Proceedings of the Sixth International Conference on Trends in Electronics and Informatics (ICOEI 2022)
IEEE Xplore Part Number: CFP22J32-ART; ISBN: 978-1-6654-8328-5

A Comparative Analysis of Prediction of Autism Spectrum Disorder (ASD) using Machine Learning

Vaibhav Vishal
Student
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
vaibhavvishal85982@gmail.com

Abhishek Singh
Student
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
linktoabhi7@gmail.com

*Y.Bevish Jinila
Associate Professor
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
ybevish@gmail.com

Kavitha.C
Assistant Professor
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
kavitha4cse@gmail.com

S.Prayla Shyry
Associate Professor
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
praylashyry@gmail.com

J.Jabez
Associate Professor
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
jabezme@gmail.com

2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) | 978-1-6654-8328-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOEI53556.2022.9777240

***Abstract*—The Autism Spectrum Disorder (ASD) is a neurological disease, which affects the mental, social and physical state of a person. A person of any age group can be found infected by it. It is very difficult to identify, if a person is the victim of this disorder. Classical approaches that find the occurrence of autism in a person is time consuming and expensive. Machine learning approaches, on the other hand, have paved the way for intelligent diagnostics. This paper focusses on identification of specific traits that helps to automate the diagnosis process and further evaluate and perform a comparative analysis of various machine learning algorithms namely K-Nearest Neighbour, Logistic Regression, SVM and Naïve Bayes, to predict the occurrence of autism disorder. Experimental analysis shows that the Naïve Bayes algorithm provides a better accuracy of 99.6% compared to other algorithms.**

Keywords—Autism Spectrum Disorder, Logistic Regression, Machine Learning, Naïve Bayes, SVM, KNN

I. INTRODUCTION

Brain is the most important organ in human body that coordinates various parts of the body. Disassociation of nerves in brain and disruption in brain development is the cause of Autism. Autism can be identified at any age, although it typically begins in childhood and is classified as a developmental condition. People with ASD finds difficult in collaborating and networking with others [1]. They have a limiting method of interest, and their habits may be repetitious, affecting their everyday routine existence." World Health Organization [2], states one in hundred and sixty children are affected by autism. With respect to this illness, some of them will be able to live independently, while some others will require care and help for the rest of their lives.

Detection of the presence of autism is time consuming. Early detection of autism is advantageous in terms of providing patients with suitable medication. Early prediction of the disease helps to avoid the patient's health from degrading further and unwanted treatments can be avoided and cost can be reduced.

This paper aims to investigate whether a person have ASD or not by using classification methods and also find out the classification algorithm which is best suitable for predicting ASD based on performance measurements such as accuracy and error rate. The Autism dataset collected are undergone through pre-processing technique like conversion of String to numerical data using Label Encoder and One Hot Encoder and then mean technique for cleaning the data. Training data is used to recognize patterns in the data, and testing data used to verify the data. Then important features are selected using feature selection.

The act of the classification algorithm is evaluated using parameters such as Accuracy, and error rate. Finally, the classification algorithm with high performance values are considered as the most suitable classification algorithm for ASD. After applying different classified model for the prediction of Autism, it is observed that Naïve Bayes is found to be producing much better results than the other algorithms.

The remainder of the paper is laid out as follows. Section II briefly describes the existing work of autism disorder. Section III describes the proposed methodology. Section IV discusses the experimental results and analysis. Section V details the conclusion of the proposed work and further discusses on the future work.

II. RELATED WORK

This section lists out the related work of autism disorder based on machine learning approaches.

In the article by Sunsirikul et al. [3] states, the planned program's goal is to create a data analysis tool that will aid physicians in future diagnostics. Attempts were made in this study to extract patterns from behavioral data and create a behavior code for patients. It is possible to detect a correlation between certain behaviors and autism symptoms if there are enough patient behavioral records. This paper explores data mining techniques with the goal of providing a set of tools to aid physicians in intelligently evaluating patient data. The benefit of this work is that it shows the association between autistic children's behavioral patterns and PDD NOS [3]. With a high level of self-esteem, a set of disabilities, as well as the type of disorder, can be improved. Because a small number of samples are frequently overloaded with the solution, the predictor error can occur in some instances. Lack of clinical data for normal children to be used in the training phase.

