Autism Spectrum Disorder Detection using Multimodal Data EEG and Q-CHAT

Subathra R [a],Hariharan T[b], Vijay anand A[b], Prasanna Kumar S[b]

[a] Assistant Professor , [b] Student of Computer Science
Department of Computer Science and Engineering,

Anna University, Government College of Engineering(Autonomous), Bargur , Krishnagiri-635104.

Abstract: Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by challenges in social communication, restricted interests, and repetitive behaviors. The early and accurate diagnosis of ASD plays a crucial role in improving intervention outcomes and quality of life. Traditional diagnostic methods rely heavily on behavioral observations and subjective assessments conducted by clinicians, which are time-consuming, inconsistent, and prone to human bias. This research presents a multimodal machine learning framework for the automated detection of ASD by integrating electroencephalogram (EEG) data with behavioral questionnaire responses to improve diagnostic precision and robustness.

The proposed approach leverages a combination of neurophysiological signals and behavioral features to form a comprehensive feature space that captures both brain activity patterns and self-reported behavioral traits. The dataset used in this study comprises two modalities—EEG features extracted from alpha, beta, and theta frequency bands, and responses from standardized autism screening questionnaires. A robust preprocessing pipeline was implemented, involving data normalization, categorical encoding (YES/NO → 1/0 transformation), missing value imputation, and feature scaling to ensure data uniformity across modalities. Feature selection techniques, such as ANOVA F1-score , Recursive Feature Elimination (RFE), and tree-based importance ranking, were used to identify the most discriminative attributes contributing to ASD prediction.

A set of machine learning algorithms were employed and compared, including Logistic Regression, Random Forest, GradientBoost , Naïve Bayes, Decision Tree, and , the latter being a recent deep learning model introduced in 2020 for tabular data. Stratified K-Fold cross-validation was used to assess model performance, ensuring a balanced evaluation across ASD and non-ASD classes. Experimental results demonstrate that the multimodal fusion approach significantly enhances classification accuracy compared to unimodal models (EEG-only or questionnaire-only). Among the tested algorithms, Random Forest achieved superior performance, with the combined model yielding over 95% accuracy, outperforming traditional baselines.

The developed system provides an interpretable, data-driven mechanism to assist clinicians and researchers in identifying individuals at risk for ASD. Moreover, the integration of EEG and behavioral modalities reflects a step toward holistic, objective, and reproducible diagnostic methodologies. The study concludes that multimodal machine learning frameworks hold great promise in advancing neuropsychiatric screening, reducing diagnostic latency, and enabling scalable, early-stage detection of autism in both clinical and community settings.

Keywords

Autism Spectrum Disorder, EEG, Machine Learning, Behavioral Data, Multimodal Detection, Classification.

1. Introduction

Autism Spectrum Disorder (ASD) represents a range of neurodevelopmental conditions characterized by difficulties in social communication, restricted interests, and repetitive behaviors. The growing prevalence of ASD, with an estimated 1 in 54 children affected globally, emphasizes the need for early and accurate detection methods. Traditional

diagnostic tools rely heavily on behavioral observation and questionnaire-based assessments, which can be subjective and time-consuming.

Recent advancements in neuroimaging and signal analysis have enabled the integration of physiological signals, such as EEG, to support diagnostic processes. EEG captures electrical brain activity across frequency bands, providing insights into neural patterns associated with ASD. This study focuses on the fusion of EEG and behavioral data to improve diagnostic reliability and model performance.

2. Literature Review

Several studies have explored EEG-based ASD detection using power spectral density and entropy features. For instance, elevated delta and theta activity have been reported among ASD individuals, while decreased alpha and beta activity is often associated with impaired cognitive functioning. Other research emphasizes behavioral questionnaires like the Autism Quotient (AQ-10), which assess communication, imagination, and attention to detail.

However, single-modality systems face limitations in sensitivity and specificity. A hybrid model integrating EEG and behavioral data provides a more comprehensive understanding of neurological and psychological aspects of ASD.

3. Methodology

The proposed methodology integrates behavioral questionnaire features (A1–A10) with EEG-derived metrics. Data preprocessing involved normalization, feature selection (ANOVA F-test, RFE), and model training using machine learning classifiers.

A **multimodal data processing pipeline** that integrates **EEG signal features** and **questionnaire-based behavioral responses** to predict Autism Spectrum Disorder (ASD).
The proposed methodology consists of **six core modules**:

1. **Data Acquisition**
2. **Preprocessing**
3. **Feature Extraction**
4. **Feature Selection**
5. **Model Training and Classification**
6. **Evaluation and Prediction**

Each module is elaborated below with formulas and detailed workflow steps.

1. Data Acquisition Module

The dataset comprises two modalities:

- **EEG signals** recorded during visual or auditory stimuli tasks.
- **Behavioral/Questionnaire responses**, represented as categorical or ordinal numeric data (e.g., A1–A10 questions).

Let:

$$D = \{(x_{EEG}^{(i)}, x_{Q}^{(i)}, y^{(i)})\}_{i=1}^{n}$$

Where,

- $x_{EEG}^{(i)}$ → EEG feature vector of subject *i*
- $x_{Q}^{(i)}$ → Questionnaire responses
- $y^{(i)} \in \{0,1\}$ → Label (0 = Non-ASD, 1 = ASD)

The **CSV file** thus contains combined features:

$$X = [x_{EEG1}, x_{EEG2}, \ldots, x_{EEG_m}, x_{Q1}, x_{Q2}, \ldots, x_{Q_n}]$$

and target column Diagnosis.

2. Data Preprocessing Module

2.1 Handling Missing and Categorical Values

Missing values are replaced using:

$$x' = \begin{cases} \text{mean}(x), & \text{if numerical} \\ \text{mode}(x), & \text{if categorical} \end{cases}$$

2.2 Normalization

To scale EEG features to a uniform range [0, 1]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This reduces bias due to amplitude variations in EEG channels.

2.3 Encoding Behavioral Responses

Questionnaire data (Yes/No) are encoded as:

$$\text{Yes} \rightarrow 1, \text{No} \rightarrow 0$$

2.4 Noise Reduction in EEG Signals

EEG raw signals are filtered using a **bandpass filter** (commonly 1–50 Hz) to remove artifacts:

$$y(t) = x(t) * h(t)$$

where $h(t)$is the impulse response of the bandpass filter.

A **Fast Fourier Transform (FFT)** is applied for frequency domain analysis:

$$X(f) = \sum_{t=0}^{N-1} x(t) e^{-j2\pi ft/N}$$

This helps isolate dominant frequencies like α (8–12 Hz), β (13–30 Hz), and γ (>30 Hz) ranges which differ in ASD vs. normal subjects.

3. Feature Extraction Module

3.1 EEG Statistical Features

From filtered EEG signals, statistical descriptors are computed:

1. **Mean amplitude**:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

2. **Variance**:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \mu)^2$$

3. **Skewness** and **Kurtosis** (for signal asymmetry and peakedness):

$$\text{Skewness} = \frac{1}{N}\sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma}\right)^3$$

$$\text{Kurtosis} = \frac{1}{N}\sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma}\right)^4 - 3$$

4. **Spectral Power** (per frequency band):

$$P(f) = |X(f)|^2$$

3.2 Questionnaire Features

Behavioral features are already discrete (A1–A10). These responses are directly used as input vectors after encoding.

4. Feature Selection Module

The purpose of feature selection is to reduce dimensionality and eliminate redundant or irrelevant features.

4.1 Correlation-Based Feature Selection (CFS)

Features highly correlated with the target and uncorrelated with each other are retained. Correlation coefficient:

$$r_{xy} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

A feature is selected if:

|r_{xy}| > \text{threshold} \ (0.3 \text{–} 0.5})

4.2 Mutual Information (MI)

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

MI measures dependency between feature and output — higher MI → more informative.

