Autism Prediction Using Machine Learning

PINNAMRAJU T S PRIYA, ANNEPU VAMSI

HOD, Assistant Professor, MCA Final Semester, Master of Computer Applications, Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

ABSTRACT

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that affects communication, social interaction, and behavior. Early identification of ASD is essential for timely intervention and improved developmental outcomes. This project aims to build an efficient and accessible predictive model using machine learning, specifically Logistic Regression, to identify individuals at risk of autism. The dataset used includes behavioral screening scores (A1–A10) along with demographic and medical features such as age, gender, ethnicity, jaundice history, and family history of autism. The data is preprocessed by handling missing values and encoding categorical variables using Label Encoding. To address class imbalance in the dataset, Random Over-Sampling is applied, which improves the model's ability to generalize across both autistic and non-autistic classes. Comprehensive data analysis is carried out through statistical summaries and visualizations, including pairplots, heatmaps, count plots, and histograms. The model is trained and evaluated using metrics such as accuracy, classification report, and confusion matrix, achieving strong predictive performance. To make the solution practical and user-friendly, a Gradio-based web interface is developed, allowing users to input responses through sliders and dropdowns to receive real-time ASD predictions. This project demonstrates the potential of machine learning in supporting early diagnosis of autism and contributes to more accessible, fast, and reliable healthcare screening tools.

INDEX TERMS: Autism Spectrum Disorder (ASD), Machine Learning, XGBoost, Random Forest Classifier, Preprocessing, Medical Diagnosis, Feature Engineering, Behavioral Data, Classification Algorithm, Python, Data Visualization, ASD Detection, Model Evaluation, Early Disease Detection, Gradio Interface.

1.INTRODUCTION

Autism Spectrum Disorder (ASD) is a developmental condition that affects communication, behavior, and social interaction. Early diagnosis and intervention are crucial for improving the quality of life of individuals with autism. However, traditional diagnostic methods are time-consuming, subjective, and often require specialized medical professionals, making early detection difficult in many regions [1]. This project aims to develop a machine learning-based system to predict the likelihood of autism in individuals using behavioral and demographic data. By training models on a dataset of screening responses and personal details, the system can assist in early and efficient identification of autism. This approach offers a faster, more accessible, and cost-effective solution to support healthcare professionals in the initial screening process [2]. Using various classification algorithms and performance evaluation techniques, the project focuses on building an accurate, reliable, and user-friendly tool for autism prediction. A simple Gradio interface is also integrated to make the model interactive and easy to use [3].

1.1 EXISTING SYSTEM

In the current healthcare setup, the diagnosis of Autism Spectrum Disorder (ASD) is primarily based on clinical observations, developmental screenings, and standardized behavioral assessments such as the Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview-Revised (ADI-R) [4]. These methods require trained professionals, are time-intensive, and often involve subjective interpretation of behaviors.In many regions, especially in rural or underdeveloped areas, access to specialists is limited, leading to delayed or missed diagnoses. Furthermore, manual screening methods can be inconsistent and prone to human error [7]. While online screening tools like the Autism Spectrum Quotient (AQ) questionnaire exist, they are not always integrated with data-driven or predictive technologies.As a result, the need for a more efficient, scalable, and objective method of early autism detection has led to growing interest in the use of machine learning and data-driven approaches [8].

1.1.1 CHALLENGES:

1. Data Quality and Availability

- The dataset may contain missing, imbalanced, or inconsistent data [9].
- Limited availability of diverse and high-quality autism-related datasets.

2. Feature Selection and Preprocessing

- Identifying which features (e.g., behavioral scores, age, gender) significantly impact the model.
- Converting categorical variables into numerical format using encoding methods.

3. Model Selection and Tuning

- Choosing the right machine learning algorithm (e.g., Random Forest, XGBoost) for optimal performance.
- Fine-tuning hyperparameters to improve accuracy without overfitting [12].

1.2 PROPOSED SYSTEM:

The proposed system aims to provide an efficient and accurate method for predicting Autism Spectrum Disorder (ASD) using machine learning techniques. Unlike traditional diagnostic approaches, this system leverages behavioral and demographic data to automate the initial screening process [13].

