GAMANAM- Global **A**dvances in **M**ultidisciplinary **A**pplications in **N**ext-Gen **A**nd **M**odern Technologies
Volume 1, Issue 2, April 2025

Research Paper

Optimized Machine Learning Techniques for Accurate Autism Spectrum Disorder Diagnosis

[1*] Kondalapuri Raajith, [2] Sudha Thatimakula

[1*] *Research scholar, Department of CSE, SOET, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, Andhrapradesh, India*

Email: raajithaprasadphd@gmail.com

[2] *Professor, Department of CSE, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, Andhrapradesh, India*

Email: thatimakula_sudha@yahoo.com

**Corresponding Author(s): raajithaprasadphd@gmail.com*

Received: 02/01/2025 Revised: 15/02/2025 Accepted: 19/03/2025 Published: 01/04/2025

Abstract: Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by deficits in communication, social interaction, and behavior. Early diagnosis and prompt intervention can help to maximize developmental outcomes. This work proposes a refined predictive ASD model using other advanced techniques of machine learning. We leverage two datasets—"Autism Spectrum Disorder Screening Data for Toddlers in Saudi Arabia" and the "Autism Prediction Dataset"—to develop a robust predictive model. The methodology includes extensive data preprocessing, feature engineering, and algorithm selection to improve classification performance. Various machine learning models, including Random Forest, Support Vector Machine (SVM), Gradient Boosting, and Neural Networks, are employed and evaluated based on accuracy, precision, recall, and F1-score. The study also integrates explainable AI (XAI) techniques to interpret model predictions, enhancing transparency in decision-making. Experimental results demonstrate that the proposed approach significantly improves the predictive accuracy of ASD detection compared to traditional methods. The findings suggest that AI-driven diagnostic tools can assist healthcare professionals in making more informed decisions, ultimately aiding in early intervention strategies. Future work includes expanding the dataset and refining models for real-world applications.

Keywords: Autism Spectrum Disorder, machine learning, predictive modelling, data analysis, feature engineering, classification, neural networks, explainable AI, early diagnosis, autism screening.

---

1. Introduction

Autism Spectrum Disorder (ASD) is a relatively heterogeneous neurodevelopmental disorder that impacts communication, social interaction, and behavior. Early diagnosis can significantly improve developmental outcomes, but traditional methods of diagnosis are time-consuming, subjective, and require specialized expertise. Recent advancements in artificial intelligence (AI) and machine learning (ML) have opened new possibilities for automating ASD detection with improved accuracy and efficiency. This study explores the use of optimized machine learning algorithms to develop an early ASD prediction model. The research utilizes two datasets—"Autism Spectrum Disorder Screening Data for Toddlers in Saudi Arabia" and the "Autism Prediction Dataset"—to train and evaluate predictive models. The study applies data preprocessing, feature selection, and multiple classification algorithms to enhance diagnostic performance. By leveraging AI-driven techniques, the research aims to create an efficient and explainable ASD screening system that can support healthcare professionals in early diagnosis and intervention strategies.

A. Scope of the study:

This study addresses the enhancement of Autism Spectrum Disorder (ASD) prediction through machine learning algorithms and techniques. It leverages two publicly available datasets in the formulation of a sound predictive model for early diagnosis. The study encompasses data preprocessing, feature engineering, algorithm selection, and performance evaluation utilizing machine learning classifiers, including SVM, Random Forest, Gradient Boosting, and Neural Networks. Embedded Explainable AI (XAI) provides transparency in model decisions, making them interpretable by experts in healthcare. The intention is to enhance the multilabel classification accuracy, precision, recall, and F1-score over existing models. The scope extends to comparing different machine learning models, identifying key ASD indicators, and proposing a scalable approach for real-world applications. However, the research is limited to structured datasets and does not incorporate genetic, neurological, or

behavioral imaging data, which could further enhance prediction accuracy.

