IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

Volume 8, Issue 10, October 2020

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 10, October 2020 ||

# Comparative Study of Autism Spectrum Disorder Detection between Machine Learning Algorithms

**Reetodeep Hazra [1], Megha Banerjee [1]**

Student, Department of Electronics and Communication Engineering, Techno International New Town, Kolkata, West Bengal, India[1]

**ABSTRACT**: India, with near about 855,000 positive cases, is the largest contributor of autistic children worldwide. It is said to be influenced by environmental and genetic causes; with no clear answer provided by the researchers till now. Brain imaging of autism affected children revealed that their brain structures are different than other neurotypical children, so it is widely accepted that damaged brain structure caused Autism Spectrum Disorder.

**KEYWORDS**: Autism Spectrum Disorder detection, Autism data visualisations, Autism classifier, K nearest neighbour, Random forest algorithm, Machine learning

## I. INTRODUCTION

Autism spectrum disorder (ASD) refers to a range of conditions characterised by some impaired social behaviour, communication and language, and limited interests in different social and general activities that are both unique to the individual and carried out repetitively. As of 2018, India was the greatest contributor of autistic children (851,000), followed by China (422,000) and Nigeria (207,000) in the 2$^{nd}$ and 3$^{rd}$ place respectively. It is estimated that one in 160 children suffers from ASD. This estimation speaks to a normal figure, and detailed predominance changes significantly across studies. Some well-controlled examinations have, nonetheless, announced figures that are generously higher. The commonness of ASD in some low-and middle pay nations is so far obscure. While a few people with ASD can live freely, others have extreme disabilities and require long lasting consideration and backing.

Cause of ASD includes several aspects like environmental, non-genetic, and genetic factors. But there is no single particular cause for autism. Brain imaging of autism affected children revealed that their brain structures are different than other neurotypical children, so it is widely accepted that damaged brain structure caused ASD.

Paper is organized as follows. Section II describes related works done in this area before. The methodology is given in Section III. Section IV presents experimental results showing results followed by conclusion in the section V.

## II. RELATED WORK

Erkan et al. [1] used k-Nearest Neighbours method (kNN), Support Vector Machine method (SVM) and Random Forest method (RF) for the detection of ASD. The final results they achieved were assessed by the average values. It was shown that SVM and RF are effective methods for ASD classification. They concluded that if the number of data samples is large enough, a high accuracy can be achieved for machine learning-based ASD diagnosis. Among three classification methods, RF achieved the best performance for ASD data classification. Kazi et al. [2] proposed an effective prediction model based on machine learning (ML) technique and to develop a mobile application for predicting ASD for people of any age. As outcomes of the research, they developed an autism prediction model by merging Random Forest-CART (Classification and Regression Trees) and Random Forest-ID3 (Iterative Dichotomiser 3) and also a mobile application was developed based on the proposed prediction model. They concluded that with the help of autism screening application, an individual can be guided at an early stage that will prevent the situation from getting any worse and reduce costs associated with delayed diagnosis. Suman et al. [3] attempted using various ML and deep learning techniques. Various performance evaluation metrics were used to analyse the performance of the models implemented for ASD detection on non-clinical dataset from three sets of age groups viz. Child, Adolescents and the Adult. They concluded that CNN based model was able to achieve highest accuracy result than all the other considered model building techniques, These results strongly suggested that a CNN based model can be implemented for detection of Autism Spectrum Disorder instead of the other conventional machine learning classifier suggested in earlier researches. Mamata et al. [4] revealed the classification between ASD and TC subjects for sMRI and fMRI using the K-NN classifier for different feature sets. The paper used GLCM method for feature extraction and forward selection of features for performance improvement. The K-NN classifier gave the maximum accuracy for sMRI with 5 features and for fMRI with 9 features for K-fold=15. Further accuracy was decreased or increased as features were added. They concluded that to improve the accuracy of classification, feature optimization is required. Fusion of sMRI and fMRI images can give better results as fused image is having the qualities of both the source images. Fatiha et al. [5] compared three different classification methods, Naive Bayes, IBk (k-nearest neighbors), RBFN (radial basis function network), and Random Forest, on UCI 2017 Autistic Spectrum Disorder Screening Data for Children dataset. They concluded that as a result of the experiment, Random Forest method has been shown to be more successful than Naive Bayes, IBk and RBFN methods. Tania et al. [6] implemented classification techniques with the transformed ASD datasets and assessed their performance. They found that SVM showed the best performance for the toddler dataset, while Adaboost gave the best results for the children dataset, Glmboost for the adolescent and Adaboost for the adult datasets. The feature transformations resulting in the best classifications were sine function for toddler and Z-score for children and adolescent datasets. After these analyses, several feature selection techniques were used with these Z-score-transformed datasets to identify the significant ASD risk factors for the toddler, child, adolescent and adult subjects. They concluded that the results of these analytical approaches indicate that, when appropriately optimised, machine learning methods can provide good predictions of ASD status. This suggested that it may possible to apply these models for the detection of ASD in its early stages.

