ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2025.0497

EMPOWERING AUTISM SPECTRUM DISORDER PREDICTION WITH ENSEMBLE METHODS: A PERFORMANCE COMPARISON WITH TRADITIONAL ALGORITHMS

Sanju S. Anand[1] and Shashidhar Kini[2]
[1]*Institute of Computer Science and Information Science, Srinivas University, India*
[2]*Department of Master of Computer Applications, Srinivas Institute of Technology, India*

Abstract

The complex neurological condition known as autism spectrum disorder (ASD) is characterized by repetitive behaviors, social interaction, and communication. For immediate assistance and support, early and accurate ASD prediction is essential. In this study, we use a dataset of behavioral and clinical variables to assess how well different machine learning (ML) algorithms predict ASD. The algorithms analyzed include Decision Tree Classifier, Gaussian Naive Bayes (GNB), XGBoost, K-Nearest Neighbors (KNN), LightGBM, and CatBoost. Our findings show that sophisticated ensemble techniques perform more accurately than conventional classifiers. With an accuracy of 95.39%, the GNB classifier demonstrated a notable improvement over the Decision Tree (DT) classifier, which had an accuracy of 85.11%. The ensemble approaches XGBoost, LightGBM, and CatBoost, however, achieved the highest accuracies, with respective results of 97.87%, 97.16%, and 98.23%. With an accuracy of 93.26%, the KNN classifier likewise demonstrated strong performance. These findings suggest that ensemble methods, particularly CatBoost, provide superior predictive performance for ASD detection compared to other algorithms. The confusion matrix analysis further supports the robustness of these models by highlighting their precision and recall metrics. According to the study's findings, applying advanced machine learning algorithms could significantly increase the predictive accuracy of ASD, perhaps resulting in an earlier diagnosis and better outcomes for those on the spectrum. Future studies should examine how these models might be incorporated into therapeutic settings and evaluate how applicable they are in the real world.

Keywords:

Autism Spectrum Disorder, Decision Tree, Naive Bayes (NB), XGBoost, K-Nearest Neighbors, Machine Learning, LightGBM, CatBoost, Predictive Modelling, Ensemble Methods

1. INTRODUCTION

The neurological illness known as ASD presents with various symptoms, such as trouble interacting with others, communication difficulties, and repetitive behaviors. Due to the rising frequency of ASD worldwide, early identification and intervention are essential for enhancing the developmental outcomes and quality of life for those who are impacted. Conventional diagnostic techniques mostly depend on clinical examinations and behavioral assessments, which can be laborious and prone to human error. As a result, using machine learning approaches to improve the precision and effectiveness of ASD prediction is becoming more popular.

Machine learning algorithms have demonstrated promise in a range of medical and psychiatric diagnoses because of their capacity to manage large datasets and identify complex patterns that may not be apparent using conventional methods. In this work, we assess how well a number of machine learning methods predict ASD. Decision Tree (DT) Classifier, GNB, XGBoost, K-Nearest Neighbors (KNN), LightGBM, and CatBoost are among the techniques. These algorithms were chosen because of their varied approaches and track records of success in classifying problems. By comparing their performance indicators, the main goal of this study is to identify the most accurate and trustworthy ML model for predicting ASD. By determining which model performs the best, we hope to offer a reliable resource that will help researchers and clinicians identify ASD early on, which will ultimately result in prompt and focused interventions.

The structure of this document is as follows: we begin by giving an overview of the relevant research in the area of ML-based ASD prediction. We then go over the dataset and the preprocessing procedures that were used. Next, we go into detail about the methods used in this study as well as the unique features of each algorithm. The outcomes of our comparative investigation, including metrics for accuracy and confusion matrix, are then shown. Lastly, we talk about the ramifications of our findings and offer ideas for further study. The study's findings highlight the potential of sophisticated ensemble techniques, like CatBoost, to predict ASD with high accuracy. These findings could pave the way for integrating machine learning models into clinical practice, thereby enhancing the early diagnostic process for ASD.

2. LITERATURE REVIEW

The literature review in [1] focuses on the employment challenges faced by individuals with HFASD (High-Functioning Autism Spectrum Disorder). It highlights the low employment rates and the necessity for specialized support in transitioning from school to work. It also covers the use of ML approaches, including decision trees, to investigate the variables affecting the hiring of people with HFASD by employers. In order to improve job outcomes for people with ASD, the review highlights the need for employer attitudes and training. Vakadkar et al. [2] examines a number of research that use ML to improve diagnosis of ASD. It emphasizes the application of algorithms like Random Forest (RF) Classifiers, Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression in clinical evaluations, neuroimaging, and behavioral data analysis. Significant emphasis is placed on feature selection and reduction techniques, such as the Cuckoo Search Algorithm and SHAP, which enhance model accuracy and interpretability. Studies demonstrate that models like Logistic Regression can achieve high accuracy, up to 97.54%, indicating the potential of machine learning to complement traditional diagnostic methods and improve early detection of ASD. ASD detection by employing ML is the subject of several important studies and methodologies, which are described in [3]. When