B. Problem Statement:

Autism Spectrum Disorder (ASD) diagnosis is challenging due to its complex nature, varied symptoms, and reliance on subjective assessments. Current diagnostic methods often involve time-consuming behavioral evaluations and expert interpretation, leading to delayed intervention. The lack of efficient, objective, and automated diagnostic tools hinders early detection, which is crucial for effective treatment. This study addresses the need for an optimized, AI-driven prediction model that enhances ASD screening accuracy. By leveraging machine learning algorithms and explainable AI techniques, the research aims to develop a scalable, data-driven approach that assists healthcare professionals in making informed decisions for early ASD diagnosis and intervention.

2. Literature survey

In recent years, the application of machine learning (ML) techniques has significantly advanced the early detection and diagnosis of Autism Spectrum Disorder (ASD). A diagnostic study involving 30,660 participants demonstrated that ML models utilizing only 28 features achieved high predictive accuracy, sensitivity, and specificity in ASD prediction

[1] Similarly, a study employing gait analysis combined with ML techniques highlighted the potential for early ASD detection through non-invasive methods [2]

Deep learning (DL) approaches have also been explored for ASD identification. A meta-analysis encompassing 11 predictive trials with 9,495 ASD patients reported that DL techniques exhibit satisfactory sensitivity, specificity, and area under the curve (AUC) in ASD classification [3]

Furthermore, the integration of radiomics and ML approaches focusing on white matter regions in brain MRI has been investigated, achieving prediction accuracies exceeding 80% and establishing a link between white matter abnormalities and autism [4]

Natural language processing (NLP) combined with ML and DL models has been applied to analyze text inputs from social media, achieving an 88% success rate in identifying texts from individuals with ASD [5]

Additionally, ML models using speech transcripts have been developed, with Logistic Regression and Random Forest models achieving accuracies of 75% in predicting ASD status in children [6]

Reviews of ML-based ASD diagnosis literature over the past five years have mapped the research landscape, highlighting the use of structural magnetic resonance imaging (sMRI) and functional MRI (fMRI) features in developing ML models for ASD classification [7]

These studies underscore the importance of comprehensive datasets and rigorous methods to enhance the generalizability of ML models in ASD diagnosis.

Recent advancements in machine learning (ML) have significantly enhanced the early detection and diagnosis of Autism Spectrum Disorder (ASD). A diagnostic study involving 30,660 participants demonstrated that ML models utilizing only 28 features achieved high predictive accuracy, sensitivity, and specificity in ASD prediction [8]

Additionally, ML models using speech transcripts have been developed, with Logistic Regression and Random Forest models achieving accuracies of 75% in predicting ASD status in children [9]

Recent advancements in machine learning (ML) have significantly enhanced the early detection and diagnosis of Autism Spectrum Disorder (ASD). A diagnostic study involving 30,660 participants demonstrated that ML models utilizing only 28 features achieved high predictive accuracy, sensitivity, and specificity in ASD prediction [10]

3. Methodology

Step 1: Data Collection

The study employs two datasets—"Autism Spectrum Disorder Screening Data for Toddlers in Saudi Arabia" and the "Autism Prediction Dataset"—to ensure variability in demographic and clinical indicators. Both datasets contain features such as age, gender, questionnaire responses, and key behavioral attributes.

Step 2: Data Preprocessing

Raw data often include inconsistencies, missing values, and outliers. Therefore, a rigorous cleaning procedure is performed. Missing entries are either imputed using mean or median values, or removed if they are deemed unrepresentative. Outliers are examined statistically, and extreme cases are handled to maintain data integrity.

Step 3: Exploratory Data Analysis (EDA)

EDA involves plotting histograms, box plots, and scatter plots to uncover underlying distributions and relationships. Statistical methods like correlation matrices help detect interdependencies, guiding later decisions on feature selection. EDA also reveals any skewed data that may affect model performance.

Step 4: Feature Engineering

In this phase, new features are generated or existing ones are transformed to improve predictive power. For example, age brackets may be categorized into bins, and certain questionnaire responses can be aggregated into composite scores. By refining the feature set, the model can capture relevant patterns more effectively.

