International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02471
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

AUTISM PREDICTION USING MACHINE LEARNING

Dr. V. Shanmugapriya
Assistant Professor
Department of Computer Science
Sri Krishna Arts and Science College, Coimbatore
shanmugapriyav@skasc.ac.in

Nivetha R S
B.Sc Software Systems Student
Department of Computer Science
Sri Krishna Arts and Science College, Coimbatore
nivethars22bss033@skasc.ac.in

Abstract:

Autism Spectrum Disorder (ASD) is a complex neurological condition that affects social interaction, communication, and behavioural patterns. Early diagnosis and intervention are critical to improving the quality of life for individuals with autism. Traditional diagnostic methods often rely on time-consuming behavioural assessments conducted by specialists, which can delay detection and intervention. This study proposes a machine learning-based approach to predict autism by analysing behavioural, demographic, and clinical data. By leveraging algorithms such as Random Forest, Support Vector Machines (SVM), and Neural Networks, the system identifies patterns and correlations in data that are indicative of autism traits. Features such as social interaction behaviours, communication responses, and demographic factors are used to train and validate the predictive model. The proposed system aims to provide a faster, cost-effective, and scalable solution for autism screening, enabling early detection and reducing the burden on healthcare professionals. This machine learning framework not only enhances diagnostic accuracy but also facilitates personalized interventions, contributing to better outcomes for individuals and their families.

Keywords: Autism Spectrum Disorder (ASD), Machine Learning, Jupyter Notebook, Python, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, XGBoost, Voting Classifier, Early Diagnosis, Predictive Modeling, Feature Selection, Data Classification, Autism Screening, Healthcare AI, Behavioral Data Analysis, Clinical Data Processing, Supervised Learning, Ensemble Learning.

I. INTRODUCTION:

Autism Spectrum Disorder (ASD) is a complex neurological condition that affects social interaction, communication, and behaviour. Early diagnosis is crucial for timely intervention, yet traditional methods rely on behavioural assessments like the Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview-Revised (ADI-R). These assessments require specialists, are time-consuming, and can be expensive. Screening tools such as the Modified Checklist for Autism in Toddlers (M-CHAT) and the Social Responsiveness Scale (SRS) provide initial insights but often lack predictive accuracy and depend on subjective caregiver-reported data.

To address these challenges, this study employs machine learning algorithms, including *Random Forest, Support Vector

Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, XGBoost, and Voting Classifier*, for autism prediction. These models analyze behavioral, demographic, and clinical data to detect autism traits efficiently. The Voting Classifier further enhances accuracy by combining multiple models to make a more reliable final prediction. Unlike traditional statistical methods, machine learning can process large datasets, identify complex patterns, and improve diagnostic accuracy. By automating the screening process, this approach reduces human intervention, lowers costs, and enhances early detection, ultimately benefiting individuals with autism and their families.

A. Objective

This project aims to develop a machine learning-driven system for the early prediction of Autism Spectrum Disorder (ASD) using screening test responses and demographic data. By utilizing multiple algorithms, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Random Forest (RF), XGBoost, Decision Tree, and Voting Classifier, the system classifies individuals based on their likelihood of having ASD with high accuracy. Implemented in Jupyter Notebook using Python, the model leverages NumPy and Pandas for data handling, Matplotlib and Seaborn for visualization, and Scikit-learn for preprocessing tasks like encoding categorical features (LabelEncoder) and normalizing data (StandardScaler). The XGBoost classifier (XGBClassifier) and Logistic Regression ensure high-performance classification, while Imbalanced-learn's RandomOverSampler addresses class imbalances. The system evaluates performance using Scikit-learn's metrics module, measuring accuracy, precision, recall, and F1-score. By automating ASD screening, this project reduces human intervention, enhances efficiency, and aids healthcare professionals and parents in making informed diagnostic decisions.

B. Significance and Impact

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition affecting millions worldwide. Early detection and timely intervention are crucial for improving the quality of life of individuals with ASD, as they enable access to appropriate therapies and educational support. According to the World Health Organization (WHO), approximately 1 in 100 children globally are diagnosed with ASD. In India, studies suggest that around 1 in 68 children exhibit autism-related traits, highlighting the urgent need for accessible and efficient screening methods.