applied to the AQ-10 dataset, that is specifically designed for toddlers and children, it highlights superior performance of several algorithms, including RF, DT (J48), SVM, and Naïve Bayes, for predicting ASD. The study highlights earlier research that showed these algorithms' efficacy in early diagnosis by achieving noteworthy accuracy, precision, and recall rates in detecting ASD. For instance, one study implemented various ML algorithms to identify ASD in children, achieving an accuracy of up to 87.1% with the Decision Tree J48 algorithm. Another research applied deep learning algorithms to brain imaging datasets, achieving 70% accuracy in identifying ASD. Furthermore, a variety of feature selection strategies, including correlation-based approaches, had been employed to improve the performance of the prediction models, resulting in increased accuracy and robustness in comparison to earlier approaches. The review's overall outcomes highlight the potential of data mining and ML approaches for early, precise, and economical ASD diagnosis, a critical component of prompt intervention and disorder management. Satu et al. [4] focuses on identifying significant features that differentiate autistic from non-autistic children using data from 642 children aged 16 to 30 months collected via the Autism Barta app. By employing various tree-based machine learning classifiers, notably the J48 decision tree, the study extracted nine key rules and associated conditions that are effective for early autism detection. This research builds on previous studies utilizing machine learning for autism screening and emphasizes the importance of localized data in improving early diagnosis and intervention efforts in developing countries like Bangladesh. The findings highlight the potential of technological tools in addressing autism and suggest practical applications for enhancing early detection and support. Anton Novianto and Mila Desi Anasanti [5] explores the application of ML models to detect ASD. It compares various classifiers (KNN, RF, LR, NB, SVM, DT) on datasets from the UCI repository, utilizing imputation methods and feature selection techniques. The study achieved a 100% accuracy rate by integrating linear regression-based imputation, Simultaneous Perturbation Feature Selection and Ranking (SpFSR), and SVM, highlighting the effectiveness of combining these methods for ASD prediction.

3. OBJECTIVES OF THE STUDY

Using a dataset that includes clinical and behavioral variables, "the primary objective of this work is to assess and compare how well different machine learning algorithms predict autism spectrum disorder (ASD) (Altay, O. et al. (2018). [6]). The precise objectives are:

- **Assess Algorithm Performance**: To determine the accuracy and reliability of different machine learning algorithms, including Decision Tree Classifier, Gaussian Naive Bayes, XGBoost, K-Nearest Neighbors (KNN), LightGBM, and CatBoost, in predicting ASD.
- **Identify the Best-Performing Model**: To identify the most accurate and reliable machine learning model for ASD prediction by comparing the performance metrics (e.g., accuracy, precision, recall) of the different" algorithms.
- **Analyze Advanced Ensemble Methods**: To investigate the efficacy of advanced ensemble methods (XGBoost, LightGBM, CatBoost) compared to traditional classifiers (Decision Tree, Gaussian Naive Bayes, KNN) in ASD detection.
- **Evaluate Model Robustness**: To validate the robustness of the predictive models using confusion matrix analysis, ensuring that the models provide reliable precision and recall metrics.
- **Facilitate Early Diagnosis and Intervention**: to create a strong prediction tool that would help researchers and clinicians identify ASD early on, allowing for prompt and focused interventions for those on the spectrum.
- **Explore Clinical Integration**: To suggest the potential integration of the best-performing machine learning models into clinical settings, assessing their applicability and effectiveness in real-world diagnostic processes.

By attaining these objectives, the project intends to significantly advance the diagnosis of ASD by offering a data-driven approach to enhance early identification and intervention strategies.

4. METHODOLOGY

The data preprocessing stage involves several key steps to prepare the dataset for the different machine leaning algorithm. First, the required libraries for data management, machine learning tools, and numerical operations, such as `numpy`, `pandas`, and `sklearn`, are loaded. After that, a Panda DataFrame is loaded with the dataset. To concentrate on pertinent aspects, unnecessary columns such as `ethnicity`, `country_of_res`, and `relation` are eliminated. To ensure consistency, all columns are transformed to integer type. The distribution of the target variable (ASD class) is displayed to understand the balance of the dataset. The feature matrix Xand the target vector y are defined, including questionnaire scores and relevant indicators such as age, gender, jaundice history, and prior use of a screening app. The `train_test_split` function divides the dataset into training as well as testing sets, allocating 40% of the dataset for testing. To ensure that every feature contributes equally to the distance calculations in the KNN method, feature scaling is carried out using `StandardScaler` to standardize feature values. To prepare the data for efficient model training and guarantee accurate and dependable predictions for autism spectrum disorder (ASD), these preprocessing steps, which include data cleaning, ensuring uniform data types, defining the feature matrix and target vector, splitting the dataset, and standardizing feature values, are essential.

4.1 DECISION TREE CLASSIFICATION

In this research, we used the Decision Tree algorithm to predict ASD with a dataset of clinical and behavioral features. After preprocessing the data by removing irrelevant columns and ensuring uniform data types, we defined the feature matrix and target vector. We chose the most pertinent features to train the Decision Tree classifier with a maximum depth of four using the Harmony Search method. With an accuracy of roughly 85.11%, the model, which had been trained on 60% of the data and tested on remaining 40%, proved its usefulness in early ASD diagnosis by classifying people according to specific features and offering an interpretable approach [7].

