Proceedings of the 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT-2024)
IEEE Xplore Part Number: CFP24CV1-ART; ISBN: 979-8-3503-2753-3

2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) | 979-8-3503-2753-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IDCIOT59759.2024.10467893

Identification of Autism Spectrum Disorder (ASD) in Adults through Various Machine Learning Algorithms

S. Yuvaraj
Department of Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India.
yuvaraj.scse@sece.ac.in

S. Sugavanaesh
Department of Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India.
sugavanaesh.s2020cse@sece.ac.in

C. Tharun Kumar
Department of Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India.
tharunkumar.c2020cse@sece.ac.in

G. Saran
Department of Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India.
saran.g2020cse@sece.ac.in

R. Subha
Department of Computer Science and Engineering,
Sri Eshwar College of Engineering,
Coimbatore, India
hodcse@sece.ac.in

***Abstract*—This research study presents a comprehensive analysis of Autistic Spectrum Disorder (ASD) in adults, addressing its prevalence, diagnostic challenges, and the diverse range of support and intervention strategies available. The study analyzes the growing recognition of ASD as a lifelong condition and explores the unique social, emotional, and cognitive needs of adults on the spectrum. Our study also considers the value of early detection and treatment in enhancing the quality of life for adults with Autism Spectrum Disorder (ASD). Datasets with behavioral aspects are now more important than ever because of the global increase of ASD incidence. However, the lack of such datasets makes it difficult to perform in-depth analysis in order to improve the effectiveness, sensitivity, specificity, and prediction accuracy of ASD screening processes. Clinical or screening-oriented datasets for autism are now noticeably lacking. To fill this gap, this study proposes a new dataset that includes 20 unique traits, which was created exclusively for screening ASD in adults. This dataset includes ten behavioral characteristics assessed using the AQ-10-Adult and an additional ten personal characteristics that, according to behavioral science specialists, have shown useful in accurately differentiating people with ASD from control participants. This dataset shows potential for additional study, particularly in identifying important autistic characteristics and improving patient classification for ASD.**

Keywords—Autism, Machine Learning, Autism Spectrum Disorder (ASD), Adults, accuracy, screening

I. INTRODUCTION

Autism spectrum disease (ASD) is a neurological and developmental ailment that affects how people connect with others, communicate, learn, and behave. Autism is classified as a "developmental disorder" since symptoms commonly appear in the first two years of life, despite the fact that it can be diagnosed at any age [5] [12]. Autism spectrum disorder (ASD) is becoming more common in persons of all ages. Recent studies have shown how crucial it is to comprehend and assist people with ASD throughout their lifetimes, despite the disorder traditionally being thought of as a childhood issue. While there has been substantial improvement in the early identification of ASD in children, adults on the autistic spectrum have received less attention [3]. Through an examination of the intricacies of ASD in the adult community, this essay seeks to close this gap. It is crucial to identify ASD in its earliest stages since doing so can result in prompt treatments and assistance, which greatly enhance a person's quality of life [9]. Machine learning has become a useful method for diagnosing ASD at the earliest stages, particularly in youngsters. However, due to the variety of the adult autistic community, using machine learning techniques to adults offers particular difficulties. Furthermore, improvements in the identification of neurological disorders utilizing methods like electroencephalogram (EEG) signals have shed light on the potential for objective metrics in the diagnosis and comprehension of neurological illnesses [13][14]. Entropy analysis is used to automate the diagnosis of neurological abnormalities, which paves the way for possible applications in the detection of ASD, especially in adults who may show small but substantial changes in EEG patterns. [11] [16] Taking into mind the particular needs of adults with ASD, a machine learning-based categorization technique for adult autism spectrum disorders that takes into account the variances in symptomatology and life events that distinguish this group [3][7]. This study builds on previous research while focusing on the particular problems and opportunities associated with ASD in adults. In this article, we want to present a thorough review of ASD in adults, covering its prevalence, diagnosis, and the variety of support and intervention techniques available. We want to add to the growing pool of information around ASD in adults by expanding on the basis of machine learning and EEG analysis and to promote a more inclusive and personalized approach to diagnosis and management [3].