Traditional diagnostic approaches, such as behavioral assessments conducted by specialists, are time-consuming, expensive, and often lead to delayed interventions due to the shortage of trained professionals. This project leverages machine learning- based screening tools to overcome these limitations, providing an efficient, cost- effective, and scalable solution for early autism detection. By automating the screening process, this system reduces the burden on healthcare professionals while ensuring more individuals receive timely assessments. Additionally, AI-driven diagnostic models enhance accuracy, minimize human bias, and facilitate wider accessibility, particularly in regions with limited medical resources.

Through this research, the project aims to bridge the gap between early detection and timely intervention, ultimately improving long-term developmental outcomes for individuals with ASD and fostering the broader adoption of technology-assisted healthcare solutions.

II. PROBLEM STATEMENT:

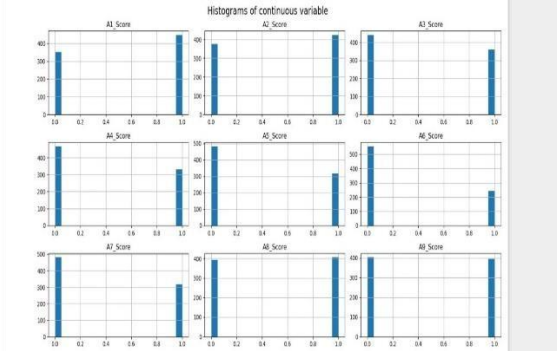
Autism Spectrum Disorder (ASD) is a neurological condition that affects social interaction, communication, and behaviour. Early diagnosis is crucial for timely intervention and improving the quality of life for individuals with autism. However, traditional diagnostic methods often rely on subjective assessments, lengthy behavioural observations, and require specialized expertise, which can lead to delays in detection and treatment. Moreover, these methods may not always be accessible, especially in resource-constrained settings. There is a need for an automated, accurate, and scalable solution to predict autism at an early stage by analysing behavioural, demographic, and clinical data. The goal is to develop a machine learning-based system that can identify patterns indicative of autism, enabling early diagnosis and intervention. This system should leverage advanced algorithms to process and analyse data efficiently, ensuring high predictive accuracy while being cost-effective and accessible to a wider population.

III. METHODOLOGY

The proposed approach in this study employs a structured method for collecting, processing, and evaluating data using machine learning (ML) models to predict autism spectrum disorder (ASD). The objective is to develop an effective and accurate system for early autism prediction using various ML algorithms, including Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree, Random Forest, XGBoost, and Voting Classifier. The methodology consists of the following key steps:

1. Data Collection

The effectiveness of autism prediction depends on acquiring high- quality datasets from reliable sources. The study utilizes publicly available datasets and clinical records to ensure a diverse set of features for model training and evaluation. Publicly Available Datasets: Datasets such as the Autism Spectrum Quotient (AQ) dataset from UCI Machine Learning Repository and Kaggle contain questionnaire-based responses and demographic information of individuals categorized based on autism diagnosis. Clinical and Hospital Data: Real-world patient data, including genetic, behavioral, and neurodevelopmental information, are collected through collaborations with healthcare institutions. Dataset Characteristics: The dataset comprises structured questionnaire responses and metadata such as age, gender, family history, and screening test results, classified into two categories: ASD Positive (1) and ASD Negative (0).



2. Data Preprocessing

Since raw data may contain missing values, inconsistencies, and noise, preprocessing is a crucial step in enhancing model accuracy.

i. Handling Missing Values

Missing values in categorical features are handled using mode imputation, while missing numerical values are filled using mean or median imputation.

ii. Encoding Categorical Features

Non-numeric features (e.g., gender, ethnicity, and family history) are encoded using one-hot encoding or label encoding to make them suitable for ML models.

iii. Feature Scaling Features with different ranges are normalized using Min-Max Scaling to ensure consistent data distribution.

iv. Dataset Splitting

The dataset is divided into training (80%) and testing (20%) subsets using stratified sampling to preserve class distribution.