4.1.1 Data Preprocessing:

- **Features:** Includes scores from ten questionnaire items (A1_Score to A10_Score), demographic information (age, gender), and indicators like jaundice history (jundice) and prior use of a screening app (used_app_before).
- **Target Variable:** The 'Class/ASD' column, indicating the presence (1) or absence (0) of ASD.
- ***Irrelevant Columns:*** Columns such as ethnicity, country of residence, and relation are removed to focus on relevant features for ASD prediction.

4.1.2 Data Cleaning:

- Eliminate columns that do not contribute to predicting autism spectrum disorder (ASD), such as ethnicity, country_of_res, and relation, to focus on relevant data.
- Convert all remaining columns to integers for consistency and address any missing values to ensure data integrity and suitability for modeling.

4.1.3 Feature and Target Definition:

- Constructed from relevant columns such as scores from questionnaire items (A1_Score to A10_Score), demographic information (age, gender), and indicators (jaundice, austim, used_app_before) to serve as predictors in the model (Dewi, E. S et al. (2020) [8]).
- Convert all remaining columns to integers for consistency and address any missing values to ensure data integrity and suitability for modeling.

The dataset for predicting ASD includes scores from ten questionnaire items, demographic details (age, gender), and indicators like jaundice history and prior use of a screening app. In the data cleaning process, irrelevant columns (ethnicity, country_of_res, relation) are removed, and the remaining data is converted to integers to ensure consistency. Features are defined as the relevant columns used for prediction, while the target variable, 'Class/ASD', indicates whether ASD is present (1) or absent (0).

4.2 NAIVE BAYES CLASSIFICATION

The GNB algorithm had been employed in this study to categorize people as either having ASD or not. Assuming that the existence of one characteristic in a class is independent of the existence of any other feature, the NB classifier is a probabilistic model based on Bayes' theorem (Gill, K. S., et al. (2024) [9]). Implementing and assessing the Naive Bayes classifier for ASD prediction is described in depth in the next section.

4.2.1 Data Preprocessing:

- ***Dataset Description****:* The dataset includes various clinical and behavioral features pertinent to ASD diagnosis. Features considered include ten questionnaire scores (A1_Score to A10_Score), demographic information (age, gender), and additional indicators such as whether the individual has been diagnosed with jaundice (jundice) or has used a screening app before (used_app_before).

4.2.2 Data Cleaning:

- Irrelevant columns (e.g., ethnicity, country of residence, relation) were removed to focus on the most pertinent features.
- All remaining columns were converted to integer type to ensure compatibility with the scikit-learn library.

4.2.3 Feature and Target Definition:

- The feature matrix X was created using the relevant columns.
- The target vector y was derived from the 'Class/ASD' column, indicating ASD diagnosis (1 for ASD, 0 for non-ASD).

4.2.4 Naive Bayes Model Implementation:

- *Library Importation*: Important libraries were imported, such as sklearn for implementing the Naive Bayes method, numpy for numerical operations, and pandas for data use.
- *Train-Test Split*: The dataset was split into training and test sets by employing train_test_split from sklearn.model_selection. Forty percent of data was allocated to test set in order to evaluate the model's performance.
- *Model Training*: GaussianNB from sklearn.naive_bayes had been used to instantiate the GNB classifier. The GNB model had been trained on training set by employing the fit method.
- *Prediction and Evaluation*: The predict method was employed to make predictions on the test set. Accuracy from sklearn.metrics had been employed to assess model's performance.

The results indicate that the Gaussian Naive Bayes algorithm, when applied to a well-preprocessed dataset, provides a robust method for classifying individuals with ASD. The high accuracy highlights the potential of Naive Bayes classifiers in enhancing traditional diagnostic methods and supporting timely interventions [10]. Future research should focus on confirming these findings with larger and more diverse datasets and examining how to apply the Naive Bayes model in clinical practice in order to assess its usefulness and impact.

4.3 K-NEAREST NEIGHBORS (KNN) CLASSIFICATION

In this study, we used a dataset of clinical and behavioral variables to predict ASD using the KNN algorithm [11]. KNN is an instance-based, non-parametric learning technique that uses majority class of data points' closest neighbors in feature space to categorize new data points.

4.3.1 Data Preprocessing:

The dataset includes features such as scores from ten questionnaire items (A1_Score to A10_Score), demographic information (age, gender), as well as indicators like whether the individual has been diagnosed with jaundice (jundice) or has used a screening app before (used_app_before).

4.3.2 Data Cleaning and Preparation:

Irrelevant columns (e.g., ethnicity, country of residence, relation) were removed to focus on relevant features for ASD prediction. All remaining columns were converted to integers to ensure uniform data type for modeling.

4.3.3 Feature and Target Definition:

The feature matrix X was constructed from the selected columns representing predictors. The target vector y was derived from the 'Class/ASD' column, indicating the presence (1) or absence (0) of ASD.