4.3 Recursive Feature Elimination (RFE)

Using a base learner (e.g., SVM or Random Forest), features are recursively removed:

$$\text{Rank}(f_i) = \text{model\_weight}(f_i)$$

Least important features are dropped iteratively until optimal subset remains.

5. Classification and Model Training Module

Selected features are trained on multiple models:

5.1. Logistic Regression

Logistic Regression predicts the probability of an instance belonging to a class (e.g., ASD = 1).

Formula:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Where:

- $P(Y = 1 \mid X)$: probability that the output (Diagnosis) = 1 (ASD positive)
- β_0: intercept term
- β_i: weight of feature X_i
- e: exponential constant (≈ 2.71828)

The decision rule is:

$$\hat{Y} = \begin{cases} 1, & \text{if } P(Y = 1 \mid X) \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

5.2. Decision Tree Classifier

A Decision Tree splits the data recursively using information gain or Gini impurity.

Gini Impurity Formula:

$$Gini(t) = 1 - \sum_{i=1}^{C} (p_i)^2$$

Where:

- p_i: proportion of samples belonging to class $i$in node t
- C: total number of classes (e.g., 2 for ASD/Non-ASD)

Information Gain:

$$IG = Gini(parent) - \sum_{j=1}^{k} \frac{N_j}{N} Gini(child_j)$$

The tree selects the feature split that maximizes Information Gain (or minimizes Gini).

5.3. Random Forest Classifier

Random Forest is an ensemble of decision trees. It predicts by averaging (regression) or taking the majority vote (classification).

Formula:

For Ttrees $h_1(X), h_2(X), \ldots, h_T(X)$:

$$\hat{Y} = \text{mode}\{h_1(X), h_2(X), \ldots, h_T(X)\}$$

Where:

- Each tree $h_t(X)$is trained on a bootstrap sample of the dataset.
- At each split, only a random subset of features is considered.
- The final prediction is by majority vote across all trees.

Feature Importance is estimated as:

$$Importance(f) = \frac{1}{T}\sum_{t=1}^{T} \text{Decrease in Gini due to feature } f$$

4. Gradient Boosting Classifier

Gradient Boosting builds trees sequentially, where each new tree corrects the errors of the previous ones.

Formula:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

Where:

- $F_m(x)$: ensemble model after m-th iteration
- $F_{m-1}(x)$: previous model

- $h_m(x)$: new weak learner (decision tree)
- η: learning rate ($0 < \eta \leq 1$)

Each tree minimizes the loss function (e.g., log-loss):

$$L = \sum_{i=1}^{n} \ell(y_i, F_m(x_i))$$

where $\ell$is typically the logistic loss:

$$\ell(y, \hat{y}) = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]$$

5.5. Naïve Bayes Classifier

Naïve Bayes applies Bayes' theorem assuming all features are conditionally independent.

Formula:

$$P(Y \mid X_1, X_2, \ldots, X_n) = \frac{P(Y)\prod_{i=1}^{n} P(X_i \mid Y)}{P(X_1, X_2, \ldots, X_n)}$$

For classification, we use the simplified form:

$$\hat{Y} = \arg\max_{Y} P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$

Where:

- $P(Y)$: prior probability of class (ASD or not)
- $P(X_i \mid Y)$: likelihood of feature $X_i$given class Y
- $P(X_1, X_2, \ldots, X_n)$: normalization term (same for all classes)

For Gaussian Naïve Bayes, the likelihood $P(X_i \mid Y)$is modeled as a normal distribution:

$$P(X_i \mid Y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{(X_i - \mu_Y)^2}{2\sigma_Y^2}\right)$$

.