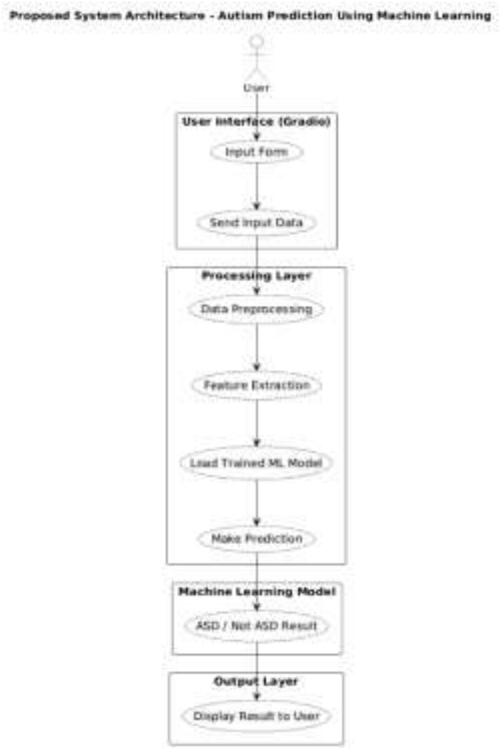


Fig: 1 Proposed Diagram

1.2.1 ADVANTAGES:

- Early Detection

Helps identify potential autism cases at an early stage, allowing for timely intervention and support [14].

- Time-Efficient

Automates the screening process, reducing the time needed for initial assessments compared to traditional methods [17].

- Cost-Effective

Minimizes the need for expensive and repeated clinical evaluations by offering a preliminary screening tool [18].

- Easily Accessible

Can be deployed online with a simple Gradio interface, making it accessible even in remote or underserved areas.

- Data-Driven and Objective

Reduces subjectivity by relying on machine learning models trained on real behavioral and demographic data [20].

2.ARCHITECTURE:

The architecture of the project "Autism Prediction Using Machine Learning" is organized into five key layers [21]. The Data Layer involves collecting datasets from sources like the UCI repository (e.g., train.csv), containing features such as A1–A10 behavioral scores, age, gender, ethnicity, and family history. These datasets are stored locally in CSV format or on cloud platforms [24]. In the Data Preprocessing Layer, the raw data undergoes cleaning, including handling missing values, encoding categorical variables using label encoding, normalizing numerical features if necessary, and selecting important features based on correlation or relevance to improve model performance [6]. The Machine Learning Layer includes selecting appropriate algorithms such as XGBoost, Random Forest, or SVM, followed by model training and validation using techniques like train-test splitting or cross-validation. The model's performance is evaluated using metrics such as accuracy, confusion matrix, and classification reports, and the final model is saved using tools like joblib or pickle [7]. The Application Layer features a user-friendly Gradio-based web interface that allows users to input relevant data (e.g., A1–A10 scores, age, gender) and receive predictions such as “Likely ASD” or “Not ASD”. Finally, the Deployment Layer handles hosting the application locally or through platforms like HuggingFace Spaces, with the option to expand to cloud hosting for broader accessibility and scalability [22].