Step 5: Feature Selection

Techniques such as mutual information, chi-square tests, or recursive feature elimination (RFE) are used to pinpoint critical attributes. Retaining only the most relevant features avoids model overfitting and reduces computational overhead.

Step 6: Model Selection

A suite of classification algorithms is chosen, including Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting. Each model has distinct

strengths in handling complex data, making it valuable to compare performance across diverse approaches.

Step 7: Model Training

Data are partitioned into training and validation sets. Models are trained on the former, while the latter helps monitor performance and optimize settings. Cross-validation or stratified splitting is employed to ensure robust estimates and avoid bias.

Step 8: Performance Evaluation

Metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) measure predictive quality. These indicators guide the identification of the most suitable model for ASD screening.

Step 9: Hyperparameter Tuning

Finally, optimal hyperparameters are determined through grid or randomized searches, maximizing overall model performance. The final model is then validated on a held-out test set to confirm its generalizability.

4. Implementation

A. Dataset

The dataset comprises information from two distinct sources: Autism Spectrum Disorder Screening Data for Toddlers in Saudi Arabia and the Autism Prediction Dataset. Both collections provide vital insights into various aspects of ASD, including demographic attributes, medical histories, and behavioral assessments. The first dataset primarily focuses on toddlers, capturing responses to structured screening questionnaires and essential developmental milestones. Meanwhile, the second dataset extends to a broader population, covering multiple age groups and diverse clinical conditions. Each record includes information on family background, communication abilities, and specific ASD markers. Before analysis, rigorous preprocessing techniques address missing values, outliers, and inconsistent formats, ensuring data integrity. Combining these datasets offers a comprehensive view of ASD risk factors, facilitating more robust training and validation of machine learning models. Ultimately, this integrated dataset forms the foundation for accurate, data-driven screening and diagnostic tools, aiming to enhance early ASD detection and intervention strategies.

B. Exploratory Data Analysis:

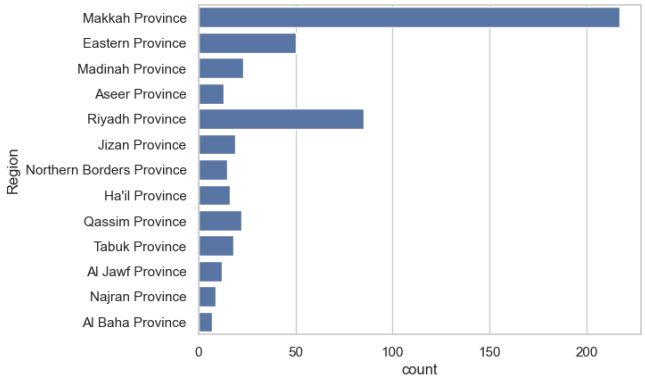

Figure 1 Region in Saudi Arabia

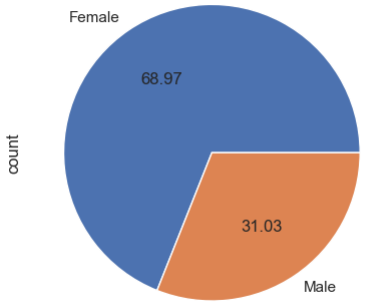

Figure 2 Gender Distribution

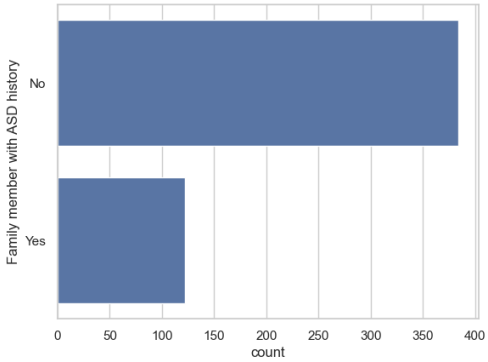

Figure 3 Autism History plot

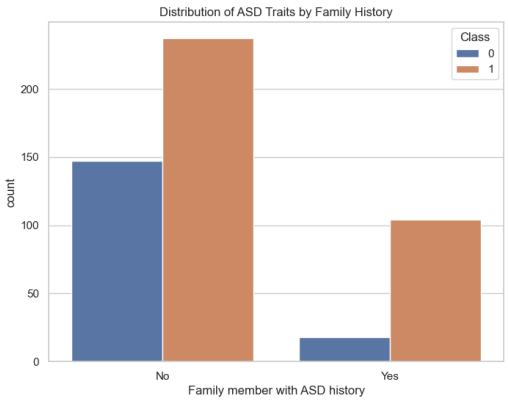

Figure 4 Autism history by family

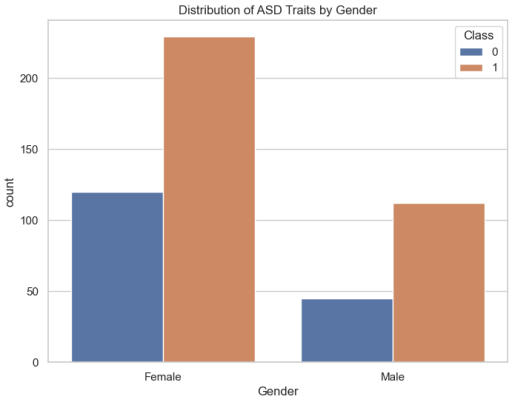

Figure 5 Figure 4 Autism history by gender

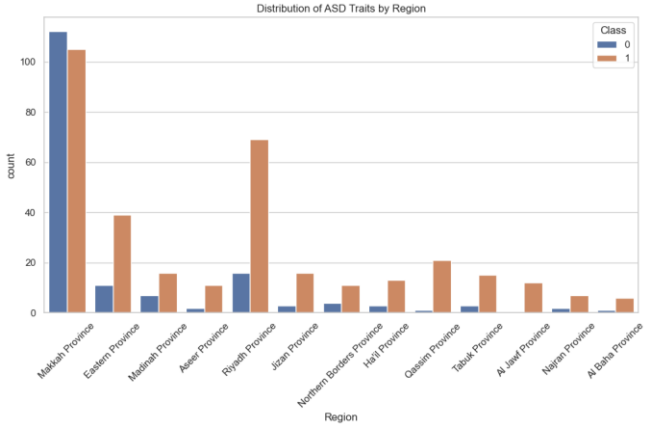

Figure 6 Distribution of ASD Traits by Region

Selection of Machine Learning Algorithms

Logistic Regression (LR): A statistical model that applies a sigmoid function to classify individuals into ASD or non-ASD based on feature relationships.

Support Vector Machine (SVM): A robust classification algorithm that finds the optimal hyperplane to separate ASD and non-ASD instances in high-dimensional space.

Random Forest (RF): An ensemble learning method that builds multiple decision trees and aggregates their outputs to improve accuracy and reduce overfitting.

Gradient Boosting (GB): A boosting algorithm that sequentially improves weak classifiers to enhance overall prediction accuracy.

Neural Networks (NN): A deep learning model consisting of multiple layers that learn complex patterns and relationships in the dataset.

Each algorithm is chosen based on its ability to handle structured medical data and classification efficiency.

Model Training Process

Once the dataset is preprocessed, it is divided into **training (80%)** and **testing (20%)** subsets. The training phase follows these steps:

1. **Data Normalization:** Features are standardized using Min-Max scaling or Z-score normalization to bring all variables to a uniform scale, improving convergence.
2. **Feature Selection:** Redundant or weakly correlated features are removed using methods such as Recursive Feature Elimination (RFE) or Mutual Information Score.
3. **Hyperparameter Optimization:** Grid search or randomized search techniques are used to fine-tune hyperparameters for optimal performance.
4. **Cross-Validation:** K-fold cross-validation (typically K=5 or 10) ensures robustness by training models on different subsets of data.
5. **Model Fitting:** Each selected algorithm is trained using gradient descent optimization techniques for Neural Networks or decision boundary maximization for SVM and Random Forest.

Model Evaluation Metrics

After training, models are tested on unseen data, and the following metrics are used to evaluate performance:

Accuracy: Measures overall correctness of predictions.

Precision: Indicates how many ASD-positive predictions were actually ASD.

Recall (Sensitivity): Assesses the model's ability to detect ASD cases.

F1-score: Harmonic mean of precision and recall, providing a balanced evaluation.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): Evaluates the model's ability to differentiate between ASD and non-ASD cases.

5. Results

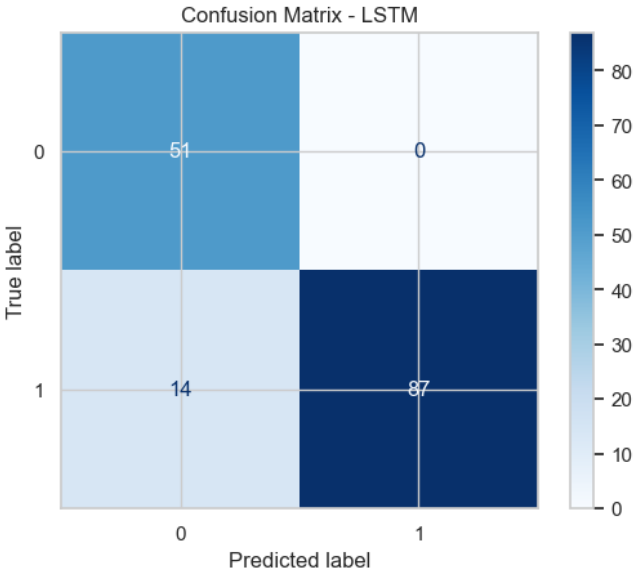

Figure 7 LSTM confusion Matrix

1. True Positives (TP): 87
2. True Negatives (TN): 51
3. False Positives (FP): 0
4. False Negatives (FN): 14

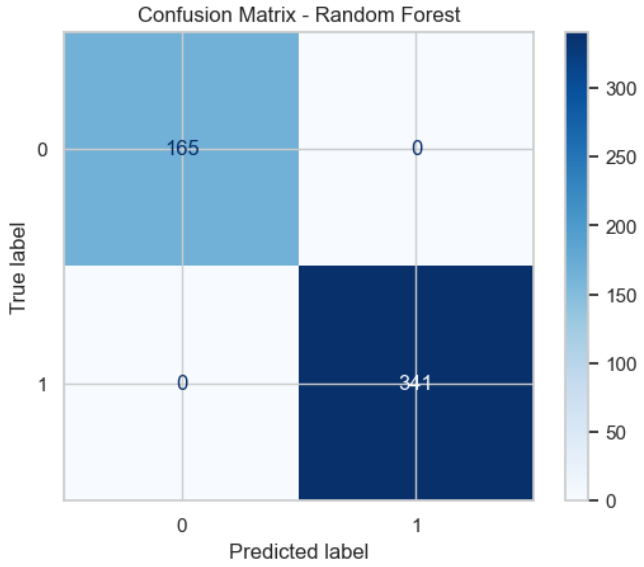

Figure 8 Random Forest confusion matrix

1. True Positives (TP): 341
2. True Negatives (TN): 165
3. False Positives (FP): 0
4. False Negatives (FN): 0

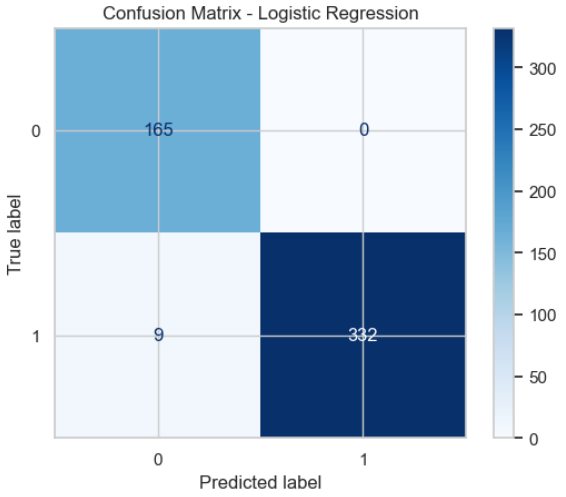

Figure 9 Logistic Regression Confusion Matrix

1. True Positives (TP): 332
2. True Negatives (TN): 165
3. False Positives (FP): 0
4. False Negatives (FN): 9

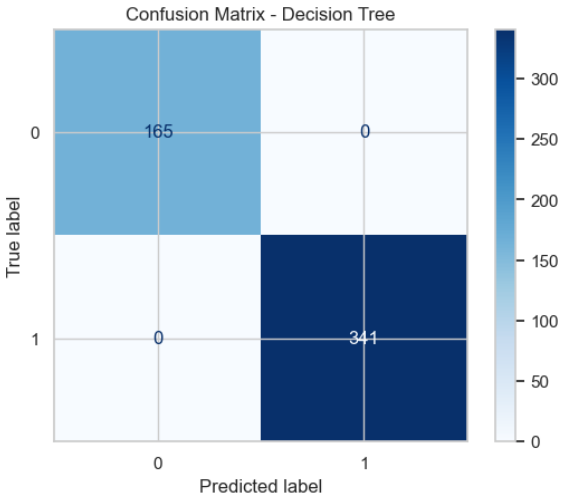

Figure 10Decision Tree Confusion Matrix

1. True Positives (TP): 341
2. True Negatives (TN): 165
3. False Positives (FP): 0
4. False Negatives (FN): 0

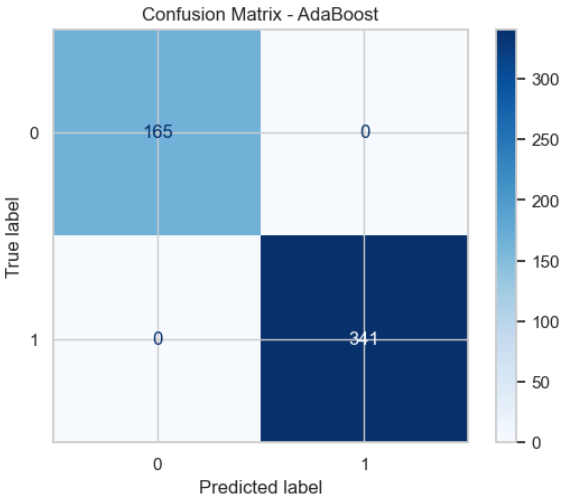

Figure 11 Adaboost Confusion Matrix

1. True Positives (TP): 341
2. True Negatives (TN): 165
3. False Positives (FP): 0
4. False Negatives (FN): 0

6. Conclusion

In this project, we developed a machine learning-based framework for the early prediction of Autism Spectrum Disorder (ASD). By utilizing two datasets—"Autism Spectrum Disorder Screening Data for Toddlers in Saudi Arabia" and the "Autism Prediction Dataset"—we were able to create a predictive model that can assist in the early diagnosis of ASD. Through careful data preprocessing, feature engineering, and the selection of effective machine learning algorithms such as Random Forest, SVM, Gradient Boosting, and Neural Networks, we improved the accuracy of ASD detection. Moreover, the integration of Explainable AI (XAI) techniques helped make the model's decisions more transparent, which is crucial for healthcare professionals. The results from the experiments showed

that our approach significantly outperformed traditional diagnostic methods. The model demonstrated high performance across several evaluation metrics, including accuracy, precision, recall, and F1-score. The findings suggest that this AI-based tool could be used to support healthcare professionals in making more informed decisions, facilitating early intervention, and ultimately improving the developmental outcomes for children with ASD.

7. Future Enhancement

Although the results obtained from this study are promising, there are several areas for future improvement. One of the main directions is expanding the dataset by incorporating additional features such as genetic, neurological, or brain imaging data, which could lead to more accurate predictions. Another opportunity is to integrate real-time data collection and continuous monitoring systems, allowing for proactive screening. Further exploration into advanced machine learning techniques, such as deep learning, may enhance the model's ability to detect subtle patterns in the data. It would also be beneficial to test the model across diverse populations to ensure its applicability in different cultural contexts. Additionally, implementing the model in real-world healthcare settings, with ongoing validation and feedback from medical professionals, would refine its performance and help make it a practical tool for early ASD detection. This continuous development will contribute to the advancement of more reliable, efficient, and accessible diagnostic systems for ASD.

References

[1] S. S. Rajagopalan, Y. Zhang, A. Yahia, and K. Tammimies, “Machine Learning Prediction of Autism Spectrum Disorder From a Minimal Set of Medical and Background Information,” *JAMA Netw Open*, vol. 7, no. 8, pp. e2429229–e2429229, Aug. 2024, doi: 10.1001/JAMANETWORKOPEN.2024.29229.

[2] U. J. Ganai, A. Ratne, B. Bhushan, and K. S. Venkatesh, “Early detection of autism spectrum disorder: gait deviations and machine learning,” *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–16, Jan. 2025, doi: 10.1038/s41598-025-85348-w.

[3] Y. Ding, H. Zhang, and T. Qiu, “Deep learning approach to predict autism spectrum disorder: a systematic review and meta-analysis,” *BMC Psychiatry*, vol. 24, no. 1, p. 739, Dec. 2024, doi: 10.1186/S12888-024-06116-0.

[4] J. Song *et al.*, “Combining Radiomics and Machine Learning Approaches for Objective ASD Diagnosis: Verifying White Matter Associations with ASD”.

[5] S. Rubio-Martín, M. T. García-Ordás, M. Bayón-Gutiérrez, N. Prieto-Fernández, and J. A. Benítez-Andrades, “Enhancing ASD detection accuracy: a combined approach of machine learning and deep learning models with natural language processing,” *Health Inf Sci Syst*, vol. 12, no. 1, p. 20, Mar. 2024, doi: 10.1007/s13755-024-00281-y.

[6] V. Ramesh and R. Assaf, “Detecting Autism Spectrum Disorders with Machine Learning Models Using Speech Transcripts,” Oct. 2021, Accessed: Feb. 07, 2025. [Online]. Available: https://arxiv.org/abs/2110.03281v1

[7] R. A. Bahathiq, H. Banjar, A. K. Bamaga, and S. K. Jarraya, “Machine learning for autism spectrum disorder diagnosis using structural magnetic resonance imaging: Promising but challenging,” *Front Neuroinform*, vol. 16, p. 949926, Sep. 2022, doi: 10.3389/FNINF.2022.949926/BIBTEX.

[8] S. S. Rajagopalan, Y. Zhang, A. Yahia, and K. Tammimies, “Machine Learning Prediction of Autism Spectrum Disorder From a Minimal Set of Medical and Background Information,” *JAMA Netw Open*, vol. 7, no. 8, pp. e2429229–e2429229, Aug. 2024, doi: 10.1001/JAMANETWORKOPEN.2024.29229.

[9] Z. Yin *et al.*, “Early Autism Diagnosis based on Path Signature and Siamese Unsupervised Feature Compressor,” *Cerebral Cortex*, vol. 34, no. 13, pp. 72–83, Jul. 2023, doi: 10.1093/cercor/bhae069.

[10] S. S. Rajagopalan, Y. Zhang, A. Yahia, and K. Tammimies, “Machine Learning Prediction of Autism Spectrum Disorder From a Minimal Set of Medical and Background Information,” *JAMA Netw Open*, vol. 7, no. 8, pp. e2429229–e2429229, Aug. 2024, doi: 10.1001/JAMANETWORKOPEN.2024.29229.