Enhancing Autism Detection through Machine Learning Models Focusing on Behavioral Analysis

***Abstract*—Autism Spectrum Disorder (ASD) is a complex and enduring condition characterized by challenges related to communication and behavior. The paper suggests a deep learning-based method that makes use of behavior to identify autism in both adults and children by analyzing their behavioral characteristics through machine learning approaches and determining a process that makes autism detection easier and cost-effective. For behavior analysis, we implemented specific models like KNN, Random Forest, CatBoost, SVM, GradientBoost, and Logistic Regression and also ensembled models by incorporating a few of our pre-trained models together to give better accuracy rates. We have acquired behavioral datasets from publicly available platforms called UC Irvine Machine Learning Repository and Kaggle.This paper aims to facilitate model comparisons and streamline the autism detection process using advanced machine learning techniques available today.**

***Index Terms*—Autism detection, Machine learning, KNN, Logistic Regression, Random Forest, CatBoost, GradientBoost, SVM, behavior analysis.**

I. Introduction

Autism spectrum disorder (ASD) is a complicated developmental disease involving extensive difficulties with social communication, narrow interests, and repetitive conduct.Autistic individuals typically exhibit symptoms that differ from those seen in neurotypical individuals. The symptoms of autism can include lack of eye contact, a limited range of interests or intense focus on specific topics, repetitive behaviors, heightened sensitivity to sensory stimuli, difficulty in social engagement, aversion to physical touch, and challenges in adapting to changes in routines. Occasionally, individuals with autism can become overwhelmed in certain situations, leading to what is known as a meltdown, they may express their distress by crying, screaming, or engaging in physical behaviors or they might completely withdraw and become unresponsive. A considerable amount of research has been dedicated to enhancing the accuracy of autism detection through the utilization of various algorithms. .In our research paper, we aim to underscore this comprehensive approach by focusing on behavior analysis of adults and children. We collected behavioral data from both autistic patients and neurotypical individuals from the available online source called UC Irvine Machine Learning Repository and Kaggle, then we conducted preprocessing and testing of the data and then applied machine learning algorithms like Logistics Regression, KNN, GradientBoost, SVM, CatBoost, and RandomForest to analyze and extract valuable insights from the dataset. We will obtain our predictions by training the models, which will enable us to identify and ascertain symptoms of autism based on the analysis of behavioral feature data. Here we have employed different algorithms to train datasets, facilitating the extraction of components associated with human behaviour and expressions. The main goal of this research is to detect autism with the help of behavioral data to alert the patient and their families to early treatment and make sure that they get special care from society. They should get all the opportunities to learn and grow like a normal child and as a person.

II. Literature Review

A. Machine learning framework for early-stage detection of autism spectrum disorders:

In this paper [1], the researcher worked on behavioral datasets of toddlers, children, ado- lescents, and adults to detect ASD. They used different types of feature scaling techniques like QT, and normalizer, for their datasets. They used 8 different ma- chine learning algorithms, such as Ada Boost (AD), Linear Discriminant Analysis (LDA), Random Forest (RF), and K-Nearest Neighbors (KNN), to all 4 features-scaled datasets. For the toddler and children datasets, AD had the highest accuracy among all the classifiers with a mean of 98.6 percent. For the adolescent and adult datasets, LDA had the highest accuracy with a mean of 98.075 percent. They tried to emphasize the claim that ASD detection is easiest during the early stages. They claimed that the limitation of their work was that they could not gather more data to build a generalized model that could be used by people of all age-groups. Lack of enough data can cause issues like overfitting, bias, etc.

B. Early detection of autism spectrum disorder (asd) using traditional machine learning models:

Looking into another paper [2], the researchers worked with a dataset that con- sisted of dialogs from actual parents of autistic children who had been undergoing communication, behavior, and speech therapy. They analyzed each sentence of the dialogs to identify potential ASD symptoms, extracted the relevant features, and then proceeded to train machine learning models to diagnose ASD with new unseen data. They used models like KNN, SVM, LR, and RF. The highest accuracy was found with SVM and LR (71%). KNN had an accuracy of 62% and RF had that of 69%. The data was collected through an online survey on social media, which might have contributed to inaccuracies in data and introduced bias. This likely resulted in the relatively low accuracy of the models