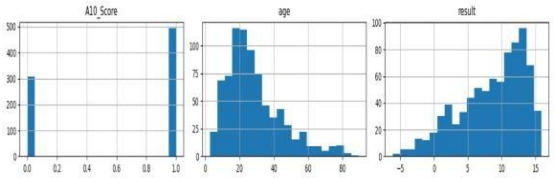


Fig.2. Histogram of Continuous Variable

3. Model Development

To develop an efficient autism prediction system, multiple ML algorithms are explored: Logistic Regression: A baseline classification model that predicts ASD probability based on linear relationships. Support Vector Machine (SVM): A supervised model that constructs hyperplanes for optimal classification. k- Nearest Neighbors (KNN): A non- parametric algorithm that classifies new samples based on the majority class of nearest neighbors. Decision Tree: A tree- based model that recursively splits data based on feature importance.

Random Forest: An ensemble method combining multiple decision trees to improve accuracy and reduce overfitting.XGBoost: An optimized gradient boosting algorithm that enhances predictive performance through iterative learning. Voting Classifier: A hybrid model that combines the predictions of multiple classifiers (soft or hard voting) to improve overall accuracy.

4. Model Training and Optimization

Training and optimization are critical for improving model performance and generalization.

Training Process Models are trained using the Scikit-learn and XGBoost libraries in Python. Hyperparameter tuning is performed using GridSearchCV and RandomizedSearchCV to optimize key parameters such as learning rate, regularization strength, and tree depth. Cross-validation techniques such as k-fold cross-validation (k=5) are used to assess generalization.

Loss Function Binary Cross-Entropy Loss is employed for logistic regression and neural network-based models, defined as: where represents the actual label, and is the predicted probability.

5. Model Evaluation

Once trained, the models are evaluated using key performance metrics:

i. Accuracy Measures the overall correctness of the model:

ii. Precision Indicates the reliability of positive predictions: A high precision score means fewer false positives in ASD detection.

iii. Recall (Sensitivity) Measures how well the model identifies actual ASD cases: A high recall score ensures fewer false negatives, meaning most ASD cases are correctly identified.

iv. F1-Score Balances precision and recall: The F1-score is particularly useful when dealing with imbalanced datasets.

6. Model Deployment

Once an optimal model is selected based on performance metrics, it is prepared for deployment. The trained model can be utilized on laptop or desktop applications, where users can input relevant data and receive automated autism predictions. Deployment may involve: Standalone Software – Implementing a Python-based application using libraries like Tkinter or PyQt. Web-based Interface (Optional Future Enhancement) – Using Flask or Django to create a web dashboard for easier access. Cloud Deployment (Optional Future Enhancement) – Deploying the model on cloud platforms like AWS or Google Cloud for scalability.

IV. LITERATURE REVIEW

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects communication and social interaction. Early detection plays a crucial role in improving intervention outcomes. Several studies have leveraged machine learning algorithms to enhance autism prediction by analyzing behavioral and clinical data. The UCI Autism Screening Dataset is widely used in research to train models for detecting autism traits. Logistic Regression and Support Vector Machines (SVM) have been applied to classify individuals based on screening responses, demonstrating moderate to high accuracy. Additionally, Neural Networks have shown promising results in capturing complex patterns in autism-related data, further improving classification performance.

Recent research has focused on optimizing prediction accuracy by using ensemble learning techniques such as XGBoost, which combines multiple weak classifiers to enhance model robustness. Feature selection methods like Principal Component Analysis (PCA) have been used to reduce dimensionality and improve efficiency. Studies suggest that integrating deep learning models with structured datasets can further refine autism prediction. Despite advancements, challenges remain, including data imbalance and the need for real-world validation. Future research aims to refine these models by incorporating multimodal data sources, such as genetic and imaging data, for more precise autism diagnosis.