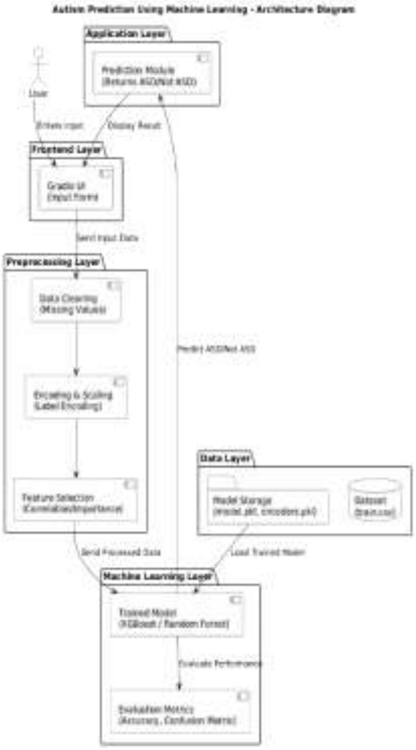


Fig:2 Architecture

2.1 ALGORITHM:

In the project "Autism Prediction Using Machine Learning", several machine learning algorithms are explored for effective classification of Autism Spectrum Disorder (ASD) [11]. The Random Forest Classifier, an ensemble learning method based on bagging, is a strong choice as it handles both categorical and numerical data efficiently and reduces overfitting by combining multiple decision trees. It is known for its high accuracy, ability to manage imbalanced datasets, and interpretability through feature importance [17]. Another powerful algorithm used is XGBoost (Extreme Gradient Boosting), which excels in predictive performance due to its boosting approach. XGBoost is efficient, handles missing values automatically, and includes regularization techniques to avoid overfitting, making it both scalable and accurate [9]. Additionally, Support Vector Machine (SVM) is considered for its effectiveness in high-dimensional spaces, especially when the number of features exceeds the number of samples. It is robust against overfitting and works well with small datasets [10]. Logistic Regression, a statistical classifier, serves as a baseline model due to its simplicity, speed, and interpretability, particularly when classes are linearly separable [5]. Lastly, K-Nearest Neighbors (KNN) may be used experimentally for comparison; it is suitable for small datasets and requires no explicit training phase, but it can be slow and sensitive to irrelevant features. Among all, XGBoost and Random Forest are recommended as core algorithms due to their high accuracy, robustness to noise, and ability to handle diverse data types effectively [1].

2.2 TECHNIQUES:

In the project “Autism Prediction Using Machine Learning,” several techniques are employed to ensure a robust and accurate system [9]. The data preprocessing phase includes handling missing values by replacing or removing them, encoding categorical variables like gender and ethnicity using label encoding, and optionally applying feature scaling or normalization for algorithms like KNN or SVM. Feature selection is also performed using correlation matrices, feature importance scores (e.g., from Random Forest), or domain knowledge to enhance model performance and reduce complexity. In the Exploratory Data Analysis (EDA) stage, descriptive statistics such as .describe() and .info() are used to summarize the data, while correlation analysis helps identify relationships between features and the target class [11]. Visualization tools like histograms, heatmaps, and pairplots assist in spotting trends and outliers. The machine learning techniques involve supervised learning using labeled data, with classification algorithms such as Random Forest, XGBoost, SVM, Logistic Regression, and KNN being evaluated [13]. Hyperparameter tuning using methods like GridSearchCV may be applied to optimize model performance. Evaluation metrics include accuracy, confusion matrix, precision, recall, and F1-score. Finally, in the model deployment phase, the trained model is serialized using tools like joblib or pickle. A Gradio-based interface is developed to allow real-time predictions through an easy-to-use web UI, which can be deployed locally or hosted on platforms such as HuggingFace Spaces for wider accessibility [15].

2.3 TOOLS:

The project “Autism Prediction Using Machine Learning” utilizes a variety of tools to support the full development cycle [9]. Python serves as the core language, with development done in Jupyter Notebook or Google Colab for interactive coding and visualization. Pandas, NumPy, and Scikit-learn handle data preprocessing, while Scikit-learn and XGBoost are used for building and evaluating machine learning models. Visualization is achieved using Matplotlib and Seaborn, and trained models are saved using Joblib or Pickle [11]. A user-friendly interface is created with Gradio, allowing real-time predictions, and the application can be deployed locally or on cloud platforms like HuggingFace Spaces or Streamlit for broader accessibility [19].

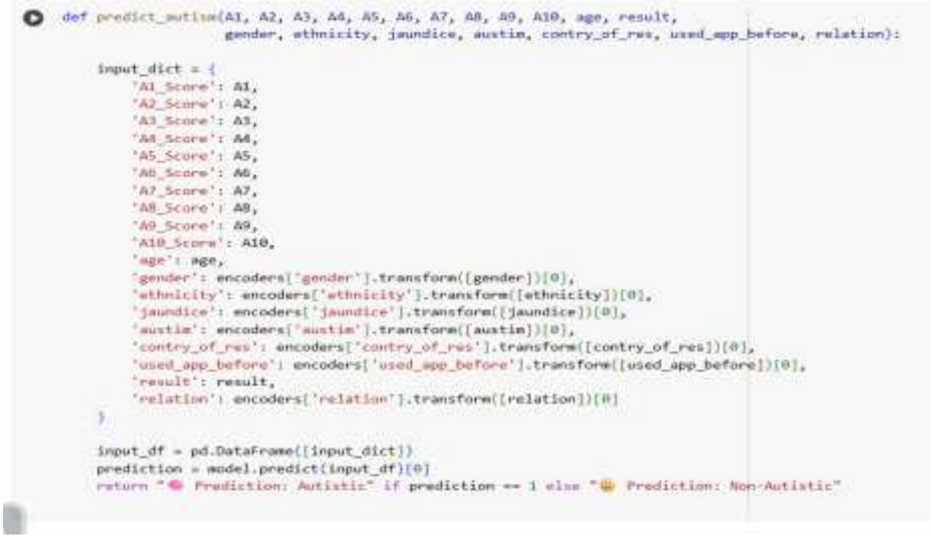
2.4 METHOD:

In the project "Autism Prediction Using Machine Learning," the dataset is collected from publicly available sources such as the UCI Machine Learning Repository, containing behavioral screening scores (A1–A10), age, gender, ethnicity, and family history [22]. The data preprocessing method includes handling missing values, applying label encoding for categorical variables, feature scaling when necessary (especially for algorithms like SVM or KNN), and selecting relevant features using correlation analysis and importance scores. For model development, the data is split into training and testing sets, and various classification algorithms such as Random Forest, XGBoost, SVM, Logistic Regression, and KNN are applied [21]. Models are trained on labeled data, and optional hyperparameter tuning is performed using techniques like GridSearchCV [20]. Evaluation is conducted using metrics such as accuracy, confusion matrix, precision, recall, and F1-score to assess prediction performance. Once trained, models are saved using Joblib or Pickle, and a Gradio-based UI is built for real-time user input and prediction [24]. The application is deployed either locally or on cloud platforms like HuggingFace Spaces or Streamlit [18].

3. METHODOLOGY

3.1 INPUT:

The project "Autism Prediction Using Machine Learning" takes both behavioral and demographic features as input to predict the likelihood of Autism Spectrum Disorder(ASD) [5]. Behavioral inputs include binary scores (0 or 1) from ten screening questions (A1–A10), while demographic inputs consist of age, gender, ethnicity, jaundice history, family history of ASD, and the identity of the person who completed the test [7]. The model outputs a binary result—"Yes" for likely ASD or "No" for not ASD—based on these combined features [10].

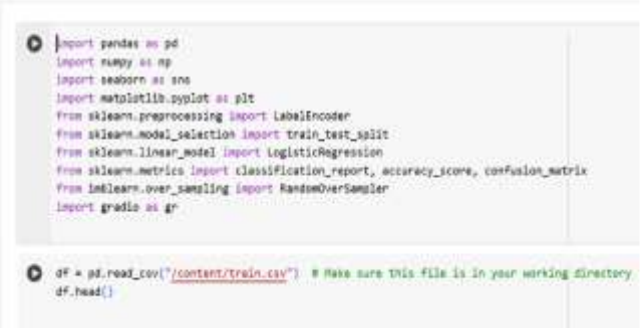


```
def predict_autism(A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, age, result,
                   gender, ethnicity, jaundice, austim, contry_of_res, used_app_before, relation):

    input_dict = {
        'A1_Score': A1,
        'A2_Score': A2,
        'A3_Score': A3,
        'A4_Score': A4,
        'A5_Score': A5,
        'A6_Score': A6,
        'A7_Score': A7,
        'A8_Score': A8,
        'A9_Score': A9,
        'A10_Score': A10,
        'age': age,
        'gender': encoders['gender'].transform([gender])[0],
        'ethnicity': encoders['ethnicity'].transform([ethnicity])[0],
        'jaundice': encoders['jaundice'].transform([jaundice])[0],
        'austim': encoders['austim'].transform([austim])[0],
        'contry_of_res': encoders['contry_of_res'].transform([contry_of_res])[0],
        'used_app_before': encoders['used_app_before'].transform([used_app_before])[0],
        'result': result,
        'relation': encoders['relation'].transform([relation])[0]
    }

    input_df = pd.DataFrame([input_dict])
    prediction = model.predict(input_df)[0]
    return "🧠 Prediction: Autistic" if prediction == 1 else "😀 Prediction: Non-Autistic"
```