C. Autistic spectrum traits detection and early screening: A machine learning based eye movement study:

Another research [3] worked with eye movement data and tried to implement ma- chine learning algorithms to predict

ASD. They designed a VR scene that included social stimuli like real-life objects, figures, landscapes, and recorded the eye move- ment to enrich their dataset. In their preliminary experiment, they recruited 107 adults, recorded their eye movements while watching the VR scene, and then had them complete a questionnaire that recorded their statements about imagination, social skills, behavior, and other important characteristics. All of these were con- verted into features, and then models like SVM, DT, RF, KNN, Naive Bayes, and Ensemble were tested on the gathered data for ASD detection. Ensemble had the highest accuracy at 73% whereas KNN had the lowest at 55%

D. Autism spectrum disorder detection in toddlers for early diagnosis using machine learning:

Another research [4] also worked with behavior data for children. They conducted a survey that collected answers to critical questions about behavior of children. Their research targeted early detection of ASD, which is why they chose a single age group for their dataset. They chose to run KNN, Random Forest, SVM and Na¨ıve Bayes algorithms on their data to accurately detect ASD in children. The highest accuracy was found for KNN (98%) and the lowest was found for SVM (83RF and Na¨ıve Bayes, their accuracies were 93% and 89%, respectively. They also mentioned the lack of large data which resulted in two of the algorithms they had tested overfitting. Consequently, the model might perform poorly on new, unseen data. Additionally, relying on subjective survey data might introduce biases which can affect the model's performance.

E. Investigation of eye-tracking scan path as a biomarker for autism screening using machine learning algo- rithms:

Another research [5] suggests a deep learning model for early ASD identification using eye-tracking scan path data. The model was trained and assessed based on eye- tracking data from kids with and without ASD. However, the authors acknowledged that more study is required to verify the model on a bigger, more varied dataset since they had a relatively small sample size. Since this is such an important tool for early ASD diagnosis, we can substantially improve the detection accuracy through longitudinal analysis, using more advanced machine learning models and a more diverse and larger dataset.

F. Analysis and detection of autism spectrum disor- der using machine learning techniques:

Authors in [6] applied Naive Bayes, Logistic Regression, K Nearest Neighbor, Sup- port Vector Machine, Artificial Neural Network and Convolutional Neural Network for autism detection of children, adolescents and adults. They got the datasets from the UCI ML repository which had 21 features except for toddlers and all the datasets had 1100 instances. For the adult dataset, the performance of the Convolutional Neural Network was better than that of other models because it had an accuracy rate of 99 percent. In the adolescent dataset, again Convolutional Neural Network performed better since the accuracy rate of the model was 95 percent but the rate of specificity was a bit low.

III. METHODOLOGY

A. Data Collection

The dataset for autism screening in Adults has been collected from Kaggle.In the Kaggle dataset, there are also 704 instances and the attribute type is categorical, continuous, and binary. There are two datasets used for autism screening in toddlers. The one Autism Screening of Toddlers dataset has 1054 cases and 18 attributes—Q-Chat-10 items (A1-A10), age, gender, ethnicity, jaundice, family history of ASD, country of residence, previously used app, and daily screen usage. The Q-Chat-10 score identifies ASD traits.Another dataset is collected from AUTISM RESEARCH: UNIVERSITY OF ARKANSAS Computer Science. In this dataset, there are 1985 instances and 28 attributes. The features of autism spectrum disorder include the Autism Spectrum Quotient, Social Responsiveness Scale, Age, Q-CHAT-10 Score, Genetic Disorders, Depressive Symptoms, Learning Disorder, Speech Delay/Language Disorder, Global Growth Delay/Intellectual Disabilities, Social/Behavioral Issues, Childhood Autism Rating Scale, Anxiety Syndrome, Sex, Ethnicity, History of Jaundice, and Family Members with ASD.

B. Data Preprocessing

We have used three datasets for our research paper. We know the raw data set has missing values, inconsistent data, and errors. So, we must preprocess the raw data to clean it and standardize it to maintain reliability and consistency. The step-by-step preprocessing is as follows:

Identify Common Columns and Merge them

Our three datasets contain different rows and columns. So, Our first task was to identify common columns across the datasets. We have found 17 columns that are similar in each dataset. We renamed all the common columns using the same name. Next, We Combine all the datasets into a single dataset. Some features engineering was done such as deriving age in years from months to maintain consistency. After combining all the data we got 3743 instances and 17 attributes.

