HBRP PUBLICATION

Advanced Innovations in Computer Programming Languages
Volume 3 Issue 3

Premature Identification of Autism Spectrum Disorder using Machine Learning Techniques

Suhas GK[1], Naveen N[2], Nagabanu M[3], Mario Edwin R[4], Nithish Kumar R[5*]
[1]*Assistant Professor, Department of CSE, HKBKCE,* Bangalore, India
[2,3,4,5]*Student, Department of CSE, HKBKCE,* Bangalore, India
[*]***Corresponding Author***
E-mail Id:-*1hk17cs107@hkbk.edu.*

ABSTRACT

Autism Spectrum Disorder (ASD) is gaining traction quicker than ever before. Autism features can be detected by screening tests, but they are costly and time consuming. Autism can now be predicted within 12 to 36 months thanks to advances in artificial intelligence and machine learning (ML). The suggested model will be tested using the AQ-10 dataset as well as 1000 real-world data from people with and without autistic symptoms. These datasets contain a collection of questions with only two possible answers: yes or no. Autistic children's parents will have noticed a specific pattern in their behavior, and they will be able to answer the questions based on these observed behavioral patterns. After collecting all the inputs in the form of excel sheet further it will be fed into the proposed system. With the usage of different algorithms such as Random Forest, Support Vector Machine and Ada boost the system will process the data efficiently and give the output. For both types of datasets, the analyzed findings demonstrate that the proposed prediction model will deliver improved outcomes in terms of accuracy, specificity, precision, and false positive rate (FPR).

Keywords:-*ASD, AQ-10 dataset, artificial intelligence, machine learning, false positive rate*

INTRODUCTION

Autism disease is a neurological illness that impairs the capacity of a person to connect, communicate and learn. Autism may be diagnosed at any age, although usually the symptoms begin in the first two years. Autism sufferers have a range of obstacles, such as concentration difficulty, learning impairments, anxiety and depression mental health concerns, mobility difficulties, sensory problems and more.

Autism is currently exploding all over the world, and it is growing at an alarming rate. According to the WHO, ASD affects around one out of every 160 children. Some persons with this illness are able to live independently, while others will need lifelong care and assistance. Autism diagnosis takes a long time and costs a lot of money. Most of the methods detected autism in its later phases, although earlier diagnosis of autism can help a lot by administering right medicine to individuals at an early stage. It has the potential to prevent the patient's health from deteriorating further, as well as reducing the long-term expenses of delayed diagnosis. As a result, a time-efficient, accurate, and simple screening test tool is desperately needed.

Early detection of autism can be quite beneficial in terms of delivering appropriate medication to patients. It has the potential to prevent the patient's health from deteriorating further, as well as reducing the long-term expenses of delayed diagnosis. Various data mining procedures and their applications were researched or investigated. Various medical data sets were linked to the use of machine learning methods. In different

medical data sets, machine learning algorithms have varying degrees of effectiveness. Conventional machine learning algorithms, as previously indicated, produced less precise results.

The aim of this study is to create an autistic predictive model based on machine learning methods to detect autistic symptoms in kids correctly. In other words, this study focuses on developing an intelligence screening model in young children from 12 to 36 months old for predicting ASD characteristics.

Our suggested method relies on unique machine learning processes for the classification and prediction of Autism Spectrum Disorder (ASD) and therefore solves the existing problem. By utilising Random Forest (RF), Support Vector Machine and Ada Booster methods, we may enhance the performance and accuracy of our model.

RELATED WORKS

- This paper [“Hidden Markov Models to Estimate the Probability of Having Autistic Children"] Using Hidden Markov models, a model was created to assess the likelihood of autistic parents producing autistic children, coupled with data on the heritability of autism, thus genetic factors were considered one of the important variables. Autistic parents may produce autistic children with a probability of about 33% for female children, and nearly 80% for male children.
- This paper [“Detecting High-functioning Autism in Adults Using Eye Tracking and Machine Learning"] the authors recorded Adult participants' eye movements with and without autism by providing them with various stimuli and activities. The findings showed that eye tracking data may be utilised to identify highly functioning autism in adults automatically.
- This paper [“A Machine Learning Approach to Predict Autism Spectrum Disorder”] the work focused was On the development of the application for autism screening for autism spectrum disorder prediction among age groups of individuals 4-12 years, 13-18 years and research was undertaken in five phases: Data gathering, synthesization of data, prediction model development, prediction model assessment and mobile application development.
- This paper [“Automatic Detection and Labeling of Self-Stimulatory Behavioral Patterns in Children with Autism Spectrum Disorder”] proposes a system that uses. They're wearable and static, two distinct sensor platforms. The technology automatically identifies self-stimulating behavioural patterns in children with autism and 91.5% of a classification rate for behavioural patterns has been obtained. In the future scope the authors wanted to increase the number of sensors used.

The architecture of the system developed for detection of autism spectrum disorder contains several components as listed below:

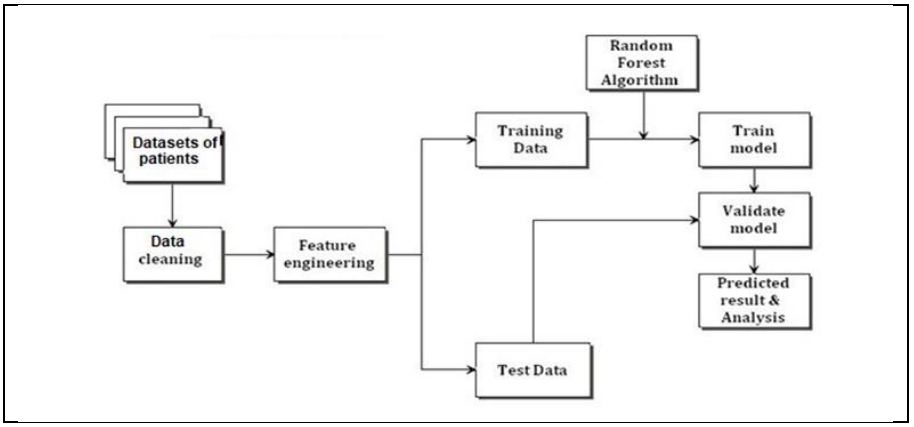
1. **Datasets of patients:** These are the vast amount of data obtained from the user in the form of Questions/Answers for the purpose of detecting autistic traits. This information is present in the form of excel sheet which can be imported into the system for various manipulations and operations to be performed for obtaining the final result. Without proper datasets, it is difficult to initiate the task of performing various operations as it acts as the driving force for feeding algorithms or model to be developed.
2. **Data cleaning:** The information given

by user can be inappropriate, inconsistent or might lack certain trends, features or behaviors which can lead to errors often while performing certain operations. This is where data cleaning performs a very important role in cleaning and obtaining data in a proper and appropriate format. Thus, in order to avoid confusions and save time in cleaning data, the system does not permit user enter the information, what it actually does is, collect from users the information in an efficient manner by giving them pre-defined set of choices to choose from.

3. **Feature engineering:** In the system built, a set of questions are generated to user in order to detect the presence of autism. These set of questions tend to act as features using which various predictions or analysis can be performed. The domain knowledge about medical field plays a vital role in performing feature engineering for a system. So, what we're doing here is that we're converting the high-end language format from the users where information is furnished in the form of Yes/No into a machine understandable format in the form of 0's and 1's.
4. **Training and testing data:** The datasets are segmented into training (80%) and testing (20%) data after the process of feature engineering. Training data helps in training the model using various datasets involving different set of records for each individual. This helps the model in the process of identification of disease in an accurate manner as more and more records are learnt, the better is the knowledge of model. Thus, the trained model can be validated by using few test cases obtained from the test data. In this way, a supervised learning process takes place in developing an efficient model.
5. **Algorithm:** The model can be built using various algorithms. The algorithms chosen for developing the model particularly here are Random Forest algorithm, Support Vector Machine algorithm and Ada-boost algorithm. Most of the traditional systems made use of Random forest and support vector machine algorithms. In the system developed, we're implementing Ada-boost algorithm in order to increase the accuracy of prediction when compared to traditional systems.
6. **Train model:** The model is trained until it achieves a good predictive accuracy. Until then, it is continuously trained in order to learn different types of datasets. In this way, the knowledge extraction of the model becomes more better and better. Ex: There are various patterns present in the questions asked to a user to detect the presence of autism. These different set of patterns can be learned by the model so that whenever it finds an instance similar or same as it, it can classify and produce an accurate decision.
7. **Validate model:** The model is validated to make sure that it produces the required correct decision in order to generate output. A part of datasets can be considered for validation so that we can estimate the accuracy of the model trained. After validation, if the results are not up to the required standards, the model can be trained iteratively again and again, so that after validation it matches the expected or estimated level of accuracy in order to produce the output.
8. **Analysis & Predicted results**: Analysis of data can be done by representing them in the form of various graphs such as Histograms, Joint plot and Heat map. These graphical representations can be used to find out any particular estimates or

trends that could help in making some important decisions. Ultimately, there are two possible predicted results produced at the end. One being if the child is tested positive for autism, they're given a pre-cautionary message and display that they're affected by autism along with safety tips to be taken care of, the other being a congratulatory message displaying the child is safe and is not affected by autism.

SYSTEM ARCHITECTURE



EXPERIMENTAL SETUP

ANACONDA NAVIGATOR: is a desktop tool included with Anaconda distribution that enables users to run applications without the need of command line commands and handle its packages, environments and channels. The navigator may search the Anaconda Cloud or the local Anaconda repository for packages and then install, execute, and update them in a particular environment.
The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- QtConsole
- Spyder
- Glue
- Orange
- RStudio
- Visual Studio Code

JUPYTER NOTEBOOK: It is a web-based interactive computer environment for the creation of Jupyter notebook documents. Based on the circumstances, the word "notebook," along with the Jupyter application form, Jupyter Web Server and Jupyter document format, may refer to a number of things.

RANDOM FOREST: is a learning method for groups. The basic idea behind the approach is that building a small decision tree with few attributes is a computationally cheap task.

If we can develop many small, weak decision trees at the same time, we can combine them into a single, strong learner by averaging or taking the majority vote. Random forests have repeatedly been shown to be the most precise learning algorithms available.

The approach is this: for each tree in the forest we choose a sample of bootstrap from S, where SI indicates the bootstrap. Then we learn a decision-making process utilising a modified decision-taking algorithm.
The algorithm was changed as follows: Instead of evaluating all potential splits at

each tree node, we randomly select a subset of characteristics f as F. The node divides the best feature in f instead of F. F is a fraction of the size of f.

ADA BOOST: This may be used with a number of other learning algorithms to enhance outcomes since certain issue types are better than others and many different parameters and settings must be changed before optimum performance is achieved on a dataset. AdaBoost (which is called the best out-of-the-box classification using decision trees as weak learners).

In conjunction with decision-tab learning, data about the relative 'hardness" of each training sample collected at every step of the AdaBoost technique are included into the tree production process, which makes it harder for subsequent trees to concentrate on samples.

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

```
function RandomForest(S, F)
    H ← ∅
    for i ∈ 1, ..., B do
        S^(i) ← A bootstrap sample from S
        h_i ← RandomizedTreeLearn(S^(i), F)
        H ← H ∪ {h_i}
    end for
    return H
end function
function RandomizedTreeLearn(S, F)
    At each node:
        f ← very small subset of F
        Split on best feature in f
    return The learned tree
end function
```

Given: $(x_1; y_1), \ldots, (x_m; y_m)$, $x_i \in X$, $y_i \in Y = \{-1, 1\}$.
Initialize $D_1(i) = 1/m$.
For $t = 1 \ldots T$:

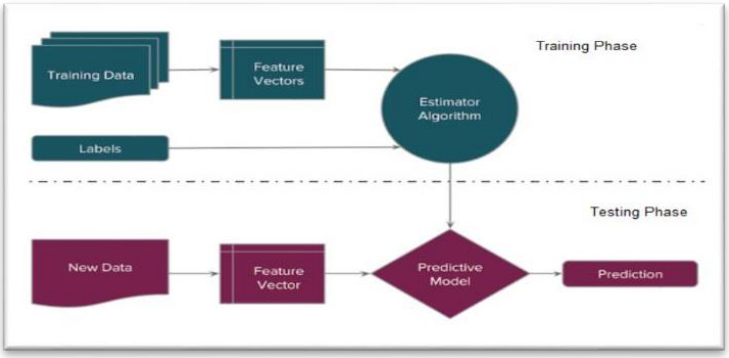
1. Train weak classifier using distribution D_t.
2. Get weak hypothesis $h_t : X \to \{-1, 1\}$ with error $\varepsilon_t = \sum_{i: h_t(x_i) \neq y_i} D_t(x_i)$
3. Choose $\alpha_t = \frac{1}{2} \log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
4. Update:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \cdot = \begin{cases} e^{-\alpha_t} & \text{if instance } i \text{ is correctly classified} \\ e^{\alpha_t} & \text{if instance } i \text{ is not correctly classified} \end{cases}$$

where Z_t is a normalization factor (chosen so that $\sum_{i=1}^{m} D_{t+1} = 1$).

Output the final hypothesis: $H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

IMPLEMENTATION



SYSTEM DESIGN

TRAINING PHASE

By importing, pre-processing and cleaning data, the datasets are being prepared to be fed into various machine learning algorithms.

1. The datasets are split into training and testing data with a ratio of 80:20 percentage.
2. The training data can be fed to any of the algorithms such as random forest, Ada boost or support vector machine algorithms.
3. If we consider random forest algorithm, it builds 'n' number of trees.
4. Where these trees interpret each and every individual record present in the datasets. In this way, the datasets are fit into the model by random forest.
5. Similar method is followed for the remaining algorithms i.e., the Adaboost and support vector machine.
6. Thus, it is a repeat of step 3 and 4.

In this way, model is trained to predict using different algorithms.

TESTING PHASE

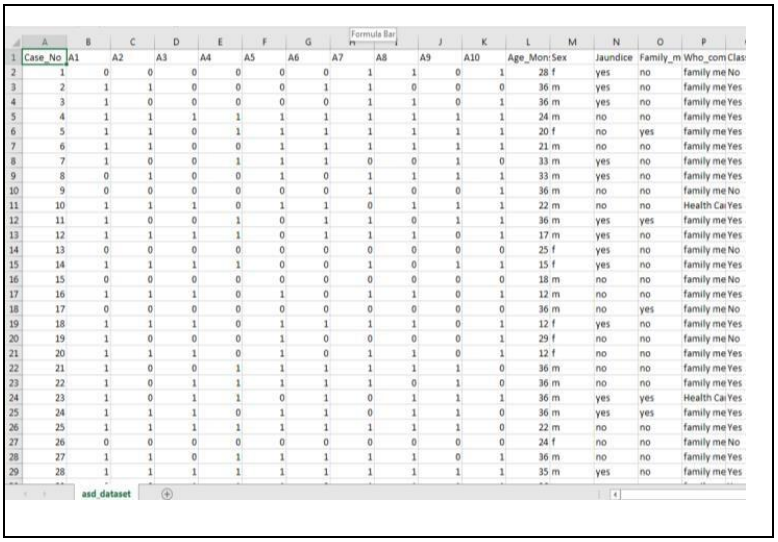
In the testing phase, the system is validated by testing the model which was trained using training data. Each of the algorithms produce their own decisions based on their method of prediction. The overall accuracy of the prediction differs from algorithm to algorithm.

A classification report is generated at the end displaying the precision, recall and f1-scores of the classified data by the algorithms. Also gives the support of each algorithm on the basis of how well it has classified with respect to correct decisions made by it. A confusion matrix is also produced to check the exact number of correct or incorrect decisions made by the model.A1-A10 refers to a set of questions that is related to the child's behaviour and along with that few other parameters such as age in months, sex, whether they have jaundice or not, family men and with whom the test was completed.

RESULTS AND ANALYSIS

The histogram is a very common graphic tool. It is used to summarise data which are either discreet or continuous and measured at an interval. The main characteristics of the data distribution are often shown in a suitable manner.The numpy.histogram() method in NumPy creates a graphical representation of data's frequency distribution. Bins are rectangles of identical horizontal size that correspond to a class interval and have a variable height that corresponds to frequency. Numpy.histogram is a Python package for creating histograms (). The input array and bins are two parameters for the NumPy. Histogram () function. Each bin's border is defined by the subsequent entries in the bin array.

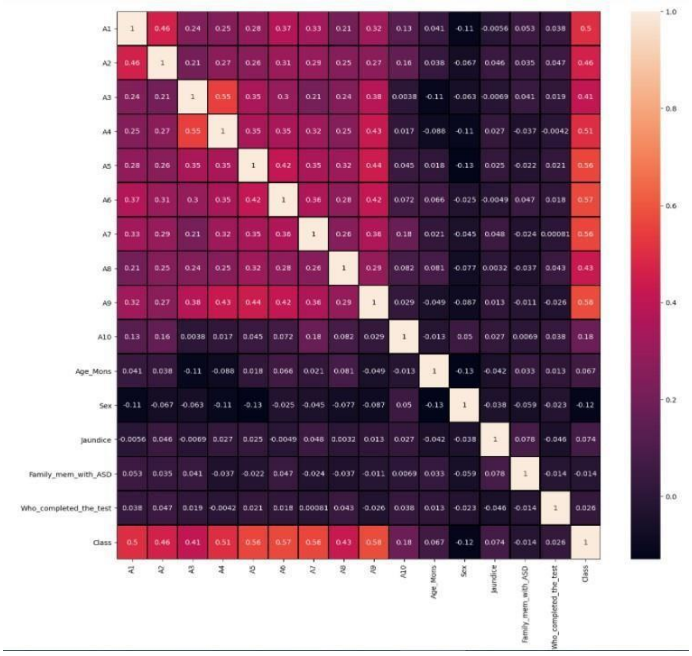
SAMPLE COLLECTED INPUT



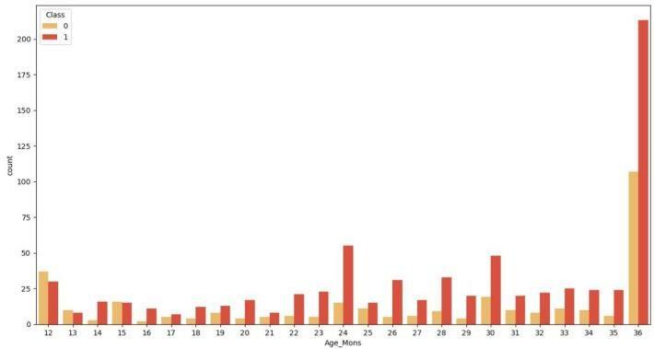
Case_No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Mon	Sex	Jaundice	Family_m	Who_com	Clas
1	1	0	0	0	0	0	1	1	1	0	1	28	f	yes	no	family me	No