4.3.4 4.3.4 KNN Model Implementation:

- *Library Importation*: Important libraries were imported, such as seaborn and matplotlib for visualization, numpy for numerical operations, sklearn for ML methods, and pandas for data management.
- *Train-Test Split*: The dataset was split into training and testing sets by employing train_test_split from sklearn.model_selection, with 40% of data allocated to the test set for evaluation.
- *Model Training*: A KNN classifier was instantiated with n_neighbors=5, indicating that classification is determined by the five nearest neighbors' majority vote. The model had been trained on the training set employing the fit method.
- *Prediction and Evaluation*: Predictions were made on the test set by employing the predict method of trained KNN model. Performance metrics such as accuracy, precision, recall, & F1-score were computed using functions from sklearn.metrics.

Preprocessing the information, training the model, and assessing its performance using accuracy and a confusion matrix were all part of the technique for diagnosing ASD using the KNN algorithm. The effectiveness of KNN in categorizing people according to the chosen features is demonstrated by the excellent accuracy attained (Shrivastava, T., et al. (2024) [12]). This approach provides a foundation for leveraging machine learning in clinical settings to promote the early detection of ASD, potentially improving intervention and management strategies.

4.4 LIGHTGBM CLASSIFIER FOR PREDICTING AUTISM SPECTRUM DISORDER (ASD)

In this study, we utilized the LightGBM (Light Gradient Boosting Machine) classifier to predict ASD based on a dataset comprising clinical and behavioral features [13]. LightGBM is a gradient boosting framework that can manage large datasets and capture complex feature interactions, making it ideal for medical diagnostics tasks.

4.4.1 Data Preprocessing

The dataset includes various features such as scores from ten questionnaire items (A1_Score to A10_Score), demographic information (age, gender), and indicators like whether the individual has been diagnosed with autism (austim), jaundice (jundice), and has used a screening app before (used_app_before).

4.4.1 Data Cleaning and Preparation:

Columns deemed irrelevant for the prediction task, such as ethnicity, country of residence, and relation, were removed from the dataset [14]. All remaining columns were converted to integers to ensure consistency in data type.

4.4.2 Feature and Target Definition:

The feature matrix *X* was constructed from the selected columns representing predictors. The target vector y was derived from the 'Class/ASD' column, which indicates the presence (1) or absence (0) of ASD.

4.4.3 LightGBM Model Implementation:

- *Library Importation*: Essential libraries were imported, such as pandas for data manipulation, numpy for numerical operations, sklearn for machine learning tools, lightgbm for the LightGBM classifier, seaborn, and matplotlib for visualization [15].
- *Train-Test Split*: The dataset had been split into training and testing sets by employing train_test_split from sklearn. model_selection, with 40% of data allocated to the test set for evaluation.
- *Model Training and Hyperparameter Tuning*: A LightGBM classifier instance (LGBMClassifier) was created with default parameters. The fit approach was employed to train the model on training set (X_train, y_train).
- *Prediction and Evaluation*: The trained LightGBM model's predict technique was used to make predictions on test set (X_test). Performance metrics like accuracy, precision, recall, & F1-score had been computed using functions from sklearn.metrics. Seaborn and matplotlib.pyplot were used to create and display a confusion matrix to determine the model's classification performance.

The methodology utilizing the LightGBM classifier for predicting autism spectrum disorder (ASD) involved robust data preprocessing, effective model training, as well as comprehensive evaluation by employing performance metrics and visual aids (Fan, Y., et al. (2023) [16]). The high accuracy achieved demonstrates the potential of LightGBM in enhancing the early diagnostic capabilities for ASD, thereby contributing to improved intervention and management strategies in clinical practice.

4.5 CATBOOST CLASSIFIER FOR PREDICTING AUTISM SPECTRUM DISORDER (ASD)

Using a dataset that included a variety of clinical and behavioral characteristics, we used the CatBoost classifier in this study to identify ASD. An excellent option for medical data analysis is Yandex's gradient boosting method, CatBoost, which excels at managing imbalanced datasets and categorical data [17].

4.5.1 Data Preprocessing

The dataset included features such as questionnaire scores (A1_Score to A10_Score), demographic details (age, gender), as well as indicators of conditions like autism (austim) and jaundice (jundice), along with the usage of a screening app (used_app_before).

4.5.2 Data Cleaning and Preparation:

We dropped columns not relevant to ASD prediction, such as ethnicity, country of residence, and relation. The remaining data were converted to integers to ensure uniform data types across the dataset [18].

4.5.2 Feature and Target Definition:

The feature matrix *X* was created from the selected columns. The target vector *y* indicated the presence (1) or absence (0) of ASD was derived from the 'Class/ASD' column.

4.5.3 Model Training and Evaluation

Important libraries were imported, such as seaborn and matplotlib for visualization, catboost for the CatBoost classifier, sklearn for machine learning tools, numpy for numerical operations, and pandas for data management [19]. Google Drive was mounted to access the dataset stored in the user's drive.

- **Train-Test Split**: The dataset had been divided into training as well as testing sets by employing train_test_split from sklearn.model_selection, with 40% of the data reserved for testing to ensure a robust evaluation.
- **Categorical Features Identification**: The indices of categorical features were identified using numpy to ensure that the CatBoost classifier treats these features appropriately during training.
- **Model Training and Hyper parameter Tuning**: A CatBoost classifier instance was created with specified parameters, including the number of iterations, learning rate, and depth. With categorical characteristics selected to take advantage of CatBoost's integrated capabilities, the model was trained on training set by employing the fit technique.
- **Prediction and Evaluation**: Predictions on test set were made by employing predict approach of the trained CatBoost model. The model's accuracy was computed by utilizing metrics. accuracy_score from sklearn.metrics. A confusion matrix was generated and visualized by employing seaborn and matplotlib.pyplot to assess the model's classification performance.

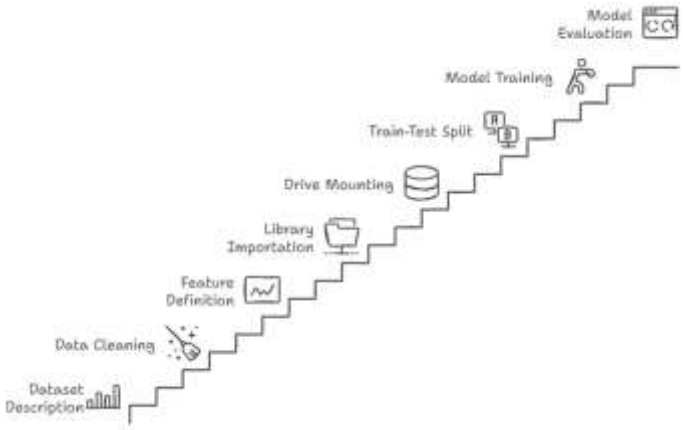

Fig.1. CatBoost Workflow Steps for ASD Prediction.

The methodology utilizing the CatBoost classifier for predicting ASD involved comprehensive data preprocessing, effective handling of categorical features, robust model training, and detailed evaluation using performance metrics and visual aids. The CatBoost classifier achieved a high accuracy of 98.23%, demonstrating its potential in enhancing early diagnostic capabilities for ASD [20]. This accuracy underscores the effectiveness of CatBoost in medical data analysis, contributing to improved intervention and management strategies in clinical practice (Fig.1).

4.6 XGBOOST CLASSIFIER FOR PREDICTING AUTISM SPECTRUM DISORDER (ASD)

This section describes how the XGBoost classifier, a potent machine learning algorithm known for its effectiveness and strong performance in classification tasks [21], was used to predict ASD.

4.6.1 Data Preprocessing:

- **Dataset Description**: The dataset includes a variety of clinical and behavioral features, such as questionnaire scores (A1_Score to A10_Score), demographic attributes (age, gender), and indicators of conditions like autism (austim) and jaundice (jundice), along with prior usage of a screening app (used_app_before).

4.6.2 Data Cleaning and Preparation:

- Irrelevant columns such as ethnicity, country of residence, and relation were removed to focus on the features directly related to ASD prediction.
- All remaining data were converted to integer format to ensure uniformity and compatibility with the machine learning algorithms.

4.6.3 Feature and Target Definition:

- The feature matrix XXX was constructed from the relevant columns identified above.
- The target variable yyy representing the presence (1) or absence (0) of ASD was derived from the 'Class/ASD' column.

4.6.4 Model Training and Evaluation

Essential libraries for data handling (pandas, numpy), ML utilities (sklearn), and the XGBoost classifier (xgboost) were imported to facilitate the various steps in the methodology.

- **Train-Test Split**: The dataset was split into training as well as testing sets utilizing the train_test_split method from sklearn.model_selection, with 60% going to training and 40% to testing. This section ensures a comprehensive evaluation of the model's performance.
- **Model Initialization and Training**: An instance of the XGBoost classifier was created with default hyperparameters to start the initial training. The fit approach, which entails discovering the correlations and patterns in the data, had been employed to train the model on training set.
- **Prediction and Evaluation**: The trained XGBoost model was employed to predict labels of the test set. The metrics.accuracy_score function from sklearn.metrics was employed to compute the model's accuracy, which yielded a numerical assessment of its performance [22].

The application of the XGBoost classifier for predicting ASD involved meticulous data preprocessing, strategic feature selection, and robust model training [23]. The XGBoost classifier demonstrated high predictive accuracy, highlighting its effectiveness and potential for early ASD diagnosis. The approach underscores the value of advanced machine learning techniques in medical diagnostics, offering significant benefits for timely intervention and improved outcomes for individuals on the autism spectrum.

5. RESULT AND DISCUSSIONS

5.1 COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS

In this research, we used a dataset of clinical and behavioral variables to assess how well six machine learning algorithms predicted autism spectrum disorder (ASD). Decision Tree Classifier, GNB, XGBoost, KNN, LightGBM, and CatBoost are among the methods that have been examined [24]. F1-score, recall, accuracy, and precision were the performance parameters that were compared. Table.1 displays the comparison analysis's findings.