Standardize the Categorical Value

	Objects	Unique values	number of unique values
0	Sex	[f, m, F, M]	4
1	Ethnicity	[White-European, Latino, ?, Others, Black, Asi...	23
2	Jaundice	[no, yes, Yes, No]	4
3	Family_mem_with_ASD	[no, yes, No, Yes]	4
4	Who_completed_the_test	[Self, Parent, ?, Health care professional, Re...	11
5	ASD_traits	[NO, YES, No, Yes]	4

Fig. 1: Values Variations

For standardizing the categorical value we first Identify the unique values present in each column to understand the inconsistency and variations. We detected variances in data entry procedures by printing unique values. For example, We standardize "Yes" and "No" in place of "Yes," "yes,"

"YES," "No," and "no."Many more columns are required to standardize it as well. This method ensures consistency, improves data quality, and reduces errors.

Handling Missing Values

Our dataset contains missing values which can negatively affect our result analysis part. That's why we have determined missing values in our column and impute them. we denoted missing values in our dataset using "NaN" which is the standard way. Then we impute missing values using the most frequent strategy. This process helps to retain data integrity, maintain dataset size, and prevent data loss.

	Missing Values
A1	0
A2	0
A3	0
A4	0
A5	0
A6	0
A7	0
A8	0
A9	0
A10_Autism_Spectrum_Quotient	0
Age_Years	2
Sex	0
Ethnicity	95
Jaundice	0
Family_mem_with_ASD	0
Who_completed_the_test	95
ASD_traits	0

Fig. 2: Missing Values

	Missing Values
A1	0
A2	0
A3	0
A4	0
A5	0
A6	0
A7	0
A8	0
A9	0
A10_Autism_Spectrum_Quotient	0
Age_Years	0
Sex	0
Ethnicity	0
Jaundice	0
Family_mem_with_ASD	0
Who_completed_the_test	0
ASD_traits	0

Fig. 3: Impute Values

Convert Categorical to Numeric

Our dataset has some columns with categorical values. However, many ML algorithms need to convert categorical values to numerical values as they only understand numerical formats. We have used the replace method to convert binary categorical values to numerical format. For instance, we converted 'No' to '0', 'Yes' to 1, M' to 1, and 'F' to 0. We have used one hot-coded encoding for multiclass categorical values. Binary columns are created for each column present in a specified column such as the 'Ethnicity' column has categories like 'Asian', 'Black', 'White', etc., One-hot encoding will produce new columns called "Ethnicity-Asian," "Ethnicity-Black," and so on, with binary values indicating the presence or absence of each category. After, one-hot encoding we have a total of 34 columns and 3743 instances. It enhances model accuracy and is compatible with machine learning algorithms.

Correlation Matrix and Feature Importance

The correlation matrix is a strong tool for understanding the relationship between different features present in our dataset. It used Pearson correlation by default which measures linear relationship between all the features. The table displays the correlation between two variables in each cell. The value ranges from -1 to 1, where 1 (red color) indicates that the two variables have a perfect, positive link and -1(blue color) denotes a perfect negative correlation between the two variables. A value of 0 (lighter color) indicates that the two variables have no relationship. From the correlation matrix we can see