II. Related Works

S.M. Mahedy Hasan, etal using multiple Machine Learning approaches, developed a viable framework for early identification of ASD. For determining the optimum Machine Learning approach, they offered four distinct Feature Scaling Strategies. They developed an efficient approach using eight different Machine Learning algorithms. [1]

S. Janifer Jabin Jui, etal focuses on the use of entropy in the analysis of electroencephalogram signals for the diagnosis of neurological illnesses. Electroencephalogram (EEG) signals are processed using their way of applying entropy. They want to know which machine learning approaches and effective entropies produce the greatest results. Multiscale and Shannon entropies are the most often used entropies for ADS.[2]

Nurul Amirah Mashudi, etal presented a system for classifying adults with ADS. A dataset with 16 libraries and 703 instances was employed. Seven of the top and best Machine Learning approaches are used in the testing, which are conducted in a simulated setting. Out of the seven Machine Learning models studied, the study found that only five models have achieved 100% accuracy.[3]

Jungpil Shin, etal researched how Handwriting analysis was used to identify ADHD in individuals with ASD. In this study, drawings or handwriting patterns are employed as a dataset. A raw dataset of 29 handwriting examples of Japanese children was compiled. Using Sequential Forward Floating Search (SFFS), they derived 30 statistical features from this raw dataset. These properties were given to seven machine learning-based algorithms to train on. They eventually developed an RF-based classifier algorithm with a high accuracy of 93.10%.[4]

Suman Raj and Sarfaraz Masood proposed that CNN is more effective for ADS among various Machine learning techniques. They took three different datasets which consisted of 63 attributes and 1100 instances in total and fed this dataset into six various Machine Learning techniques. Their research indicates that CNN-based models produce the best results. Despite the excellent accuracy the SVM model provides in one dataset. However, the CNN-based model performs better in all three datasets.[5]

Charlotte Kupper etal They proposed a method for detecting ADS in a clinical sample of individuals. They assembled a dataset of approximately 1300 adult examples by hand. Using the machine learning technique Support Vector Machine (SVM), they discovered five behavioural features in the sample. The outcome is shown to be the same as the existing ADS diagnosis for children.[6]

Astha Baranwal and Vanitha M researched about ADS screening with help of various Machine Learning models. They used three different ADS screening dataset for different cases like children, adults and adolescents. They used five different Machine Learning models for prediction. According to their research, ANN provides best results for adult ADS, logical regression provides best for adolescent ADS and SVM provides best for child ADS. However, they can't provide a clear solution, because their dataset is really small to derive an effective model.[7]

Mochammad Farrell, etal categorize the ADS, a combination of the Firefly Algorithm and Random Forest was designed.[17][18] RF often yields great accuracy but a low F1-score. They chose the Firefly algorithm over the RF algorithm because it is more accurate. They demonstrated that FARF outperforms the original RF. FARF achieves 90.78% accuracy and an F1-score of 34.09% when compared to RF. RF has an accuracy of 94.32% and a F1-Score of 35.67%. [8]

Devika Varshini G and Chinnaiyan R proposed an idea for optimized Machine Learning Classification approaches for ADS prediction.[9] They passed the dataset across multiple Machine Learning algorithms like Classification, Logistic Regression, KNN, Random Forest.[19][20][21] Finally they got a maximum accuracy for KNN as 69.12%, logical regression as 68.60% and random forest classifiers as 67.78%. From their result they conclude that KNN has the highest accuracy which is calculated as the experimental results.[10].

III. Methodology

A. Dataset Description:

Our dataset, which was obtained from the UC Irvine Machine Learning Repository, contains every required data to detect ASD. The feature types, which cover 20 features and include integers, Boolean, floats, and objects, have about 704 instances. We can use this information to detect ASD class or non-ASD class.