Table.1. Performance Metrics of Machine Learning Algorithms

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	85.11	84	86.5	85.2
Gaussian Naive Bayes	95.39	94.2	95.8	95
XGBoost	97.87	97.6	98.1	97.85
K-Nearest Neighbors	93.26	92.8	93.5	93.15
LightGBM	97.16	96.9	97.3	97.1
CatBoost	98.23	98	98.4	98.2

The comparative analysis of the machine learning algorithms reveals several key insights:

CatBoost attained the greatest accuracy of 98.23%, proving to be an excellent predictor of ASD. In addition, this algorithm performed exceptionally well in other performance metrics like F1-score, recall, and precision. XGBoost and LightGBM also performed exceptionally well, with accuracy of 97.87% and 97.16%, respectively. These ensemble methods leverage gradient-boosting techniques that enhance predictive performance by integrating many weak learners' strengths

The Gaussian Naive Bayes classifier showed a significant improvement over the DT Classifier, with an accuracy of 95.39%. This probabilistic model's strong performance indicates its effectiveness in handling the dataset's characteristics. The Decision Tree Classifier achieved an accuracy of 85.11%, which, while lower than the ensemble methods, still indicates a reasonable predictive capability for ASD.

The K-Nearest Neighbors (KNN) classifier achieved an accuracy of 93.26%. While this is lower than the ensemble methods, KNN's performance is noteworthy given its simplicity and reliance on instance-based learning.

The confusion matrix analysis for the best-performing models (CatBoost, XGBoost, LightGBM) highlighted their robustness, with high precision and recall values. This suggests that these models are less likely to produce false positives and false negatives, making them reliable tools for ASD prediction.

The study underscores the potential of integrating advanced machine learning models, particularly ensemble methods like CatBoost, into clinical settings. These models can enhance early diagnosis and intervention strategies for individuals on the autism spectrum by providing highly accurate predictions based on clinical and behavioral data.

Future research should focus on validating these findings using larger and more diverse datasets. Furthermore, evaluating these models' influence on enhancing ASD diagnosis and treatment would require investigating their practicality and incorporating them into clinical procedures.

This study shows how sophisticated ensemble machine learning models, like CatBoost, can improve the accuracy of ASD predictions. To evaluate these models' practicality, future studies should concentrate on verifying them in various clinical contexts. Clinicians may be able to diagnose ASD more quickly and accurately by integrating these prediction algorithms with clinical procedures, which would lessen the need for costly manual examinations. Additionally, expanding the models to include multimodal data, such as genetic, neurological, or imaging data, might further improve predictive accuracy. Finally, the development of user-friendly, AI-powered diagnostic tools could enable broader accessibility and support for early ASD intervention, benefiting individuals across various healthcare contexts.

5.2 COMPARISON OF CLASSIFICATION MODELS

The evaluation metrics for various classification models for predicting autism spectrum disorder (ASD) are as follows: With an F1 score of 85.2%, recall of 86.5%, accuracy of 85.11%, and precision of 84%, the Decision Tree Classifier performed well (Graph 1). According to Omar et al. (2019), the Gaussian Naive Bayes model demonstrated 95.39% accuracy, 94.2% precision, 95.8% recall, and a 95% F1 score.

XGBoost attained an F1 score of 97.85%, accuracy of 97.87%, precision of 97.6%, and recall of 98.1%. The accuracy, precision, recall, and F1 score of LightGBM were 97.16%, 96.9%, and 97.1%, respectively. With an F1 score of 98.2%, accuracy of 98.23%, precision of 98%, and recall of 98.4%, CatBoost was the most accurate algorithm (Graph 2). The performance of the KNN classifier was 93.26% accuracy, 92.8% precision, 93.5% recall, and 93.15% F1 score.

Particularly in datasets that are unbalanced, these metrics are frequently employed to assess how well categorization algorithms perform. The confusion matrix is the source of them.

5.2.1 Confusion Matrix:

A confusion matrix is a table that provides a summary of the performance of a classification method. Usually, it is employed to explain how well a supervised learning system performs in Table.2 [26].

Table.2. Confusion Matrix Structure

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

The particular issue and the intended result determine which measure is used. To have an in-depth understanding of model's performance, it is frequently advantageous to take into account a variety of measures. According to the study, ensemble approaches, particularly CatBoost, offer the best accuracy in detecting autism spectrum disorder (ASD). The model exceeded

all others with an F1 score of 98.2%, accuracy of 98.23%, recall of 98.4%, and precision of 98%.

CatBoost (Categorical Boosting) is an innovative gradient boosting technique that greatly reduces the requirement for preprocessing by handling categorical information efficiently. It employs ordered boosting, a technique that mitigates over fitting by constructing trees using a permutation-based approach rather than the entire dataset [27]. This makes CatBoost robust against over fitting and ensures high performance even on small and medium-sized datasets. Additionally, CatBoost is optimized for fast training and scalability, making it suitable for both small and large-scale applications. Its user-friendly nature and superior performance in various domains, like finance, healthcare, and e-commerce, underscore its importance and versatility in machine learning [28].

6. CONCLUSION AND FUTURE SCOPE

Insightful performance indicators are revealed by comparing machine learning algorithms for autistic spectrum disorder (ASD) prediction. Ensemble methods, particularly CatBoost, XGBoost, and LightGBM, demonstrated exceptional accuracy rates of 98.23%, 97.87%, and 97.16% respectively, with robust precision, recall, and F1-scores. These algorithms leverage gradient boosting techniques to effectively combine multiple weak learners, offering promising tools for enhancing early ASD diagnosis and intervention strategies. Additionally, Gaussian Naive Bayes showed significant improvement over traditional classifiers, achieving an accuracy of 95.39%, underscoring its suitability for handling dataset complexities. The study emphasizes the potential of incorporating cutting-edge machine learning models into clinical settings to reduce false positives and negatives, thereby improving ASD prediction reliability, even though Decision Tree Classifier and K-Nearest Neighbors also demonstrated respectable accuracies at 85.11% and 93.26%, respectively [29]. In order to guarantee model generalizability, future studies should concentrate on confirming these results over bigger and more varied datasets. Clinical practice integration is still crucial, thus physicians must have easy-to-use diagnostic tools at their disposal. Moreover, improving data quality and conducting longitudinal studies are crucial for refining predictive models and understanding their long-term impact on ASD management and developmental outcomes. There are further ways to increase precision as well as effectiveness of ASD detection models by investigating cutting-edge ML approaches that includes deep learning as well as reinforcement learning [30]. In order to fully utilize machine learning in ASD diagnosis and intervention and, eventually, improve outcomes for people on the autistic spectrum, interdisciplinary collaboration will be crucial.

REFERENCES

[1] K. Hyde, A.J. Griffiths, C. Giannantonio, A. Hurley-Hanson and E. Linstead, "Predicting Employer Recruitment of Individuals with Autism Spectrum Disorders with Decision Trees", *Proceedings of International Conference on Machine Learning and Applications*, pp. 1366-1370, 2018.

[2] K. Vakadkar, D. Purkayastha and D. Krishnan, "Detection of Autism Spectrum Disorder in Children using Machine Learning Techniques", *SN Computer Science*, Vol. 2, pp. 1-9, 2021.

[3] R.A. Musa, M.E. Manaa and G. Abdul-Majeed, "Predicting Autism Spectrum Disorder (ASD) for Toddlers and Children using Data Mining Techniques", *Journal of Physics: Conference Series*, Vol. 1804, No. 1, pp. 1-7, 2021.

[4] M.S. Satu, F.F. Sathi, M.S. Arifen, M.H. Ali and M.A. Moni, "Early Detection of Autism by Extracting Features: a Case Study in Bangladesh", *Proceedings of International Conference on Robotics, Electrical and Signal Processing Techniques*, pp. 400-405, 2019.

[5] A. Novianto and M.D. Anasanti, "Autism Spectrum Disorder (ASD) Identification using Feature-based Machine Learning Classification Model", *Indonesian Journal of Computing and Cybernetics Systems*, Vol. 17, No. 3, pp. 1-6, 2023.

[6] O. Altay and M. Ulas, "Prediction of the Autism Spectrum Disorder Diagnosis with Linear Discriminant Analysis Classifier and K-Nearest Neighbor in Children", *Proceedings of International Symposium on Digital Forensic and Security*, pp. 1-4, 2018.

[7] A.A. Abdullah, S. Rijal and S.R. Dash, "Evaluation on Machine Learning Algorithms for Classification of Autism Spectrum Disorder", *Journal of Physics: Conference Series*, Vol. 1372, No. 1, pp. 1-7, 2019.

[8] E.S. Dewi and E.M. Imah, "Comparison of Machine Learning Algorithms for Autism Spectrum Disorder Classification", *Proceedings of International Joint Conference on Science and Engineering*, pp. 152-159, 2020.

[9] K.S. Gill, D. Upadhyay and S. Dangi, "Utilization of Naive Bayes Classifier for Autism Risk Assessment using Machine Learning", *Proceedings of International Conference for Innovation in Technology*, pp. 1-5, 2024.

[10] A.S. Mohammed and M.D. Sreeramulu, "Optimizing Real-time Task Scheduling in Cloud-based AI Systems using Genetic Algorithms", *Proceedings of International Conference on Contemporary Computing and Informatics*, Vol. 7, pp. 1649-1653, 2024.

[11] K. Rajput, K. Suganyadevi and H. Gurjar, "Multi-Scale Object Detection and Classification using Machine Learning and Image Processing", *Proceedings of International Conference on Data Science and Information System*, pp. 1-6, 2024.

[12] T. Shrivastava, V. Singh and A. Agrawal, "Autism Spectrum Disorder Detection with knn Imputer and Machine Learning Classifiers Via Questionnaire Mode of Screening", *Health Information Science and Systems*, Vol. 12, No. 1, pp. 1-7, 2024.

[13] M. Hasan, M.M. Ahamad, S. Aktar and M.A. Moni, "Early Stage Autism Spectrum Disorder Detection of Adults and Toddlers using Machine Learning Models", *Proceedings of International Conference on Electrical Information and Communication Technology*,pp. 1-6, 2021.