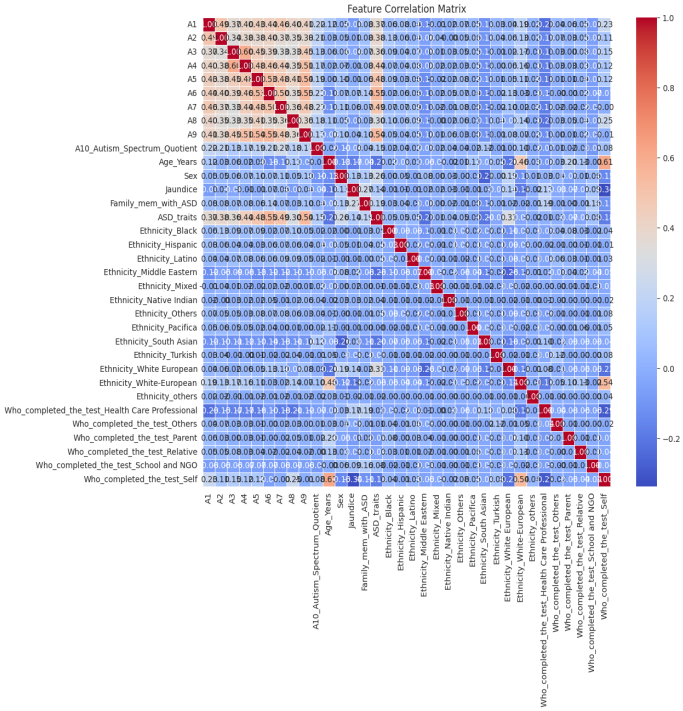


Fig. 4: Correlation Matrix

A1 to A9 have higher inter-correlation with each other and Autism Spectrum Quotient (A10) and Family-mem-with-ASD have moderate correlations with A1 to A9. So, these features are connected to the Autism spectrum quotient. However, other columns show a low correlation with each other and are not strongly related to other features. Ethnicity and sex show a low correlation with autism-related features. Then we derive the feature's importance from our model to highlight the contribution of each feature to predict the model performance. From the graph, We can see the top features of our model which shows the highest scores are A9, age-years, A6, A7, and A5., on the other hand, Ethnicity-Turkish, Ethnicity-Native Indian, and Ethnicity-others show little significance, indicating that they have little bearing on prediction. The correlation matrix and feature importance give a thorough understanding of the relationships and contributions of the variables in our

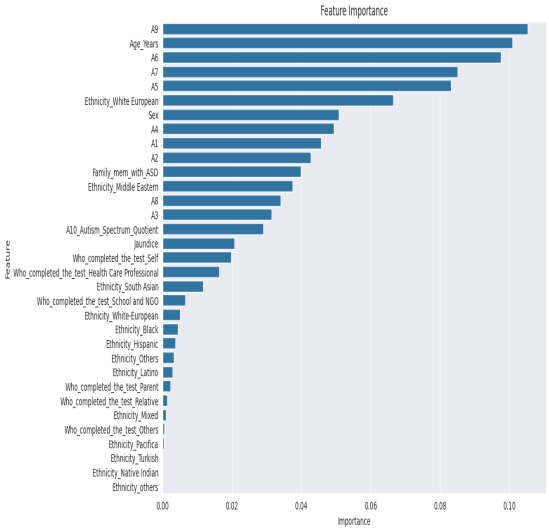


Fig. 5: Feature Importance

study and the underlying data structure and model dynamics. Then we dropped some columns that have less contribution to our dataset.

Principal component Analysis

Principal Component Analysis is a machine-learning technique that uses the dimensionality reduction method to reduce the size of a big data set to a small one while preserving the majority of the data's variance. It makes the process of processing data points for machine learning algorithms much quicker and easier. For applying PCA, we first standardized the data by arranging the data such that its mean is 0 and its standard deviation is 1 for every feature. After that, we use Principal Component Analysis (PCA), which keeps enough components to account for 95 percent of the variation. Our graph shows the explained variance ratio by principal components.

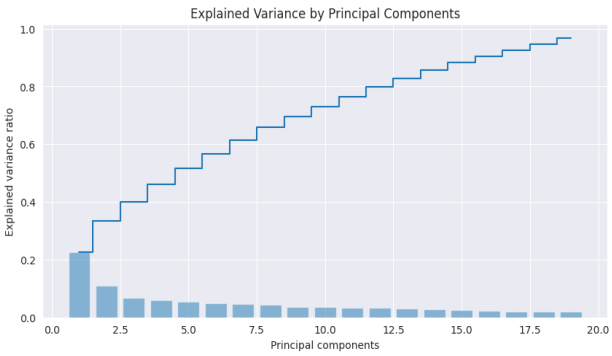


Fig. 6: PC Analysis

The X-axis shows the principal components arranged from the first principal component (PC1) to the twentieth principal component (PC20). The Y-axis represents the ratio of the overall variance explained by every principal component. Each component's explained variance ratio denoted how much of the variance in the entire dataset it accounts for. The line plot displayed the ratio of cumulative explained variance. It is calculated by adding the principal component explained variance ratios consecutively. From the graph, we can see that the first component captures the most variance which is around 20 percent. The bar height is dropped moving from the first to the latter primary components. This indicates that the variance captured by each following principle component is smaller than that of the previous one. The cumulative explained variance line starts at the first principal component's variance and increases as more components are added. After 10 components it might capture around 75 percent variance. As the right-sided components capture less variance, the line levels decrease gradually.