Twenty properties shared by these databases are used to make predictions. These qualities include the following: Table 1 shows Dataset description

TABLE 1: DATASET DESCRIPTION

Instance Id	Instance Description	Description Type
1	Age	Number (in years)
2	gender	String
3	Nationality	String
4	Jaundice by birth	Boolean (yes or no)
5	Family member with PDD	Boolean (yes or no)
6	Who is fulfillment the experiment	String
7	The country in which the user lives	String
8	Used the screen app before	Boolean (yea or no)

9	Screening method type	Integer
10-19	Screening method used	Binary (0,1)
20	Screening Score	Integer

B. Data Preparation

Data must first be prepared for use as an input by machine learning algorithms by being cleaned, processed, and occasionally modified. There are numerous incorrect or missing records in the collection, and other attributes also need to be modified. Almost all learning algorithms greatly benefit from this preprocessing in terms of performance and ability to predict future events.

C. Data Cleaning

The most crucial aspect of machine learning is data cleaning. It is critical while creating a model. It isn't the most opulent aspect of machine learning, but there are no mysteries or hidden traps. The level of data purification, on the other hand, will influence project success. Professional data scientists normally put a lot of effort into this step since better data "beats fancier algorithms," as they say.

D. Quick Visualization

Large datasets may be quickly and effectively interpreted with the help of visualization. It is essential for providing a deeper understanding of data, making it possible to quantify its impact on a business, and successfully communicating insights to both internal and external stakeholders [5][7]. We used the strength of two well-known Python visualization tools, seaborn and matplotlib, during the data processing process**.** Fig 1 shows Jaundice classification visualization. Fig 2 shows Gender classification visualization

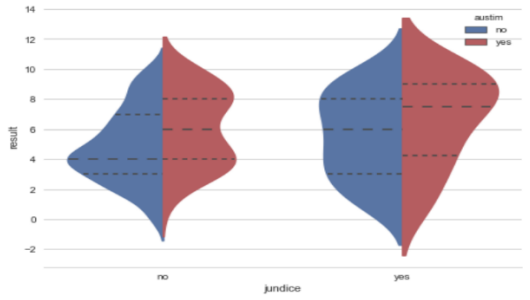

Fig 1: Jaundice classification visualization

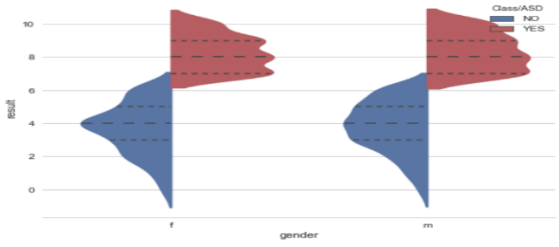

Fig 2: Gender classification visualization

E. One Hot Coding

A method used in machine learning models for expressing categorical variables into number values. The following are some of the advantages of employing one hot encoding:

- It enables the use of categorical variables in models that need numerical input.
- By giving the model more information about the category variable, one can improve model performance.

F. Train and Test Split Data

The entire dataset has been split into two parts, with an 80:20 split between training and testing used for each half. For the purpose of cross-validation, the training data has been split into two halves once more. Eighty percent of the total comes from the training dataset, while twenty percent comes from the validation dataset. Fig 3 shows testing and training split.

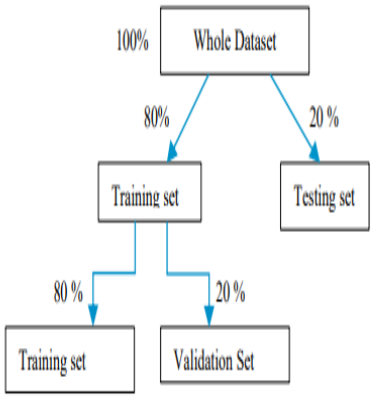

Fig 3: Testing and Training split

G. Model

Because the model is used to identify adults with ASD, we must categorize in line with the dataset in order to forecast the future. There are several machine learning algorithms accessible for classification [5], therefore we choose SVM, LOGISTIC REGRESSION, and NAIVE BAYES. We proceed after consulting with them.