## III. METHODOLOGY

### A) DATA SOURCE

The dataset was collected from the UCI Repository and is named as Autistic Spectrum Disorder Screening Data for Adult [7]. The rapid growth of ASD cases is increasing worldwide but such datasets are not available publicly. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. In this dataset, 20 features are utilised for further analysis especially in determining influential autistic traits and improving the classification of ASD cases. In this dataset, ten behavioural features are also recorded with ten individual's characteristics. The attributes of the dataset are given in the table below.

| Class | Number of samples |
|---|---|
| No. of instances (records in the dataset) | 704 |
| No. of attributes (fields within each record) | 21 |

### B) VISUALIZATION

Several visualizations were performed using the dataset. Around 1% of the world population has ASD, but during exploratory data analysis for this sample we got around 27% for Adults and 69% for Toddlers of the data with positive ASD. It's so because the test parameters feature only qualitative properties of ASDs. The visualization of the missing adult and missing toddler is shown on Fig. 3(a) and Fig. 3(b).

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 10, October 2020 ||

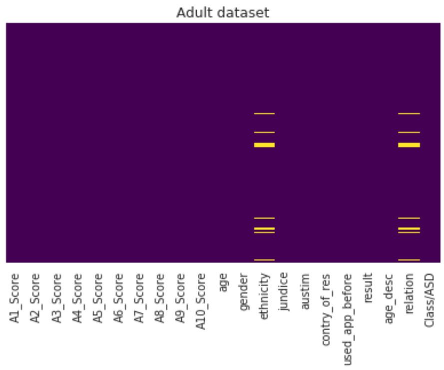


Fig. 3(a)

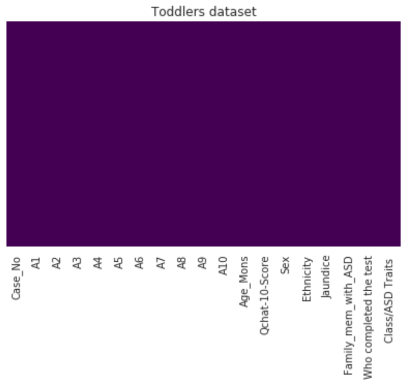


Fig. 3(b)

Fig. 3 (a) Missing Adult (b) Missing Toddler

Next we visualized the ASD positive adults and toddlers born with jaundice with respect to genders. We can see here that almost 6-7 times more in adults and 2-3 times more in toddlers of non-jaundice were born with ASD positive whereas according to reports that is around 10 times. Jaundice born child has a weak link with ASD. Also, according to reports, ASD is more common among boys (around 4-5 times) than among girls. But here in adults we see a lower ratio, whereas in toddlers it's nearly 4 times boys than girls, which is quite close to the actual ratio. Fig. 4(a) shows the ASD positive adults born with jaundice with respect to gender and Fig. 4(b) shows the ASD positive toddlers born with jaundice with respect to gender.