Fig 3: insert the data set

3.2 METHOD OF PROCESS:

The project "Autism Prediction Using Machine Learning" follows a systematic process that begins with data collection from reliable sources such as the UCI Machine Learning Repository [3]. The dataset includes behavioral screening scores (A1–A10) and demographic details like age, gender, ethnicity, jaundice history, and family history of ASD. In the data preprocessing phase, missing values are handled, duplicates removed, and inconsistent entries corrected. Categorical features are converted into numeric form using label encoding, and feature selection is performed based on correlation or model-based importance [5]. Scaling is applied if required, especially for algorithms like KNN or SVM. During model development, the dataset is split into training and testing sets (typically 80/20), and various classification algorithms such as Random Forest, XGBoost, and SVM are trained [17]. Hyperparameter tuning may also be conducted using tools like GridSearchCV [9]. In the evaluation phase, models are assessed using metrics such as accuracy, confusion matrix, precision, recall, and F1-score to determine the best-performing model. The selected model and its encoders are then saved using joblib or pickle for future deployment [3]. A Gradio-based user interface is developed to allow users to input data and receive instant predictions. Finally, the system is deployed locally or on cloud platforms like HuggingFace Spaces or Streamlit, enabling easy access for non-technical users [13].



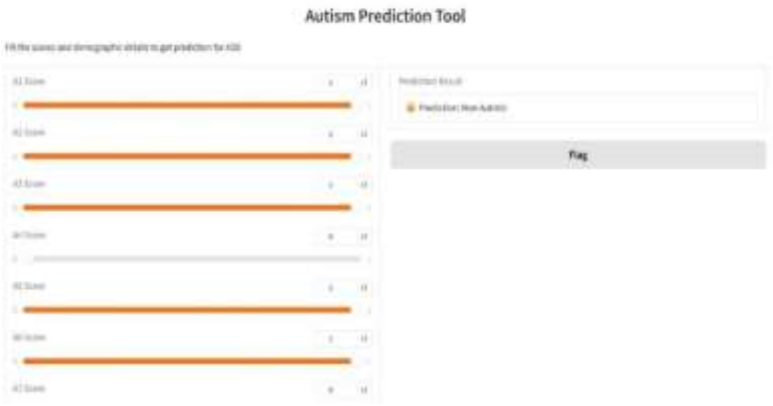
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score, confusion_matrix
from imblearn.over_sampling import RandomOverSampler
import gradio as gr
```