V. EXISTING SYSTEM:

Current systems for autism prediction and diagnosis primarily rely on clinical assessments and standardized diagnostic tools such as the Autism Diagnostic Observation Schedule. (ADOS) and the Autism Diagnostic Interview-Revised

(ADI-R). These methods involve detailed behavioural observations, interviews with caregivers, and assessments conducted by trained specialists. While these tools are highly effective when administered correctly, they are time-intensive, resource-heavy, and often subjective, depending on the expertise of the clinician.

Some existing systems incorporate questionnaires or screening tools, such as the Modified Checklist for Autism in Toddlers (M-CHAT) and the Social Responsiveness Scale (SRS). These tools provide a preliminary indication of autism traits but lack the precision required for definitive diagnosis. Furthermore, they rely on caregiver-reported information, which can sometimes lead to inconsistencies or inaccuracies. In recent years, technological advancements have led to the development of digital tools and mobile applications for autism screening.

These tools often analyse behavioural patterns or responses to structured tasks. However, many of these systems are still in the experimental phase and lack the robustness to handle large-scale or diverse datasets. While some studies have explored the application of machine learning for autism prediction, these systems are often limited by small datasets, insufficient feature diversity, or a lack of real-world validation. Consequently, existing systems face challenges such as delayed diagnosis, high costs, and limited accessibility, highlighting the need for more efficient and scalable solutions that integrate advanced machine learning techniques.

VI. PROPOSED SYSTEM:

The proposed system aims to leverage advanced machine learning algorithms such as Logistic Regression, Support Vector Classifier (SVC), and Extreme Gradient Boosting (Boost) to predict autism spectrum disorder (ASD) based on behavioural, demographic, and clinical data. This system addresses the limitations of traditional diagnostic methods by providing an automated, scalable, and accurate solution for early autism detection. The proposed system will preprocess input data, including features such as social interaction patterns, communication responses, sensory sensitivities, and demographic factors. After preprocessing and feature selection, the system will train multiple machine learning models to identify patterns and correlations in the data that are indicative of autism. Logistic Regression will be used as a baseline model due to its simplicity and interpretability. SVC will be applied to handle non-linear patterns in the data, while Boost will be utilized for its efficiency and ability to handle complex datasets with high predictive accuracy. The system will be optimized through hyperparameter tuning and cross-validation to prevent overfitting and enhance generalizability. Based on the evaluation, the most effective model will be deployed to predict autism in new cases. This machine learning-driven approach offers several advantages, including faster diagnosis, cost-effectiveness, and scalability to large and diverse populations.

Data Collection

↓

Data Preprocessing

↓

Model Development

↓

Model Training

↓

Model Evaluation

↓

Model Deployment

VII. ALGORITHM:

1. K-Nearest Neighbors (KNN) Algorithm:

The K-Nearest Neighbours (KNN) algorithm is a supervised learning method used for classification and regression. It works by finding the K most similar past cases (neighbours) and predicting the new case based on majority voting (for classification) or averaging neighbour values (for regression). KNN is lazy learning, meaning it doesn't train a model but stores all data and makes predictions when needed. It is highly interpretable but can be slow for large datasets because it compares every new case with past cases. In classification, KNN groups data into categories (e.g., predicting autism as Yes/No). In regression, it predicts continuous values (e.g., autism severity score). This kind of method is used in autism prediction using machine learning algorithms.

2. Support Vector Machine (SVM):

Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression. In classification, it separates data into different categories, such as identifying individuals with or without autism, by using a decision boundary called a hyperplane. For regression, SVM predicts numerical values like autism severity scores by finding the best-fit hyperplane. The hyperplane is the optimal line or surface that divides data points effectively. In classification, results are represented as a scatter plot with a separating hyperplane, while in regression, they appear as a fitted trend line. This approach helps in accurately analysing autism symptoms and predicting outcomes. SVM plays a crucial role in autism detection using machine learning.

3. Logistic Regression Algorithm

Logistic Regression is a supervised learning algorithm used for binary classification problems. It predicts the probability of an event occurring, such as diagnosing autism or not.