H. Evaluation Of Model Performance

1) Metrics

$$Accuracy = \frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Negative} \text{ ----- [15]}$$

$$Sensitivity\ or\ Recall\ = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \text{ ----------[15]}$$

$$Specificity\ = \frac{True\ Negative}{True\ Negative\ +\ False\ Positive} \quad (1)$$

$$F1\ score = 2\ .\ \frac{Precision*Recall}{Precision\ +\ Recall} \quad (2)$$

Fig 4 shows Area Under Curve (AUC) visualization.

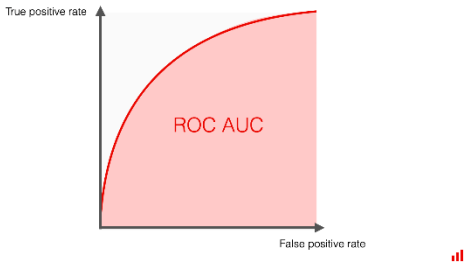

Fig 4. Area Under Curve (AUC) visualization [19]

2) Support Vector Machine

A powerful machine learning method called Support Vector Machines (SVM) can handle both linear and nonlinear problems, including classification, regression, and even outlier identification. SVMs are used in a variety of fields, such as handwriting recognition, spam detection, text classification, and picture classification. The primary intent of the algorithm known as SVM is to find the optimum hyperplane in an N-dimensional space, allowing the data points to be split into different categories in the feature space. The main goal of this hyperplane is to increase the distance between the nearest data points across various classes. The amount of input characteristics is correlated with the hyperplane's dimensionality. For instance, the hyperplane is basically a line when there are just two input characteristics. The hyperplane adopts the appearance of a 2-D plane while dealing with three input characteristics. **Fig 5** shows Example of SVM working.

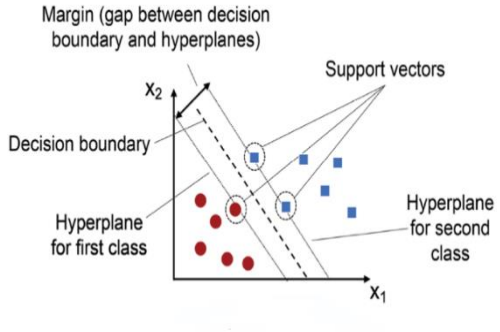

Fig 5: Example of SVM working [21]

3) Naive Bayes

The Naive Bayes Classifier Algorithm, which incorporates the "naive" assumption of conditional independence between any pair of characteristics, belongs to the class of probabilistic algorithms that make use of Bayes' theorem.

The Bayes theorem is used to determine the probability represented as P(c|x), where 'c' stands for the class of likely outcomes and 'x' represents the provided instance that has to be identified based on certain specified attributes. **Fig 6** shows Naive bayes working.

$$P(c|x) = P(x|c) * P(c) / P(x) \quad (3)$$

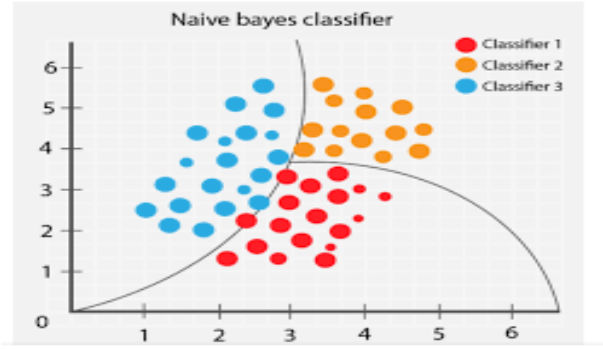

Fig 6: Naive bayes working [17]

4) Logistic Regression

To determine the likelihood that a given instance belongs to a specific class, the logistic regression supervised machine learning conduct is used. In classification problems, this technique, known as logistic regression, is frequently utilized. The reason why it uses a sigmoid function to estimate the likelihood of a certain class based on the results of a linear regression function is why the term "regression" is employed. Utilizing a collection of independent factors, logistic regression is used to predict categorical dependent variables. **Fig 7** shows Logistic Regression working.