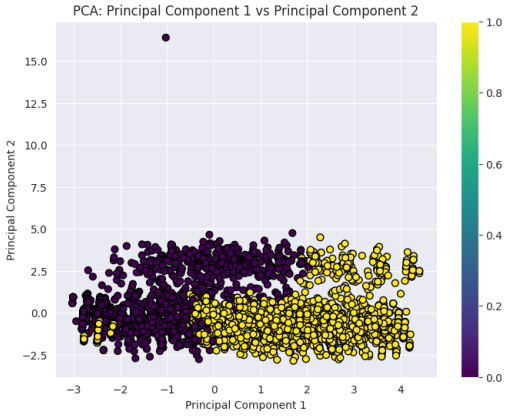


Fig. 7: Relationship Between Two Components

The scatter plot represented the relation between the two components (PC1 and PC2). The X-axis represents the values of the first principal component and Y axis represents the values of the second principal component. PC1 separates the two groups (yellow and purple) with a spread from -3 to +4, suggesting that the data differs significantly in this direction. PC2's spread is less for the bulk of points showing that while it is less significant than PC1 for most points.it does capture variance in a way that makes certain data points stand out. The sample regarding the second principal component that stands out considered unique and distinct from the others is the one with a high PC2 value.

Data Splitting

We have a total of 3743 instances in our dataset. Our dataset is divided into testing and training sets. We have set 80 percent data for the training set and 20 percent data for the testing set.

IV. PROTOTYPE IMPLEMENTATION

A. Logistic regression

The logistic regression model is a type of supervised machine learning algorithm that predicts the probability of an event or outcome by performing binary classification. Only

two possible outcomes can be generated in the logistic regression model. The outcome can either be binary or dichotomous. Relation between one or more inde- pendent variables can be examined in a logistic regression model and classification can take place. It is simply used for prediction purposes.In our processed dataset, we used the logistic regression model on the test dataset and it gave us an accuracy of 87.05binary classification task, this model actually gave us a decent accuracy. Out of 749 test data, the model predicted true positive 338 instances and true negative 314 instances. False positive and false negative instances were 48 and 49 respectively. For class 0 (Non-Autistic), among 362 instances, precision, recall and F1 scores are 87 percent each. For class 1 (Autistic), among 387 instances, precision, recall, F1 scores are 88 percent, 87 percent, 87 p respectiercentvely. The macro average and weighted average for precision, recall, and F1 score are all 0.87, indicating balanced performance across both classes.

B. KNN

KNN is a prevalent machine learning algorithm that operates by assessing proxim- ity to make predictions for various points within datasets. This algorithm majorly depends on the labeled training datasets to make predictions and is widely known for its service as a non-parametric supervised learning classifier as it does not focus on the underlying datasets. K-Nearest Neighbors (KNN) is the next model we used on the test dataset . We achieved an accuracy of 93.59 percent. For a binary classification task this model provided a high level of accuracy and proved to be good. Among 749 test data instances, the model successfully predicted 336 true positives and 365 true negatives. The number of false positives and false negatives were 26 and 22 respectively. For class 0 (Non- Autistic), among 362 instances, the precision, recall, and F1 scores were 94 percent, 93 percent, and 93percent respectively. For class 1 (Autistic), among 387 instances, the precision, recall and F1 scores were 93 percent, 94 percent, and 94 percent respectively. The macro average and weighted average for precision, recall, and F1 score were all 94 percent, indicating a well- balanced performance across both classes.

C. support Vector Machine

SVM is a supervised learning model that can be used for solving both classification and regression problems. It is mainly used for solving tasks of binary classification. For resolving the classification task, it needs to sort out the elements to classify into two groups of a dataset. Data points are called support vector which are crucial for finding out the decision boundary. Also, it helps in finding out the position and orientation of a hyperplane.In our paper, we utilized the SVM model with a linear kernel to categorize autism traits, evaluating its effectiveness on a test dataset. The SVM model showcased an accuracy of 0.86, signifying that 86 percent of its predictions accurately distinguished individuals with and without ASD traits based on the provided features.