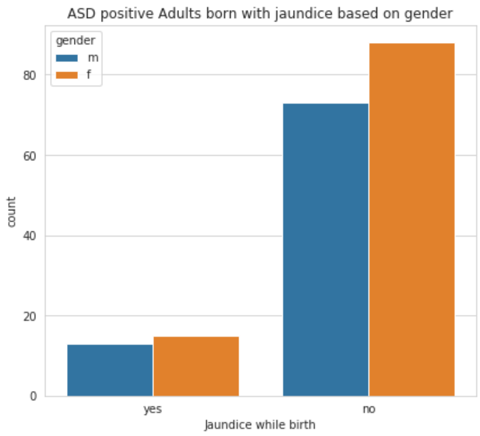


Fig. 4(a)

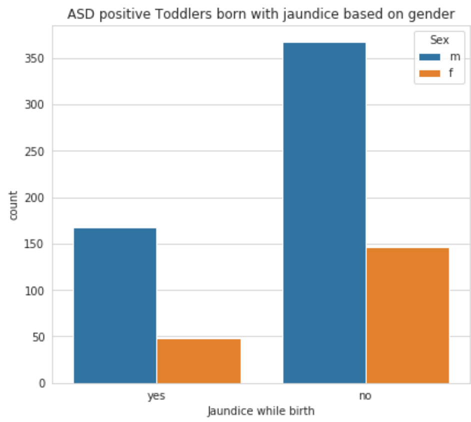


Fig. 4(b)

Fig. 4 (a) ASD positive adults born with jaundice (b) ASD positive toddler born with jaundice

Now we visualized and studied the age distribution of ASD positive in adults and toddlers. For adults most of the ASD positive is seen to be around 20 to 30 years of age, whereas for toddlers most of them are around 36 months. We can see in adults that as the age increases the number decreases, whereas in toddlers as the age increases the number increases which is vice-versa. It goes well with the research. For adults, people with autism develop strategies to help them age better. For toddlers, the significant signs of autism reveals around 3 years of age. Fig. 5(a) shows the age distribution of ASD positive in adults and Fig. 5(b) shows the age distribution of ASD positive in toddlers.

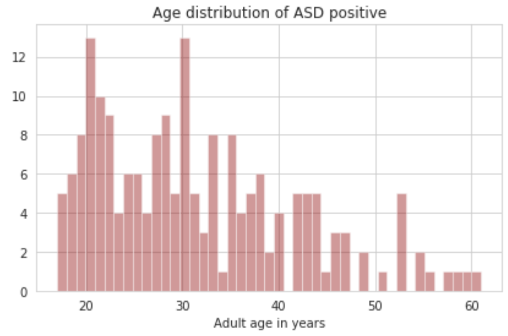


Fig. 5(a)

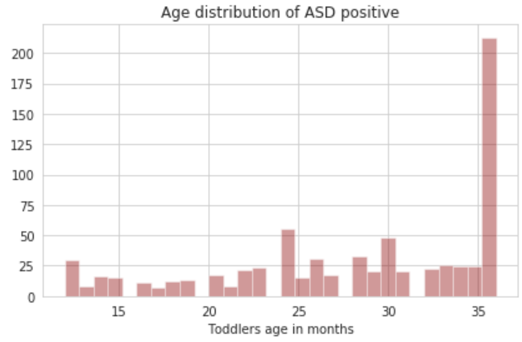


Fig. 5(b)

Fig. 5 (a) Age distribution of ASD positive adults (b) Age distribution of ASD positive toddler

The next visualization was done to study the ASD positive adults based on the top 15 countries. The visualization shown does quite well describing the report. Developed countries like UK, US, Australia and Canada indeed are the most affected ones. But we see the female population is distinguishable compared to males, which is quite contrary. So the white and the European ethnicities overshadow the rest which is quite close to the studies done. Fig. 6 shows the positive ASD of white and European ethnicities. US, UK, Australia, NZ and Canada affirm their positions as top contributors of positive ASD.