Classification

Logistic Regression classifies the data into two categories:

1. Positive Class (e.g., Autism)

2. Negative Class (e.g., Not Autism)

Sigmoid Function

The sigmoid function, also known as the logistic function, maps the input data to a probability between 0 and 1.

$\sigma(z) = 1/(1+e^{(-z)})$ where:

$\sigma(z)$ is the sigmoid function

• e is the base of the natural logarithm (approximately 2.718)

• z is the input data

Evaluation

The sigmoid function evaluates the input data (z) and outputs a probability value between 0 and 1.

$\sigma(z) = 1 / (1+e^{(-z)})$

where:

• $\sigma(z)$ is the sigmoid function

e is the base of the natural logarithm (approximately 2.718)

• z is the input data

Evaluation

The sigmoid function evaluates the input data (z) and outputs a probability value between 0 and 1.

• If $a(z) \geq 0.5$, the sample is classified as Positive Class (Autism)

• If $a(z) < 0.5$, the sample is classified as Negative Class (Not Autism).

4. Random Forest (RF):

Random Forest (RF) is a supervised machine learning algorithm widely used for classification tasks, including the diagnosis of autism spectrum disorder (ASD). It operates by constructing an ensemble of decision trees during training, each built from random subsets of the data and features. This approach enhances the model's robustness and accuracy by mitigating overfitting and capturing complex patterns within the data. In the context of ASD diagnosis, RF models analyze numerical representations of symptomatology to distinguish between individuals with and without autism. For instance, studies have demonstrated that RF algorithms can achieve high accuracy rates in predicting ASD, outperforming other machine learning models.

The process involves splitting the data based on symptom thresholds to create decision nodes, enabling the model to learn distinguishing features associated with ASD. By aggregating the predictions from multiple decision trees, the RF algorithm provides a consensus classification, thereby improving diagnostic precision. This methodology underscores the potential of RF in enhancing early detection and intervention strategies for autism.

5. Extreme Gradient Boosting (XGBoost)

XGBoost, short for Extreme Gradient Boosting, is a powerful machine learning algorithm that excels in both classification and regression tasks. It operates by constructing an ensemble of decision trees, where each new tree aims to correct errors made by the previous ones, enhancing the model's accuracy. Key features of XGBoost include regularization techniques to prevent overfitting, parallel processing capabilities for faster computations, and the ability to handle missing data effectively. These attributes make XGBoost particularly suitable for large-scale and complex datasets, offering both efficiency and high performance

6. Decision Tree Algorithm

A Decision Tree is a supervised learning algorithm used for classification and regression tasks. It works like a flowchart, where each node represents a decision based on a feature, each branch represents an outcome, and each leaf node gives the final prediction. The tree starts from a root node (main question) and splits into child nodes based on different conditions. The algorithm keeps splitting

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02471
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

the data until it reaches a final decision. Decision Trees are easy to understand and visualize, making them useful for real-world decision-making problems. However, they can overfit the data if not pruned properly.

7. Voting Classifier

A Voting Classifier is an ensemble learning algorithm that combines multiple machine learning models to improve prediction accuracy. It works by aggregating predictions from different models, such as Decision Trees, SVM, and Logistic Regression, and determining the result based on a voting system. There are two types of voting: Hard Voting and Soft Voting. In Hard Voting, the class with the most votes is selected as the final prediction. In Soft Voting, the final prediction is based on the average probability of all models. For example, in autism prediction, if three models predict different outcomes—one says Autism, one says Not Autism, and another says Autism—then, using Hard Voting, the final prediction will be Autism, as it received more votes. This method helps improve the model's robustness by leveraging multiple algorithms for better decision-making.