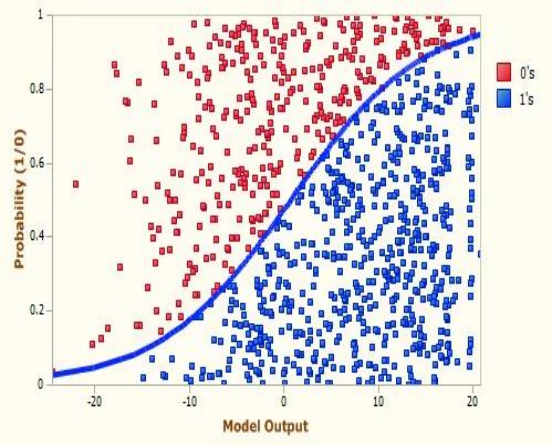

Fig 7: Logistic Regression working [20]

I. Model Tuning

Model tuning is an experimental methodology intended to identify the ideal hyperparameter settings that enhance a model's performance. Hyperparameters are a set of variables that are crucial to the regulation of the training process but whose values cannot be determined from the training data isolated. Model optimization is a different term for model tweaking.

J. Extracting Feature Importance

The use of ensembles of decision tree methods such as gradient boosting has the advantage of automatically providing estimates of feature relevance from a trained predictive model.

The following are the steps to extracting feature importance:

- If the supervised learning model is different from the one that was formerly used, import it.
- Use the whole training set to train the supervised model.
- Using feature importance extraction

Gradient Boosting Classifier or AdaBoost Classifier may be utilized when importing a model of supervised learning with feature importance.

K. Feature Selection

Because it involves reducing the amount of input variables, feature selection is an important step in the creation of predictive models. This accomplishes two goals: first, it minimizes the computing load associated with modelling, and second, in some cases, it may improve the model's overall performance [1].

Statistically based feature selection approaches use statistical methodology to assess the relationship between each input variable and the target variable. The goal is to find and keep input variables that are closely related to the target variable. Remember that the proper statistical measures are determined by the data types of the input and output variables [7]. These solutions, on the other hand, have the added benefit of being both effective and efficient.

L. Feedforward Neural Network

One of two kinds of artificial neural networks that may be distinguished by the direction of information flow between its layers is the feedforward neural network (FNN). The information in the model simply flows forward from the input nodes to the hidden nodes (if any), and then to the output nodes, in contrast to recurrent neural networks that include cycles and loops. Modern feedforward neural networks, also referred to as "vanilla" neural networks, are trained using backpropagation. **Fig 8** shows Feedforward neural network (FNN) working. **Fig 9** shows Steps for ASD classification using machine learning.

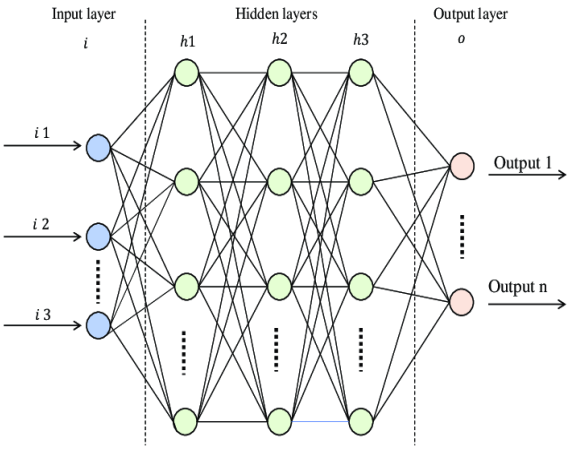

Fig 8: Feedforward neural network (FNN) working [18]