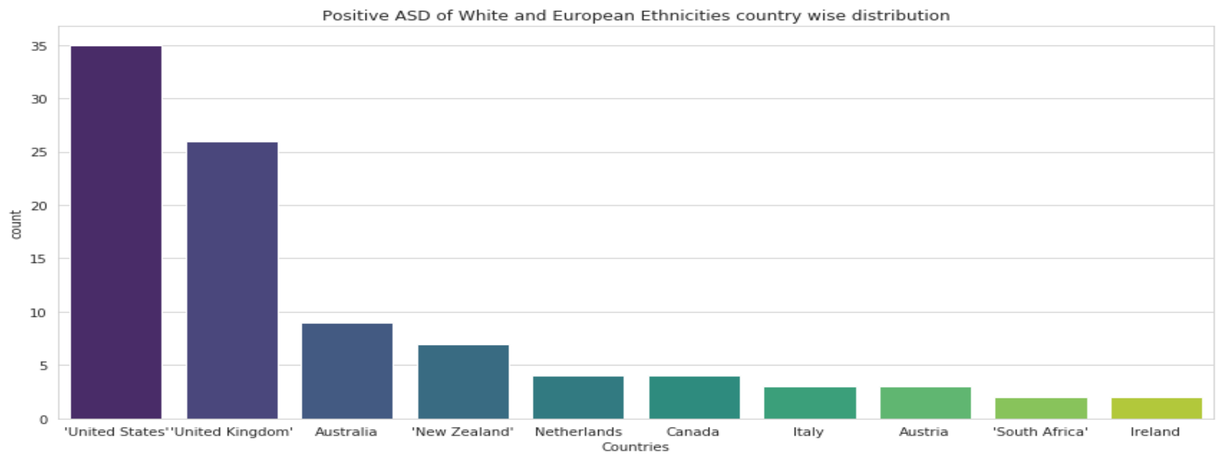


Fig. 6

Fig. 6 Positive ASD of White and European ethnicities country wise distribution

Next we studied and visualized the distribution of autism in family with different ethnicities. We can observe that both in adults and toddlers, white and Europeans ethnicities have a very high chance of being ASD positive if they have it in their genes. Black and Asians follow the next though with smaller ratios. So anything cannot be concluded firmly, but through the visualization it can be concluded that there is a genetic link for ASD positive as backed by studies. Fig. 7(a) shows the positive ASD adult relatives for different ethnicities and Fig. 7(b) shows the positive ASD toddler relatives for different ethnicities.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 10, October 2020 ||

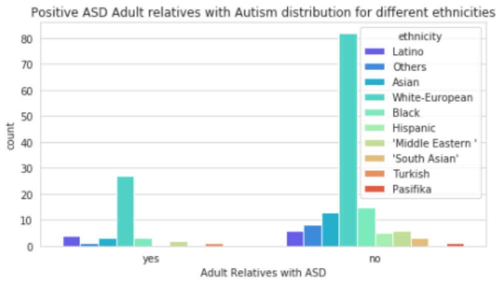


Fig. 7(a)

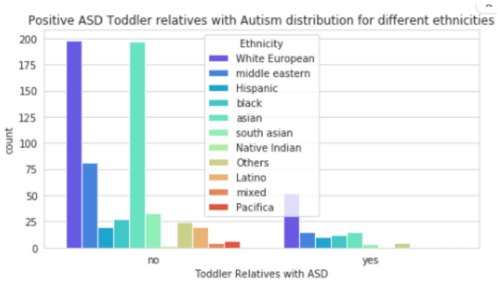


Fig. 7(b)

Fig. 7 (a) Positive adult relatives with autism (b) Positive toddler relatives with autism

## C) APPLICATION OF MACHINE LEARNING ALGORITHMS

We applied 2 machine learning algorithms for identifying whether a toddler has ASD or not. We used Random Forest classifier and K-Nearest Neighbor (KNN) classifier here.

**i)** **Random Forest -** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction**.** One big advantage of random forest is that it can be used for both classification and regression problems. Here we used RF for studying the autism traits and the approach achieved an overall accuracy of 70%.

i) **K-Nearest Neighbor-** K-Nearest Neighbors is one of the most basic yet essential classification algorithms in ML. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors. Here we used elbow methods to choose the value of k. The value of k was chosen to be 13 and the accuracy achieved using the KNN method is 71%. The graph of error rate vs. K-value is shown in Fig. 7.