2	2	1	1	0	0	0	1	1	0	0	0	36	m	yes	no	family me	Yes
3	3	1	0	0	0	0	1	1	1	0	1	36	m	yes	no	family me	Yes
4	4	1	1	1	1	1	1	1	1	1	1	24	m	no	no	family me	Yes
5	5	1	1	0	1	1	1	1	1	1	1	20	f	no	yes	family me	Yes
6	6	1	1	0	0	1	1	1	1	1	1	21	m	no	no	family me	Yes
7	7	1	0	0	1	1	1	0	0	1	0	33	m	yes	no	family me	Yes
8	8	0	1	0	0	1	0	1	1	1	1	33	m	yes	no	family me	Yes
9	9	0	0	0	0	0	0	1	0	0	1	36	m	no	no	family me	No
10	10	1	1	1	0	1	1	0	1	1	1	22	m	no	no	Health Cai	Yes
11	11	1	0	0	1	0	1	1	0	1	1	36	m	yes	yes	family me	Yes
12	12	1	1	1	1	0	1	1	1	0	1	17	m	yes	no	family me	Yes
13	13	0	0	0	0	0	0	0	0	0	0	25	f	yes	no	family me	No
14	14	1	1	1	1	0	0	1	0	1	1	15	f	yes	no	family me	Yes
15	15	0	0	0	0	0	0	0	0	0	0	18	m	no	no	family me	No
16	16	1	1	1	0	1	0	1	1	0	1	12	m	no	no	family me	Yes
17	17	0	0	0	0	0	0	0	0	0	0	36	m	no	yes	family me	No
18	18	1	1	1	0	1	1	1	1	0	1	12	f	yes	no	family me	Yes
19	19	1	0	0	0	1	0	0	0	0	1	29	f	no	no	family me	No
20	20	1	1	1	0	1	0	1	1	0	1	12	f	no	no	family me	Yes
21	21	1	0	0	1	1	1	1	1	1	0	36	m	no	no	family me	Yes
22	22	1	0	1	1	1	1	0	1	1	0	36	m	no	no	family me	Yes
23	23	1	0	1	1	0	1	0	1	1	1	36	m	yes	yes	Health Cai	Yes
24	24	1	1	1	0	1	1	0	1	1	0	36	m	yes	yes	family me	Yes
25	25	1	1	1	1	1	1	1	1	1	0	22	m	no	no	family me	Yes
26	26	0	0	0	0	0	0	0	0	0	0	24	f	no	no	family me	No
27	27	1	1	0	1	1	1	1	1	0	1	36	m	no	no	family me	Yes
28	28	1	1	1	1	1	1	1	1	1	1	35	m	yes	no	family me	Yes

asd_dataset

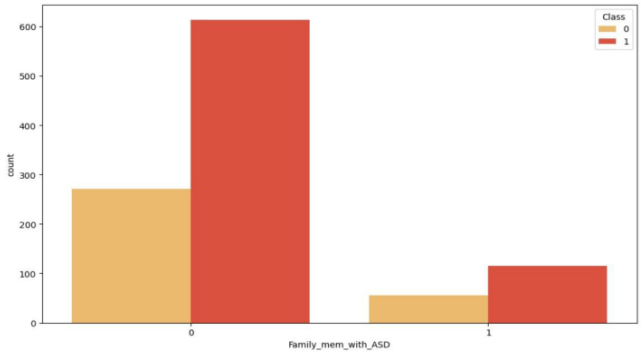
Histogram using sex vs. count variable.



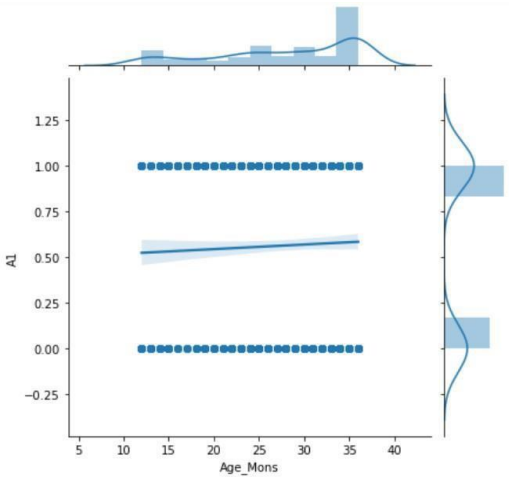
Histogram using Age_in_months vs. count variable