M. Flowchart

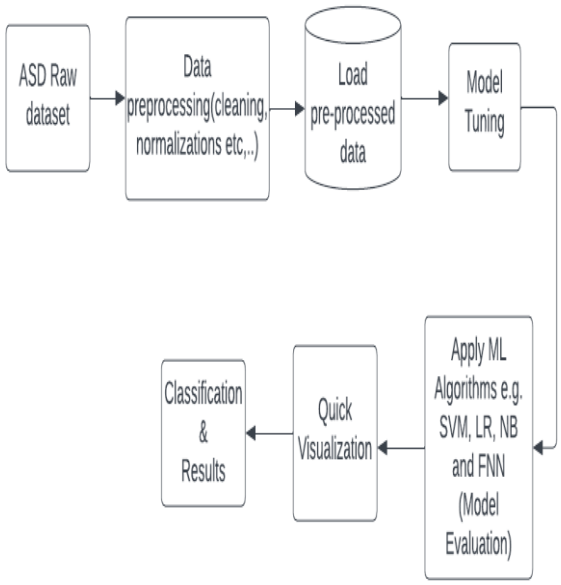

Fig 9: Steps for ASD classification using machine learning

IV. Comparison & Analysis

These are some of the supervised machine learning algorithms that we applied for the dataset to choose the appropriate one [2]. Table 2 represents the performance metrics of the proposed model for ASD classification.

TABLE 2: PERFORMANCE METRICS OF MODEL FOR ASD CLASSIFICATION

Algorithm	F1 Score	Accuracy	AUC	specificity
Decision Tree	0.78	0.87	0.84	0.93
Random Forest	0.85	0.91	0.98	0.96
SVM	0.86	0.92	0.98	0.96
Logistic Regression	0.97	0.98	0.99	0.98
Naïve Bayes	0.85	0.94	0.94	0.91
Proposed Model	0.97	0.98	0.99	0.99

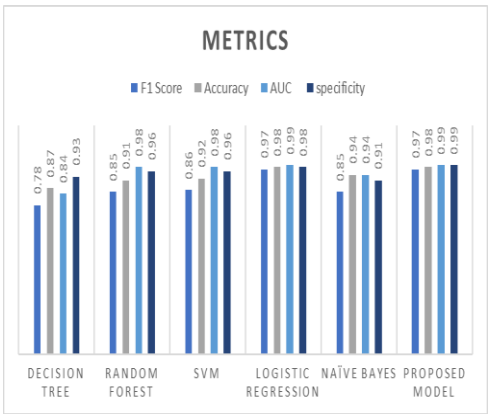

Fig 10: Graphical data for performance metrics

V. CONCLUSION

A progressive learning curve is necessary for someone with autism spectrum disorder. The long-term effects are better the earlier ASD is discovered. Until a child becomes adult, ASD is frequently misdiagnosed. This study has used a range of machine learning approaches to identify Autism Spectrum Disorder (ASD) with supervised machine learning algorithms. Here, the data analysis is performed in-order to find the correlation. This study has examined the performance of several models using a dataset of adult non-clinical data and a range of evaluation characteristics. To improve the accuracy and efficacy of ASD identification, this study has pre-processed the data and employed supervised learning classifiers with an emphasis on feature importance.

REFERENCES

[1] S. M. Mahedy Hasan, M. P. Uddin, M. A. Mamun, M. I. Sharif, A. Ulhaq, and G. Krishnamoorthy, "A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders," IEEE Access, vol. 11, pp. 15038–15057, 2023, doi: 10.1109/access.2022.3232490.

[2] S. J. J. Jui, R. C. Deo, P. D. Barua, A. Devi, J. Soar, and U. R. Acharya, "Application of Entropy for Automated Detection of Neurological Disorders with Electroencephalogram Signals: A Review of the Last Decade (2012–2022)," IEEE Access, vol. 11, pp. 71905–71924, 2023, doi: 10.1109/access.2023.3294473.