Authorized licensed use limited to: Amrita School of Engineering. Downloaded on September 06,2023 at 06:10:20 UTC from IEEE Xplore. Restrictions apply.

In 2016, Osman, et al. [5] proposed a method using data mining techniques. This proposed system is utilized to diagnose an Autism patient. Autism Spectrum Disorder (ASD) has a negative impact on a person's health. Lack of social engagement and communication, repetitive behaviors, and engrained interests and hobbies are all major markers of ASD. In this paper, classification algorithms are used to try to figure out if youngsters have ASD. "As a result of classification, a kid might be classified as having ASD or not having ASD. The LDA algorithm delivered 90.8 percent accuracy, whereas the KNN approach delivered 88.5 percent accuracy.

Cincy Raju et al. suggested that, heart disease is one of the most dangerous diseases that can lead to death. It suffers from a severe long-term impairment. This sickness strikes with such ferocity. The goal of this research is to use data mining techniques to provide an effective remedy for restorative circumstances. Many different categorization algorithms are employed. Support Vector Machine (SVM) is the finest of these methods" [4].

In [6], a machine Learning Adaption and DSM-5-based ASD screening model is used. "Erik et al. [8] proposed a management system for autism detection. The purpose of this study is to introduce, an integrated health-care system to capture, analyze, and manage data related to the diagnosis and treatment of children with Autism Spectrum Disorder. AMP is a smart web interface and statistical platform that allows doctors and specialists to collect and extract patient data in real time, as well as provide proper feedback to learn data filter preferences automatically. Similar works on prediction has been carried out in [7,9,10,11].

Canon et al. [12] has provided empirical evidence of prediction of autism disorder. In this paper, some additional evidences required are identified and presented. However, the functional impacts of the individuals have to be studied.

Karunakaran et al. [13] has presented an approach that combines adaptive functioning classifier and early learning techniques. This approach handles less noisy data which is one of the limitations. Machine learning analysis [14] and path way analysis [15] are the other two approaches proposed for prediction of autism disorder.

The existing approaches is not promising in prediction of autism disorder. Few important parameters are not considered for analysis. So, it is important to consider all essential parameters to improvise the proposed algorithm.

III. Proposed Methodology

This section demonstrates the proposed scheme for analyzing the autism disorder. The process adopts the following steps:

(a) Data Collection
(b) Data Pre-processing
(c) Construction of model
(d) Training and Testing

Fig.1 is the flow chart which is showing the process to be considered for the whole procedure.

(a) Data Collection

The dataset for predicting autism spectrum disorder is selected with 1054 no of different cases, necessary attributes of dataset are used for training our model on it. The different columns in our dataset depict different symptoms, age, family history, residence etc".

(b) Data Pre-processing

Data is usually incomplete, inconsistent, and contains a lot of errors and noisy data. Data pre-processing is a proven way to solve such problems. In this work the Data is Pre-Processed using Label Encoding and One Hot Encoding. This process is used to convert non-numerical data into numerical data.

In Autism Screening dataset pre-processing of data is done using data cleaning techniques, and then Feature selection is used for filtering irrelevant or redundant features from dataset with that features we do classification using algorithms like SVM, LDM and NAÏVE BAYES then we predict the results, and we analyze which algorithm provides better accuracy.

(c) Construction of Model

Different algorithms are used in construction of Model. Which helps us to determine the appropriate Model to be used for desired result. The algorithms used are SVM, LDM, and Naïve Bayes.

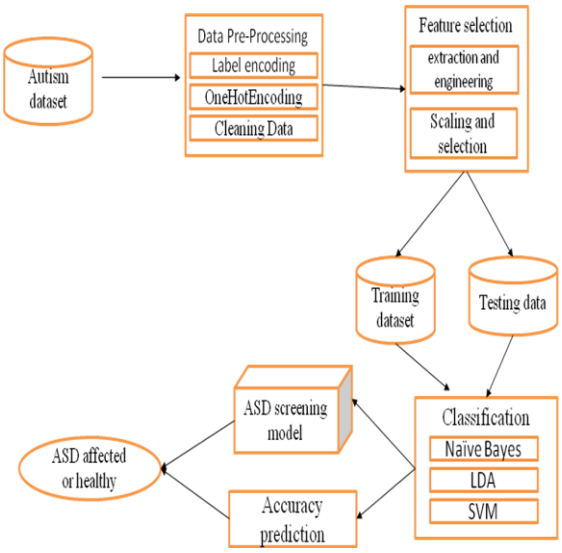

Fig.1 Proposed Process Flow

(i) LDM

LDM is a supervised machine learning approach. This method helps in predicting a categorical dependent variable from a collection of independent variables. Linear Regression and Logistic Regression are very similar in terms of how they are employed."

(ii) SVM

For dealing with unlabelled data, supervised learning is impossible, so an unsupervised technique is required, in which the data is naturally clustered into groups, and then fresh data is mapped to these created groups, which is where SVM comes in. It categorizeses unlabelled data using support vector statistics generated in the support vector

Authorized licensed use limited to: Amrita School of Engineering. Downloaded on September 06,2023 at 06:10:20 UTC from IEEE Xplore. Restrictions apply.

machines technique. The efficiency of SVM is determined on the kernel chosen, its parameters, and the soft margin parameter. A Gaussian kernel, which has only one parameter, is a popular choice. Potential drawbacks of the SVM include the following aspects:

- Input data must be fully labelled.
- Probabilities of belonging to a class that aren't calibrated SVM is based on Vapnik's theory, which eliminates the need to estimate probabilities on finite data.
- Only two-class tasks are directly applicable to the SVM. As a result, techniques must be used to reduce the multi-class task to a series of binary problems.
- It's tough to interpret the parameters of a solved model.

(iii) Naïve Bayes

Naive Bayes is a prominent method for building classifiers. The "event model" of the naive Bayes classifier is made up of assumptions about feature distributions. Multinomial and Bernoulli distributions are common for discrete features like those found in document categorization (including spam filtering). These assumptions result in two unique models that are frequently misunderstood.

Samples (feature vectors) in a multinomial event model represent the frequency with which certain events have been generated by a multinomial.

(d) Training and Testing

After selecting different models, the dataset is trained based on different Construction Model. Then the data is tested based over all the Model and one with better accuracy is considered to be used further.

IV. Experimental Results And Discussion

In this section experimental results are shown and discussed. The structure of the dataset is visualized and the performance of the algorithms is analyzed for predicting the autism disorder.

A. Dataset Description

The dataset considered for autism disorder identification has 1054 number of cases and 19 different attributes namely symptoms, age, gender, native, ethnicity etc. Fig. 2 shows the attributes available in the dataset.

```
Index(['Case_No', 'A1', 'A2', 'A3', 'A4', 'A
5', 'A6', 'A7', 'A8', 'A9', 'A10',
       'Age_Mons', 'Qchat-10-Score', 'Sex', 'E
thnicity', 'Jaundice',
       'Family_mem_with_ASD', 'Who completed t
he test', 'Class/ASD Traits '],
      dtype='object')
```

Fig. 2 Attributes

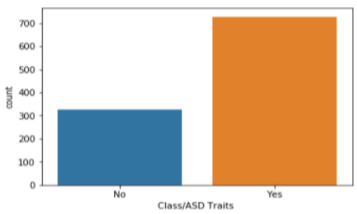

Fig.3 Count of Traits

The input dataset is preprocessed and suitable features are selected to construct the model. Before, model selection and building, the pattern of the data is understood by visualization. The fig.3 shows the count of traits that have autism or not. That is, the number of people affected and not affected by autism based on various attributes.

Fig.4 shows the correlation between the attributes in the dataset.

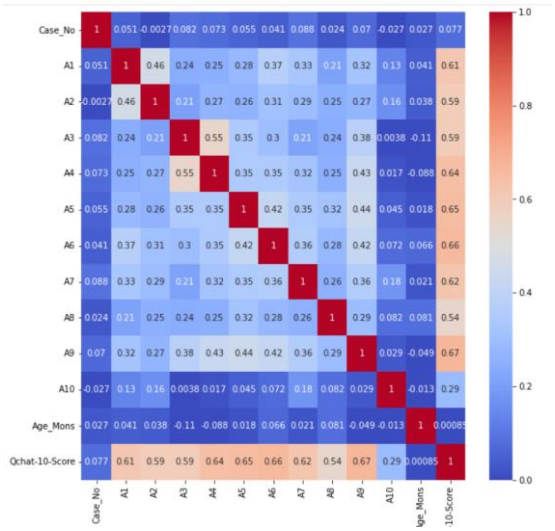

Fig. 4 Heat Map

B. Comparative Analysis

The proposed work on predicting the autism disorder is compared with various metrics like accuracy, and error rate.

Fig. 5 shows the error rate obtained in Naïve Bayes. Fig. 6 shows the error rate in SVM. Fig. 7 shows the accuracy comparison of the models constructed namely, SVM, Naïve Bayes and Logistic Regression. From the figure, it is apparent that the accuracy of prediction of autism is higher in Naïve Bayes compared to other models.

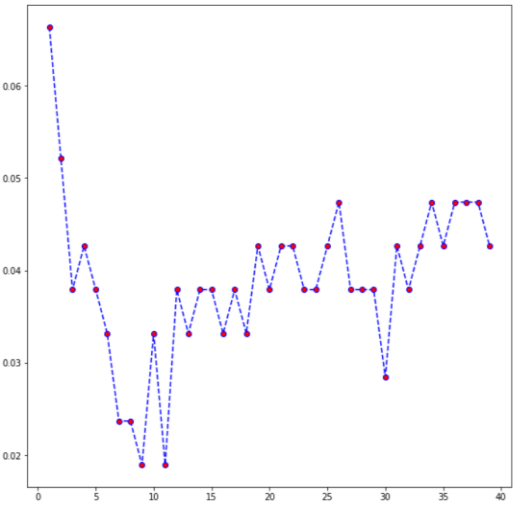

Fig. 5 Error rate in Naïve Bayes

Authorized licensed use limited to: Amrita School of Engineering. Downloaded on September 06,2023 at 06:10:20 UTC from IEEE Xplore. Restrictions apply.

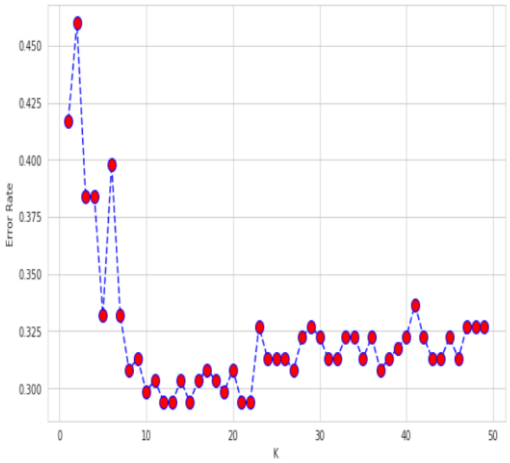

Fig.6 Error Rate in SVM

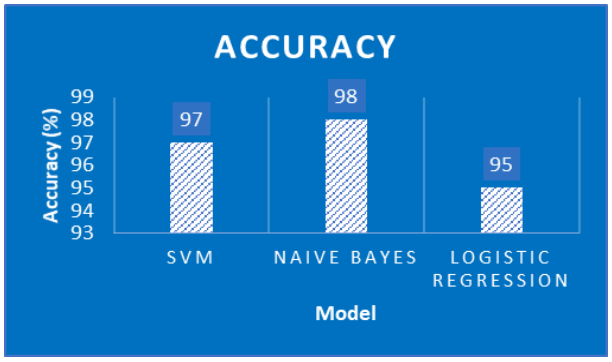

Fig. 7 Accuracy

Table 1 shows the summary of the comparison of existing and proposed approaches.