D. Gradient Boosting

Gradient Boosting is an effective machine learning technique. It is mainly used for regression and classification tasks. It forms a powerful ensemble model by iteratively combining the strengths of several weak learners, usually decision trees.An overall accuracy of 96.93% using Gradient Boosting was achieved, which indicates the model's strong performance. From the confusion matrix, we found out that the model correctly identified 348 out of 362 instances of the Non-autistic class and 378 out of 387 instances of the Autistic class. Both classes having low false positive rates and high true positive rates indicate a balanced performance across classes.

E. Random Forest

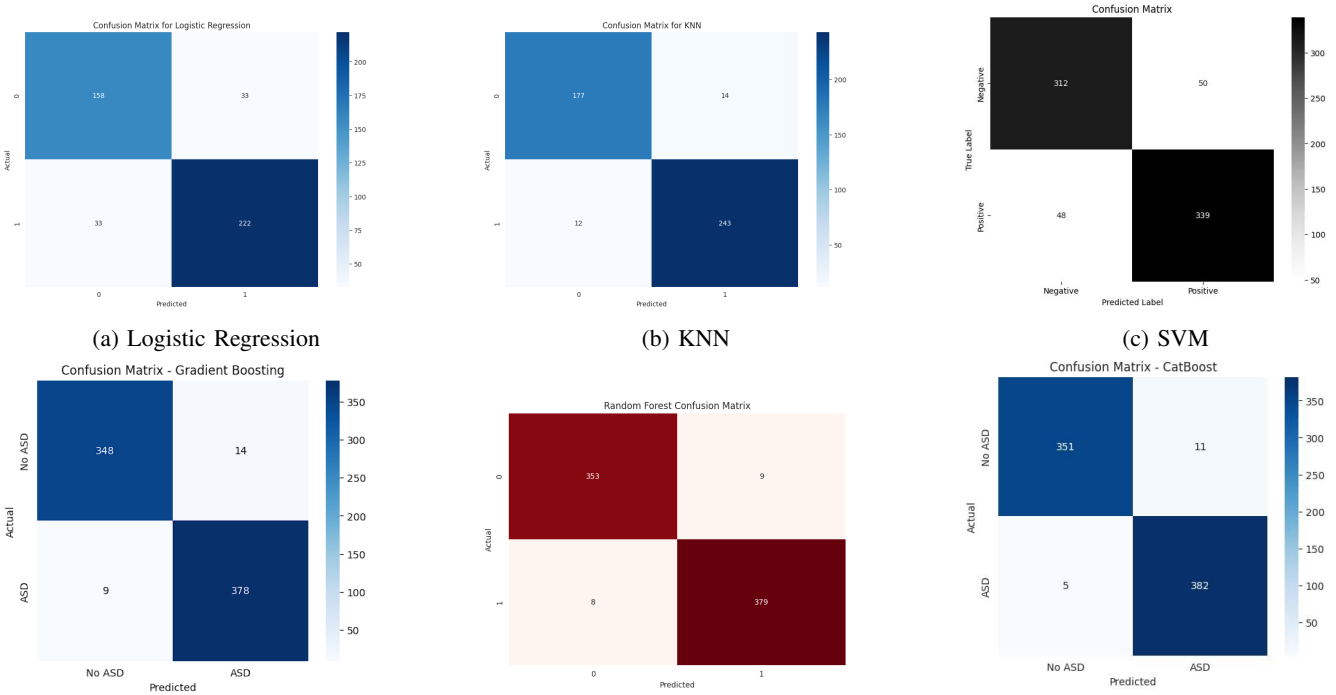
Random forest is a machine learning algorithm that operates by creating an ensemble of decision trees at training time for generating a singular accurate prediction. It is a type of supervised machine-learning algorithm which is very simple. By merging multiple decision trees, random forest can construct a robust model. We applied the Random Forest model to the test dataset and got an impressive accuracy of 97.73%. For a binary classification task, this accuracy is exceptionally high. The result also indicates the model's robust performance. The model correctly predicted 353 true positives and 379 true negatives cases. The false positives and false negatives were 9 and 8 respectively.