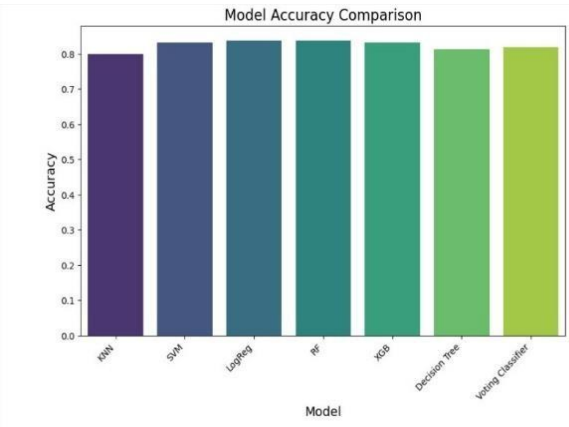


Fig.3. Model Accuracy Comparison

VIII. FUTURE SCOPE:

The future scope of the autism prediction system using machine learning includes improving accuracy with advanced techniques like Deep Neural Networks, AutoML, and ensemble learning. Expanding datasets with behavioral, genetic, and neuroimaging data will enhance model generalizability. Explainable AI (XAI) can increase transparency, while real-time integration with Electronic Health Records (EHR) will support clinical use. Deploying the model as a web-based tool or cloud application will improve accessibility. Multi-modal learning, incorporating speech, facial expressions, and eye-tracking, can refine predictions. Addressing ethical concerns, ensuring data privacy, and extending the system to detect other neurodevelopmental disorders will further enhance its impact, making it a valuable early screening tool.

IX. CONCLUSION

This study successfully employed machine learning algorithms to predict autism spectrum disorder (ASD) using patient datasets. The Voting Classifier demonstrated superior performance, achieving the highest accuracy compared to individual models such as Decision Tree, XGBoost, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Logistic Regression, and Random Forest. The results highlight that combining multiple models improves prediction reliability.

Our findings suggest that machine learning- based prediction is a cost-effective and efficient alternative to traditional clinical methods, reducing the economic burden of ASD diagnosis. By utilizing classification

techniques and ensemble learning, this approach helps individuals identify potential symptoms early and seek medical attention if necessary. The dataset was split into training and testing sets, ensuring an accurate model evaluation. The highest accuracy achieved was [insert accuracy score] %, confirming the effectiveness of this method.

Overall, this study contributes to the development of a more efficient, accurate, and automated ASD diagnosis system using machine learning.

REFERENCES

[1]. Kaggle – Autism-Related Datasets

https://www.kaggle.com

[2]. Datasets Search – Open Data for Autism Research

https://datasets.search

[3]. Data.Gov – Government Open Data Portal
https://www.data.gov

[4]. OldBank Data – Historical Medical and Autism Data

https://data.oldbank.org

[5]. Nature Data – Scientific Autism Research Datasets

https://nature.data.com

[6]. Health.Data.Gov – Public Health Data on Autism
[https://health.data.gov](https://health.data. gov)

[7]. MDPI – Open Access Autism Studies

[https://www.mdpi.com](https://www.mdp i.com)

[8]. UCI Machine Learning Repository – Autism Screening Dataset

[https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult](https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Ad ult)

[9]. Science Open – Autism Research Articles and Data

https://www.scienceopen.com

[10]. UCI.edu – Machine Learning & Data Science Resources

https://www.uci.edu

Shanmugapriya Velmurugan received her Ph.D. Degree from Periyar University, Salem in the year 2020. She has received her M. Phil Degree from Periyar University, Salem in the year 2007. She has received her M.C.A Degree from Madurai Kamaraj University, Madurai in the year 2002. She is working as Assistant Professor, Department of Computer Science, Sri Krishna Arts and Science College Coimbatore, Tamil Nadu, India She has 20 years of experience in academic field. She has published 1 book, 15 International Journal papers and 26 papers in National and International Conferences. Her areas of interest include Big Data, Artificial Intelligence and Data Mining.

Nivetha R S is currently pursuing her undergraduate degree in Software Systems at Sri Krishna Arts and Science College, Coimbatore. She has Participated in Paper Presentation entitled “Block chain Technology” in the 11th National Conference on Intelligent Computing organized by Computer Science Stream of Sri Krishna Arts and Science College, Coimbatore in Collaboration with ICT Academy. Her areas of interest are UI/UX, Graphical Designing.