Table 1 : Comparative Analysis

Approach	Accuracy (%)	Error Rate (%)
KNN	98.2	0.06
Logistic Regression	97.2	0.07
SVM	98.4	0.05
Naïve Bayes	99.6	0.02

V. Conclusion

To predict the occurrence of autism in adults, a comparative analysis of several algorithms such as logistic regression, SVM, and Nave Bayes is undertaken in this paper. In terms of autism disorder prediction, the experimental research shows that Nave Bayes beats the other models. The system can be improved in the future to improve accuracy and reduce error rates.

References

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

[2] WHO, Autism spectrum disorders, 2017 [Accessed August 22, 2018]. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/ autism-spectrum-disorders

[3] Sunsirikul, Siriwan, and Tiranee Achalakul. "Associative classification mining in the behavior study of Autism Spectrum Disorder." In Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, vol. 3, pp. 279-283. IEEE, 2010

[4] Raju, Cincy, E. Philipsy, Siji Chacko, L. Padma Suresh, and S. Deepa Rajan. "A Survey on Predicting Heart Disease using Data Mining Techniques." IEEE Conference on Emerging Devices and Smart Systems (ICEDSS), pp. 253-255, 2018.

[5] Altay, Osman, and Mustafa Ulas. "Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children." 6th International Symposium on Digital Forensic and Security (ISDFS), 2018, pp. 1-4. IEEE.

[6] Thabtah, Fadi, and David Peebles. "A new machine learning model based on induction of rules for autism detection." Health informatics Journal 26, no. 1, pp. 264-286, 2020.

[7] D.Vaishnavi, S.Suwetha, Y.Bevish Jinila, R.Subhashini, S.Prayla Shyry (2021), " A Comparative Analysis of Machine Learning Algorithms on Malicious URL Prediction", 5th International Conference on Intelligent Computing and Control Systems (ICICCS 2021) May 6-8.

[8] Linstead, Erik, Ryan Burns, Duy Nguyen, and David Tyler. "AMP: A platform for managing and mining data in the treatment of Autism Spectrum Disorder." IEEE 38th Annual International Conference In Engineering in Medicine and Biology Society (EMBC), 2016, pp. 2545-2549.

[9] Shyry, S. P., & Jinila, Y. B., "Detection and Prevention of Spam Mail with Semantics-based text classification of Collaborative and Content Filtering", Journal of Physics: Conference Series (Vol. 1770, No. 1, p. 012031). IOP Publishing, 2021

[10] Madhukeerthana, Y. Bevish Jinila, "A review on rough set theory in medical images", Research Journal of Pharmaceutical, Biological and Chemical Sciences, Vol.7, Issue 1, pp.815-822, ISSN : 0975-8585, 2016

[11] Madhukeerthana, Y. Bevish Jinila, Deepika, "Enhanced rough set theory for denoising brain MR images using bilateral filter design ",Research Journal of Pharmaceutical, Biological and Chemical Sciences, Vol.7, Issue 3, ISSN : 0975-8585, 2016.

[12] Cannon, J., O'Brien, A. M., Bungert, L., & Sinha, P., Prediction in autism spectrum disorder: a systematic review of empirical evidence. Autism Research, 14(4), 604-630, 2016

[13] Kojovic, N., Natraj, S., Mohanty, S. P., Maillart, T., & Schaer, M., Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. Scientific Reports, 11(1), 1-10, 2021

[14] Karunakaran, P., & Hamdan, Y. B., Early prediction of autism spectrum disorder by computational approaches to fmri analysis with early learning technique. Journal of Artificial Intelligence, 2(04), 207-216, 2020

[15] Chaitra, N., Vijaya, P. A., & Deshpande, G., Diagnostic prediction of autism spectrum disorder using complex network measures in a machine learning framework. Biomedical Signal Processing and Control, 62, 102099, 2020

[16] Skafidas, E., Testa, R., Zantomio, D., Chana, G., Everall, I. P., & Pantelis, C., Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. Molecular psychiatry, 19(4), 504-510, 2014

Authorized licensed use limited to: Amrita School of Engineering. Downloaded on September 06,2023 at 06:10:20 UTC from IEEE Xplore. Restrictions apply.