F. CatBoost

CatBoost is a supervised machine learning method. It is applied by the Train Using AutoML tool. Decision trees are used for classification and regression purposes in this model. Two important features of CatBoost are managing categorical data and applying gradient boosting. In the gradient boosting technique, many decision trees are built at each iteration. The result of the previous trees can be enhanced by sequential trees. So, CatBoost refined the gradient boosting technique for better results. Also, categorical features are encoded with the help of ordered encoding. For replacing the categorical features, target statistics are used from the rows in order to calculate the value.With an accuracy of 0.97, the CatBoost model demonstrated its effectiveness in classifying ASD traits, with 72 instances of ASD traits and 92 instances of non-ASD traits correctly identified. However, it misclassified 8 instances of non-ASD as ASD and 6 instances of ASD as non-ASD, providing comprehensive insights into its performance.

RESULT ANALYSIS

We used Logistic Regression, KNN, Gradient Boosting, Random Forest, SVM and CatBoost model to detect autism from the behavior dataset. We got some good numbers in some of these models. The accuracy with the CatBoost model was the highest among all. We got the accuracy of 97.86% and for class-1 (autistic) precision, recall and F1 score of 97%, 98% and 98% with the CatBoost model. With Random Forest we got the second highest accuracy of 97.73%. The precision, recall and F1 scores for class-1 with the Random Forest are



(a) Logistic Regression (b) KNN (c) SVM

(d) Gradient Boosting (e) Random Forest (f) CatBoost

Fig. 8: Confusion Matrix for the Machine Learning Models

98%. Gradient Boosting also gave us an excellent result with an accuracy of 96.92%. The Precision, recall and F1 score of Gradient Boosting are 96%, 98% and 97%. With KNN we got the accuracy of 93.59% and the precision, recall and F1 score are 93%, 94% and 94% respectively. SVM model gave us an accuracy of 86.91% and precision, recall and F1 score of 87%, 88% and 88%. Logistic Regression gave us an accuracy of 87.04%. Random Forest and CatBoost model was the most successful among these models. KNN and Gradient Boosting also performed very well with our dataset having accuracy above 95%. CatBoost is the best model for detecting autism from the behavior dataset from our research. In Table 6.1, a comparison between these machine learning is presented. Precision, recall and F1 scores are for class-1(Autistic)

TABLE I: Comparison of Machine Learning Models' Evaluation Metrics

Models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	87.04%	88%	87%	87%
KNN	93.59%	93%	94%	94%
Gradient Boosting	96.92%	96%	98%	97%
Random Forest	97.73%	98%	98%	98%
SVM	86.91%	87%	88%	87%
CatBoost	97.86%	97%	98%	98%

V. CONCLUSION

In our paper, we aimed to introduce diversity by exploring behavioral datasets through machine learning algorithms. These methods were designed to identify autism from behavioral datasets. We tried to ensure that there were fewer chances of misdiagnosis by proper training of machine learning algorithms. The confusion matrix assesses the performance of models that we are using in this research which helps in evaluating the negative and positive classes. Application of different machine learning models like Logistic Regression, Random Forest, CatBoost, GradientBoost, SVM, and KNN helped us detect the characteristics of autism beneficially from the behavioral dataset. Furthermore, leveraging autism features extracted from individual modalities of data will aid in enhancing the accuracy rate.

REFERENCES

[1] S. M. Hasan, M. P. Uddin, M. Al 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, 2022.

[2] P. Mukherjee, S. Sadhukhan, M. Godse, and B. Chakraborty, "Early detection of autism spectrum disorder (asd) using traditional machine learning models," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[3] Y. Lin, Y. Gu, Y. Xu, S. Hou, R. Ding, and S. Ni, "Autistic spectrum traits detection and early screening: A machine learning based eye movement study," *Journal of Child and Adolescent Psychiatric Nursing*, vol. 35, no. 1, pp. 83–92, 2022.

[4] S. Islam, T. Akter, S. Zakir, S. Sabreen, and M. Hossain, "Autism spectrum disorder detection in toddlers for early diagnosis using machine learning. in 2020 ieee asia-pacific conference on computer science and data engineering (csde). 2020," *Gold Coast, Australia*, pp. 1–6.

[5] M. R. Kanhirakadavath and M. S. M. Chandran, "Investigation of eye-tracking scan path as a biomarker for autism screening using machine learning algorithms," *Diagnostics*, vol. 12, no. 2, p. 518, 2022.

[6] 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.