[3] N. A. Mashudi, N. Ahmad, and N. M. Noor, "Classification of adult autistic spectrum disorder using machine learning approach," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 10, no. 3, p. 743, Sep. 2021, doi: 10.11591/ijai.v10.i3.pp 743-751.

[4] J. Shin, Md. Maniruzzaman, Y. Uchida, Md. A. M. Hasan, A. Megumi, and A. Yasumura, "Handwriting-Based ADHD Detection for Children Having ASD Using Machine Learning Approaches," IEEE Access, vol. 11, pp. 84974–84984, 2023, doi: 10.1109/access.2023.3302903.

[5] S. Raj and S. Masood, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," Procedia Computer Science, vol. 167, pp. 994–1004, 2020, doi: 10.1016/j.procs.2020.03.399.

[6] C. Küpper et al., "Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning," Scientific Reports, vol. 10, no. 1, Mar. 2020, doi: 10.1038/s41598-020-61607-w.

[7] A. Baranwal and M. Vanitha, "Autistic Spectrum Disorder Screening: Prediction with Machine Learning Models," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Feb. 2020, doi: 10.1109/ic-etite47903.2020.186.

[8] M. Farrell, K. N. Ramadhani, and S. Suyanto, "Combined Firefly Algorithm-Random Forest to Classify Autistic Spectrum Disorders," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Dec. 2020, doi: 10.1109/isriti51436.2020.9315396.

[9] F. Thabtah, "Autism Spectrum Disorder Screening," Proceedings of the 1st International Conference on Medical and Health Informatics 2017, May 2017, doi: 10.1145/3107514.3107515.

[10] Devika Varshini, G., and R. Chinnaiyan. "Optimized machine learning classification approaches for prediction of autism spectrum disorder." Ann Autism Dev Disord. 2020; 1 (1) 1001 (2020).

[11] M. Altınkaynak et al., "Diagnosis of Attention Deficit Hyperactivity Disorder with combined time and frequency features," Biocybernetics and Biomedical Engineering, vol. 40, no. 3, pp. 927–937, Jul. 2020, doi: 10.1016/j.bbe.2020.04.006.

[12] Y. H. Wada et al., "Mental health in Nigeria: A Neglected issue in Public Health," Public Health in Practice, vol. 2, p. 100166, Nov. 2021, doi: 10.1016/j.puhip.2021.100166.

[13] M. H. Aliefa and S. Suyanto, "Variable-Length Chromosome for Optimizing the Structure of Recurrent Neural Network," 2020 International Conference on Data Science and Its Applications (ICoDSA), Aug. 2020, doi: 10.1109/icodsa50139.2020.9213012.

[14] R. Rucco et al., "Brain connectivity study by multichannel system based on superconducting quantum magnetic sensors," Engineering Research Express, vol. 2, no. 1, p. 015038, Feb. 2020, doi: 10.1088/2631-8695/ab7869.

[15] S. Yuvaraj, J. Vijay Franklin, V.S. Prakash, A. Anandaraj, "An Adaptive Deep Belief Feature Learning Model for Cognitive Motion Recognition", IEEE Xplore, , June 2022, DOI: 10.1109/ICACCS54159.2022.9785267.

[16] https://vitalflux.com/wp-content/uploads/2022/08/support-vector-machine-1-640x354.

[17] https://kdagiit.medium.com/naive-bayes-algorithm-4b8b990c7319

[18] https://miro.medium.com/v2/resize:fit:1400/1*ub-ifcgdi9xgryqvo0_GRA.png

[19] https://uploads-ssl.webflow.com/6266b596eef18c1931f938f9/64760408cdd765c3f8e364c3_classification_metrics_002-min.png

[20] https://blog.goodaudience.com/machine-learning-using-logistic-regression-in-python-with-code-ab3c7f5f3bed

[21] https://vitalflux.com/classification-model-svm-classifier-python-example/