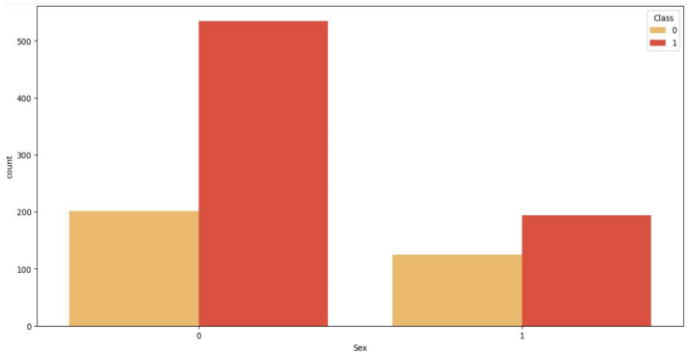
Histogram using Family_members_with_ASD vs. count variable



Jointplot using Age_in_months vs. A1 variable



HEAT MAP



CONCLUSION

We have proposed method to develop a transparent and Universal system that helps in detection of ASD between age groups 12-36 months. And incorporating a large number of datasets to ensure the precision of values yield higher accuracy for enhancing the detection of ASD utilizing Random Forest, Support Vector Machine, Ada Boost algorithms. Along with this, A cost-efficient and a simpler method of working in detection of ASD is our idea as of now for the current model that will be developed. Our future work further relies on increasing the rate of accuracy, precision values for the dataset, and a cost-effective system with an uncomplicated approach for working with respect to detection.

REFERENCES

1. Frith, U., & Happé, F. (2005). Autism spectrum disorder. *Current biology*, *15*(19), R786-R790.
2. WHO, Autism spectrum disorders, 2017 [Accessed August 22, 2018].
3. Amendah, D., Grosse, S. D., Peacock, G., & Mandell, D. S. (2011). The economic costs of autism: A review. *Autism spectrum disorders*, 1347-1360.
4. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*, 117693510600200030.
5. Khan, N. S., Muaz, M. H., Kabir, A., & Islam, M. N. (2017, December). Diabetes predicting mhealth application using machine learning. In *2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)* (pp. 237-240). IEEE.
6. Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). 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. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202-212.
7. Thabtah, F., Kamalov, F., & Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. *International journal of medical informatics*, *117*, 112-124.
8. Wall, D. P., Dally, R., Luyster, R., Jung, J. Y., & DeLuca, T. F. (2012). Use of artificial intelligence to shorten the behavioral diagnosis of autism.
9. Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi- instrument fusion. *Journal of Child Psychology and Psychiatry*, *57*(8), 927-937.
10. 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.

11. 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.
12. Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, *17*, 16-23.
13. Skafidas, E., Testa, R., Zantomio, D., Chana, G., Everall, I. P., & Pantelis, C. (2014). Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Molecular psychiatry*, *19*(4), 504-510.
14. Yaneva, V., Eraslan, S., Yesilada, Y., & Mitkov, R. (2020). Detecting high-functioning autism in adults using eye tracking and machine learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(6), 1254-1261.
15. 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.
16. Zhao, Z., Zhang, X., Li, W., Hu, X., Qu, X., Cao, X., ... & Lu, J. (2019). Applying machine learning to identify autism with restricted kinematic features. *IEEE Access*, *7*, 157614-157622.
17. Shuvo, S. B., Ghosh, J., & Oyshi, A. S. (2019, July). A data mining based approach to predict autism spectrum disorder considering behavioral attributes. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.
18. Obafemi-Ajayi, T., Settles, L., Su, Y., Germeroth, C., Olbricht, G. R., Wunsch, D. C., ... & Miles, J. (2017, November). Genetic variant analysis of boys with Autism: A pilot study on linking facial phenotype to genotype. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 253-257). IEEE.
19. Min, C. H. (2017, July). Automatic detection and labeling of self-stimulatory behavioral patterns in children with Autism Spectrum Disorder. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 279-282). IEEE.
20. Altay, O., & Ulas, M. (2018, March). Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. In *2018 6th International Symposium on Digital Forensic and Security (ISDFS)* (pp. 1-4). IEEE.