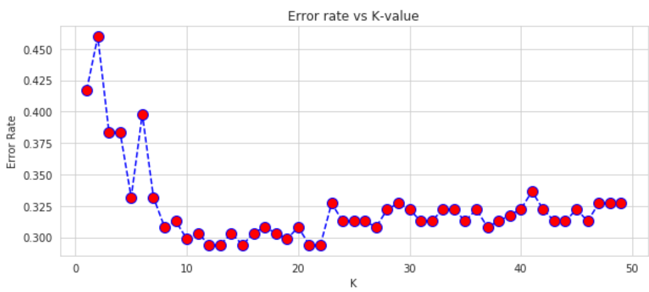


Fig. 7

Fig. 7 Error Rate vs. K-value of KNN classifier

## IV. EXPERIMENTAL RESULTS

### A) METRICES

Confusion matrix is performance estimation for machine learning statistical classification problem where output can be at least two classes. It is a table with 4 distinct blends of predicted and actual classes. The confusion matrix of the proposed model using Random Forest classifier and KNN classifier is shown below. Table 1. shows the confusion matrix of Random Forest classifier and Table 2. shows the confusion matrix of KNN classifier.

| | | |
|---|---|---|
| **Autism** | 24 | 54 |
| **Normal** | 9 | 124 |
| | **Autism** | **Normal** |

Table 1.

| | | |
|---|---|---|
| **Autism** | 27 | 51 |
| **Normal** | 11 | 122 |
| | **Autism** | **Normal** |

Table 2.

Table (1) Confusion matrix of Random Forest classifier; Table (2) Confusion matrix of KNN

### B) CLASSIFICATION RESULTS

The classification results comprising of Precision, Recall and F1 scores of Random Forest classifier and KNN classifier is shown in the tables below. Table 3 shows the classification results of Random Forest classifier and Table 4 shows the classification results of KNN classifier.

| Category | Precision | Recall | F1 Score |
|---|---|---|---|
| **Normal** | 0.73 | 0.31 | 0.43 |
| **Autism** | 0.70 | 0.93 | 0.80 |

Table 3.

| Category | Precision | Recall | F1 Score |
|---|---|---|---|
| **Normal** | 0.71 | 0.35 | 0.47 |
| **Autism** | 0.71 | 0.92 | 0.80 |

Table 4.

Table (3) Classification result of Random Forest classifier; Table (4) Classification result of KNN.

## V. CONCLUSION

In the article, we studied different method to detect ASD using the machine learning methods viz KNN and RF. The experiment was done with a dataset collected from the UCI Repository and is named as Autistic Spectrum Disorder Screening Data for Adult. The acquired results indicate that KNN method produced high classification scores, indicated by accuracy, precision, recall and F1 score. In our experiment, we found that the KNN method is more effective than the RF for ASD data classification. We also observed that the early diagnosis of ASD is absolutely possible. If the number of data samples is large enough, the accuracy of the diagnosis of ASD by machine learning methods will be higher. The accuracy also depends on the completion level of the collected data. If the data of the ASD dataset is complete, the accuracy of the early diagnosis of ASD will be high.

Finally, it can be confirmed that with the KNN method, we can detect ASD easily, fast and accurately. Therefore, ASD can subsequently be treated effectively which in turn can improve the quality of life for patients with ASD and their families.

## REFERENCES

[1] Erkan, Uğur & Thanh, Dang, "Autism Spectrum Disorder Detection with Machine Learning Methods," Current Psychiatry Reviews, vol. 15, pp. 297 – 308, 2011.

[2] Islam, Muhammad Nazrul & Omar, Kazi & Mondal, Prodipta & Khan, Nabila & Rizvi, Md., "A Machine Learning Approach to Predict Autism Spectrum Disorder," ResearchGate, 2019.

[3] Suman Raj, Sarfaraz Masood, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," International Conference on Computational Intelligence and Data Science (ICCIDS 2019), pp. 996-1004, 2019.

[4] Mamata V. Lohar, Suvarna S. Chorage, "Detection of Autism Spectrum Disorder (ASD) using Machine Learning Techniques: A Review," International Journal of Future Generation Communication and Networking, vol. 13, no. 1, pp. 426-438, 2020.

[5] F. N. Büyükoflaz and A. Öztürk, "Early autism diagnosis of children with machine learning algorithms," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, pp. 1-4, 2018.

[6] T. Akter et al., "Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorders," in IEEE Access, vol. 7, pp. 166509-166527, 2019.

[7] Fadi Fayez Thabtah, "Autism Spectrum Disorder Screening Data for Adult," 2017, Available at: https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult


INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.488

INTERNATIONAL STANDARD SERIAL NUMBER INDIA


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING