. Li, J. Feng, S. Cheng, and J. Cheng, "A LPSO-SGD algorithm for the Optimization of Convolutional Neural Network," *2019 IEEE Congr. Evol. Comput. CEC 2019 - Proc.*, 2019.
[8] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inform.*, vol. 2, pp. 59-77, 2006.
[9] D. P. Wall, R. Dally, R. Luyster, J.-Y. Jung, and T. F. DeLuca, "Use of artificial intelligence to shorten the behavioral diagnosis of autism," *PLoS One*, vol. 7, no. 8, pp. 1–8, 2012.
[10] C. Allison, B. Auyeung, and S. Baron-Cohen, "Toward brief 'red flags' for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 51, no. 2, pp. 202–212, 2012.
[11] D. P. Wall, J. Kosmicki, T. F. Deluca, E. Harstad, and V. A. Fusaro, "Use of machine learning to shorten observation-based screening and diagnosis of autism," *Transl. Psychiatry*, vol. 2, no. 4, pp. e100–e100, 2012.
[12] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: pitfalls and promises," *J. Autism Dev. Disord.*, vol. 45, pp. 1121–1136, 2015.
[13] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *NeuroImage Clin.*, vol. 17, pp. 16–23, 2018.
[14] N. S. Khan, M. H. Muaz, A. Kabir, and M. N. Islam, "Diabetes predicting mhealth application using machine learning," in *IEEE international WIE conference on electrical and computer engineering (WIECON-ECE)*, 2017.
[15] B. van den Bekerom, "Using machine learning for detection of autism spectrum disorder," in *Proc. 20th Student Conf. IT*, 2017.
[16] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Res.*, vol. 9, no. 8, pp. 888–898, 2016.
[17] F. Thabtah, "Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment," in *Proceedings of the 1st International Conference on Medical and health Informatics*, 2017.
[18] F. Hauck and N. Kliewer, "Machine learning for autism diagnostics: applying support vector classification," in *Int'l Conf. Heal. Informatics Med. Syst*, 2017.
[19] D. Bone, S. L. Bishop, M. P. Black, M. S. Goodwin, C. Lord, and S. S. Narayanan, "Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion," *J. Child Psychol. Psychiatry*, vol. 57, no. 8, pp. 927–937, 2016.
[20] J. A. Kosmicki, V. Sochat, M. Duda, and D. P. Wall, "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning," *Transl. Psychiatry*, vol. 5, no. 2, pp. e514–e514, 2015.
[21] S. M. Gorade, A. Deo, and P. Purohit, "A study of some data mining classification techniques," *Int. Res. J. Eng. Technol.*, vol. 4, no. 4, pp. 3112–3115, 2017.
[22] C.-S. Rau *et al.*, "Machine learning models of survival prediction in trauma patients," *J. Clin. Med.*, vol. 8, no. 6, p. 799, 2019.
[23] V. Kumar, "Evaluation of computationally intelligent techniques for breast cancer diagnosis," *Neural Comput. Appl.*, vol. 33, no. 8, pp. 3195–3208, 2021.
[24] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv Prepr. arXiv2010.16061*, 2020.
[25] R. H. Salman, N. A. Shiltagh, and M. Z. Abdullah, "Development of a Job Applicants E-government System Based on Web Mining Classification Methods", *Iraqi Journal of Science*, vol. 62, no. 8, pp. 2748–2758, Aug. 2021.
[26] J. D. Kelleher, B. Mac Namee, and A. D'arcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, 2020.
[27] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Stat. Surv.*, vol. 4, pp. 40–79, 2010.

[28] D. Berrar, "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. April, pp. 542–545, 2018.

[29] N. Jumaa, A. Salman, and D. R. Al-Hamdani, "The autism spectrum disorder diagnosis based on machine learning techniques," *J. Xi'an Univ. Archit. Technol*, vol. 12, pp. 575–583, 2020.

[30] B. Juba and H. S. Le, "Precision-recall versus accuracy and the role of large data sets," in *Proceedings of the AAAI conference on artificial intelligence*, 2019