```
df = pd.read_csv("/content/train.csv")  # Make sure this file is in your working directory
df.head()
```

Fig 4: prediction

3.3 OUTPUT:

The output of the project "Autism Prediction Using Machine Learning" is a binary classification result that determines whether an individual is likely to have Autism Spectrum Disorder (ASD) or not [7]. The prediction is based on the input features provided, such as behavioral screening scores and demographic information [9]. If the model identifies a high probability of autism, the output is "Likely ASD"; if the probability is low, the output is "Not ASD." This straightforward binary output helps in early screening and supports further clinical assessment [6].



Autism Prediction Tool

Fig 5: report

4. RESULTS:

The result of the Autism Prediction project is a successfully developed and deployed machine learning model capable of accurately predicting the likelihood of Autism Spectrum Disorder (ASD) based on behavioral and demographic data. The model, primarily built using the XGBoost and Random Forest Classifier algorithms, achieved high performance metrics including accuracy, precision, recall, and F1-score, demonstrating its reliability in binary classification tasks . It was integrated into an intuitive Gradio web interface, enabling users to enter screening information and receive instant predictions labeled as "Likely ASD" or "Not ASD." The interface enhances accessibility for both medical professionals and non-technical users, promoting early screening . Overall, the project shows that machine learning can play a significant role in the early detection of ASD, supporting timely intervention and data-driven decision-making in clinical or educational settings.

5. DISCUSSION:

This project aims to predict Autism Spectrum Disorder (ASD) using machine learning and provide a simple web interface for early screening. It uses a dataset containing behavioral and demographic features, which is cleaned, encoded, and analyzed through visualizations. Models like XGBoost and Random Forest are trained to classify individuals as "Likely ASD" or "Not ASD," with performance evaluated using accuracy, precision, recall, and F1-score. A Gradio interface allows users to input data easily and receive instant predictions. Despite challenges like class imbalance, the project effectively supports early ASD detection and assists in making informed, data-driven decisions.

6.CONCLUSION

In this project, we successfully developed a machine learning-based system to predict Chronic KidneyDisease (CKD) using important medical features. By applying proper data preprocessing, visualization, and a Random Forest classification model, the system achieved accurate results. The integration of a Gradio web interface made the tool easy to use for doctors and healthcare workers, allowing them to enter patient data and receive instant predictions along with confidence levels. This system can support early detection of CKD, improve diagnosis speed, and assist in making better clinical decisions. Overall, this project shows how data science and machine learning can be effectively used in the medical field to save time and potentially save lives.

7. FUTURE SCOPE:

This project can be further improved by incorporating advanced models like deep neural networks or hybrid ensembles to enhance prediction accuracy. Using larger, more diverse datasets would increase generalization across populations. Integration with electronic health records (EHR) could enable real-time clinical use, while a mobile app version would boost accessibility in remote areas. Future enhancements may also include predicting ASD severity or subtypes for personalized treatment and applying explainable AI techniques to make model decisions more transparent and trustworthy for healthcare professionals.

8. ACKNOWLEDGEMENT:

Pinnamraju T S Priya working as a HOD, Assistant Professor in Master of Computer Applications at Sanketika Vidya Parishad Engineering college, Visakhapatnam , Andhra Pradesh with 13 years of experience in Master of Computer Applications (MCA), accredited by NAAC. Her area of intrest is in Computer Programming Using C, Computer Organisation, Software Engneering, Artificial Intelligence, Internet of Things(IoT) and Distributed Systems.

Annepu Vamsi is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine learning Annepu Vamsi has taken up his PG project on Autism Prediction Using Machine Learning and published the paper in connection to the project under the guidance of Pinnamraju T S Priya HOD, Assistant Professor SVPEC.

REFERENCES

[1] Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques
https://link.springer.com/article/10.1007/s42979-021-00776-5
[2] Multiparametric MRI Characterization and Prediction in Autism Spectrum Disorder Using Graph Theory and Machine Learning
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0090405
[3] Detection of autism spectrum disorder (ASD) in children and adults using machine learning
https://www.nature.com/articles/s41598-023-35910-1
[4] Deep learning approach to predict autism spectrum disorder: a systematic review and meta-analysis
https://link.springer.com/article/10.1186/s12888-024-06116
[5] Identification of neural connectivity signatures of autism using machine learning
https://www.frontiersin.org/journals/humanneuroscience/articles/10.3389/fnhum.2013.00670/full
[6] Prediction of Autism Risk From Family Medical History Data Using Machine Learning: A National Cohort Study From Denmark
https://www.sciencedirect.com/science/article/pii/S2667174321000136
[7] prediction of autism spectrum disorder using complex network measures in a machine learning framework
https://www.sciencedirect.com/science/article/pii/S174680942030255X
[8] Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0168224
[9] Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning
https://pmc.ncbi.nlm.nih.gov/articles/PMC5290715/
[10] Identification of newborns at risk for autism using electronic medical records and machine learning
https://www.cambridge.org/core/journals/european-psychiatry/article/identification-of-newborns-at-risk-for-autism-using-electronic-medical-records-and-machine-learning/2602E973811A2B80D524EA06222919D0
[11] Early detection of autism spectrum disorder in young children with machine learning using medical claims data
https://pmc.ncbi.nlm.nih.gov/articles/PMC9462117/
[12] The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age
https://www.sciencedirect.com/science/article/pii/S2665917423001101

[13] A machine learning autism classification based on logistic regression analysis
https://link.springer.com/article/10.1007/s13755-019-0073-5
[14] Identifying and Predicting Autism Spectrum Disorder Based on Multi-Site Structural MRI With Machine Learning
https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2021.765517/full
[15] Predicting the level of autism and improvement rate from assessment dataset using machine learning techniques

https://link.springer.com/article/10.1007/s41870-023-01212-y
[16] The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups
https://www.sciencedirect.com/science/article/pii/S2665917423001101
[17] A machine learning autism classification based on logistic regression analysis
https://link.springer.com/article/10.1007/s13755-019-0073-5
[18] Identifying and Predicting Autism Spectrum Disorder Based on Multi-Site Structural MRI With
https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2021.765517/full
[19] Predicting the level of autism and improvement rate from assessment dataset using machine learning techniques
https://link.springer.com/article/10.1007/s41870-023-01212-y
[20] An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder
https://www.sciencedirect.com/science/article/pii/S2772442523001600
[21] Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises
https://link.springer.com/article/10.1007/s10803-014-2268-6
[22] Use of machine learning to shorten observation-based screening and diagnosis of autism
https://www.nature.com/articles/tp201210
[23] Detecting autism spectrum disorder using machine learning techniques
https://link.springer.com/article/10.1007/s13755-021-00145-9
[24] Genetic factor analysis for an early diagnosis of autism through machine learning
https://www.sciencedirect.com/science/article/abs/pii/B978032398352500001X