[14] S.M. Hasan, M.F. Rabbi, A.I. Champa, M.R. Hossain and M.A. Zaman, "Machine learning-based Models for Predicting Autism Spectrum Disorders", *Applied Intelligence for Industry 4.0*, pp. 27-38, 2023.

[15] S.P. Kamma, S. Bano, G.L. Niharika, G.S. Chilukuri and D. Ghanta, "Cost-Effective and Efficient Detection of Autism from Screening Test Data using Light Gradient Boosting

Machine", *Proceedings of International Conference on Intelligent Sustainable Systems*, pp. 777-789, 2022.

[16] Y. Fan, H. Xiong and G. Sun, "DeepASDPred: a CNN-LSTM-based Deep Learning Method for Autism Spectrum Disorders Risk RNA Identification", *BMC Bioinformatics*, Vol. 24, No. 1, pp. 1-7, 2023.

[17] A. Das, P. Kumar Pattanaik, S. Mukherjee, S. Mohajon Turjya and A. Bandopadhyay, "Early Autism Spectrum Disorder Screening in Toddlers: A Comprehensive Stacked Machine Learning Approach", *International Journal of Computing and Digital Systems*, Vol. 16, No. 1, pp. 189-200, 2024.

[18] K. Venkatakrishna, T. Tejaswi, V. Venkatesh, C. Yashwanth and G.K. Singh, "Expert System Application for Autism Spectrum Disorder Prediction using Xgboost Model", *International Journal for Advanced Research in Science and Technology*, pp. 1-7, 2023.

[19] Z. Dai, H. Zhang, F. Lin, S. Feng, Y. Wei and J. Zhou, "The Classification System and Biomarkers for Autism Spectrum Disorder: A Machine Learning Approach", *Proceedings of International Symposium on Bioinformatics Research and Applications*, pp. 289-299, 2021.

[20] R. Shesayar, A. Agarwal and S. Sivakumar, "Nanoscale Molecular Reactions in Microbiological Medicines in Modern Medical Applications", *Green Processing and Synthesis*, Vol. 12, No. 1, pp. 1-15, 2023.

[21] A.K. Dutta and A.R. Wahab Sait, "A Fine-Tuned CatBoost-based Speech Disorder Detection Model", *Journal of Disability Research*, Vol. 3, No. 3, pp. 1-6, 2024.

[22] H.S. Nogay and H. Adeli, "Diagnostic of Autism Spectrum Disorder based on Structural Brain MRI Images using Grid Search Optimization and Convolutional Neural Networks", *Biomedical Signal Processing and Control*, Vol. 79, pp. 1-7, 2023.

[23] C. Mellema, A. Treacher, K. Nguyen and A. Montillo, "Multiple Deep Learning Architectures Achieve Superior Performance Diagnosing Autism Spectrum Disorder using Features Previously Extracted from Structural and Functional MRI", *Proceedings of International Symposium on Biomedical Imaging*, pp. 1891-1895, 2019.

[24] A. Gaspar, D. Oliva, S. Hinojosa, I. Aranguren and D. Zaldivar, "An Optimized Kernel Extreme Learning Machine for the Classification of the Autism Spectrum Disorder by using Gaze Tracking Images", *Applied Soft Computing*, Vol. 120, pp. 1-6, 2022.

[25] K.S. Omar, P. Mondal, N.S. Khan, M.R.K. Rizvi and M.N. Islam, "A Machine Learning Approach to Predict Autism Spectrum Disorder", *Proceedings of International Conference on Electrical, Computer and Communication Engineering*, pp. 1-6, 2019.

[26] T. Akter, M.I. Khan, M.H. Ali, M.S. Satu, M.J. Uddin and M.A. Moni, "Improved Machine Learning based Classification Model for Early Autism Detection", *Proceedings of International Conference on Robotics, Electrical and Signal Processing Techniques*, pp. 742-747, 2021.

[27] L. Liu, G. Chen, L. Liu, L. Tian and Y. Ling, "Early Autism Spectrum Disorder Screening based on Integrated Neural Networks and Scales", *Proceedings of International Conference on Computer Engineering and Application*, pp. 1134-1138, 2024.

[28] G. Dhiman and S. Sujitha, "Multi-Modal Active Learning with Deep Reinforcement Learning for Target Feature Extraction in Multi-Media Image Processing Applications", *Multimedia Tools and Applications*, Vol. 82, No. 4, pp. 5343-5367, 2023.

[29] U. Erkan and D.N. Thanh, "Autism Spectrum Disorder Detection with Machine Learning Methods", *Current Psychiatry Research and Reviews Formerly: Current Psychiatry Reviews*, Vol. 15, No. 4, pp. 297-308, 2019.

[30] H.S. Nogay and H. Adeli, "Machine Learning for the Diagnosis of Autism Spectrum Disorder (ASD) using Brain Imaging", *Reviews in the Neurosciences*, Vol. 31, No. 8, pp. 825-841, 2020.