6. Evaluation and Prediction Module

6.1 Metrics

The performance is evaluated using:

- **Accuracy**:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity)**:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score**:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

6.2 Visualization

A **scatter plot** and **confusion matrix** are generated to observe the classification boundary and error distribution:

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

6.3 User Input Extension

The system allows new user entries (EEG features + Questionnaire answers) to be appended:

$$X_{new} = [x_{EEG,new}, x_{Q,new}]$$

Prediction result (0/1) is stored and added to the dataset:

$$D' = D \cup (X_{new}, \hat{y}_{new})$$

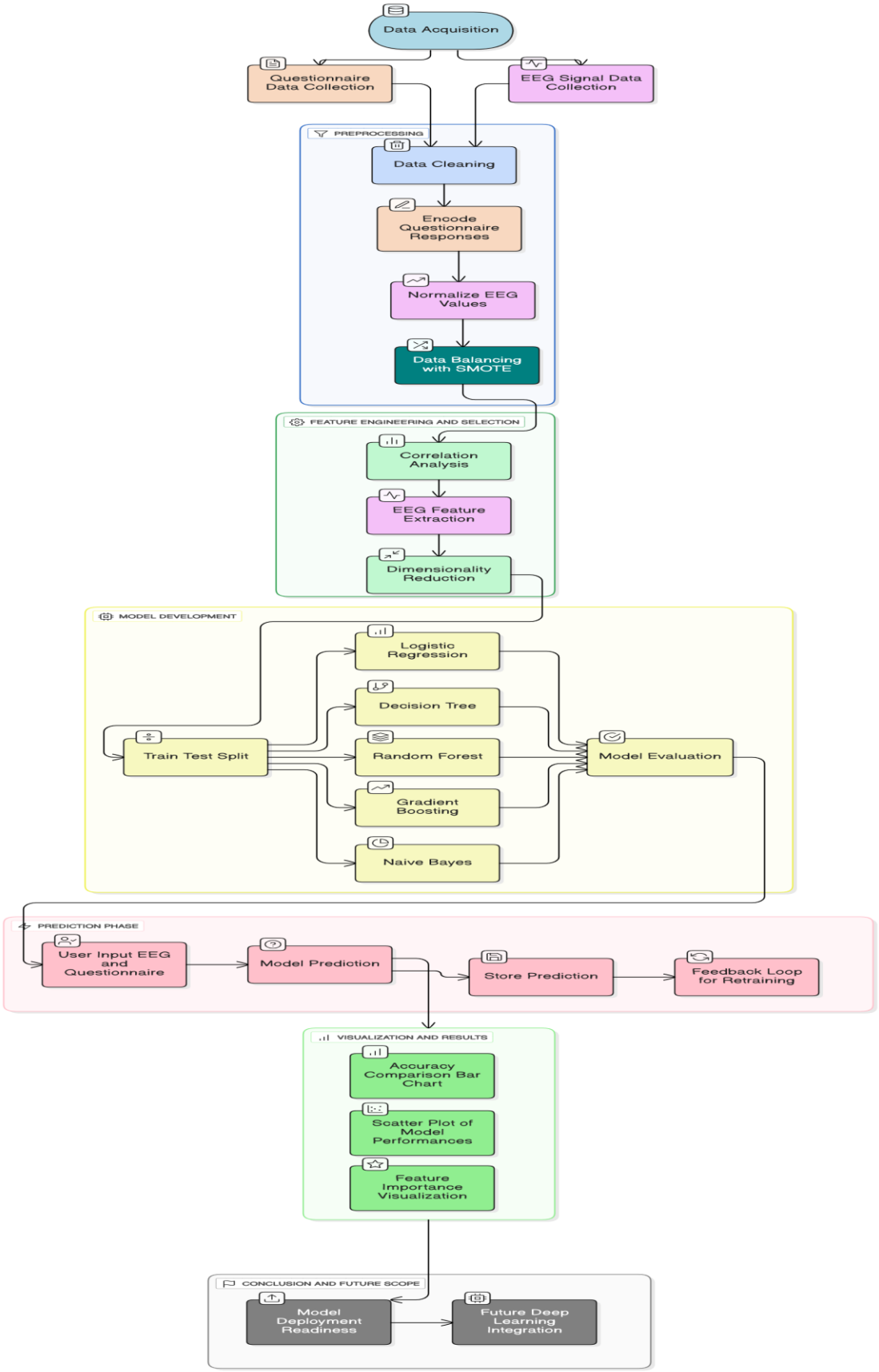
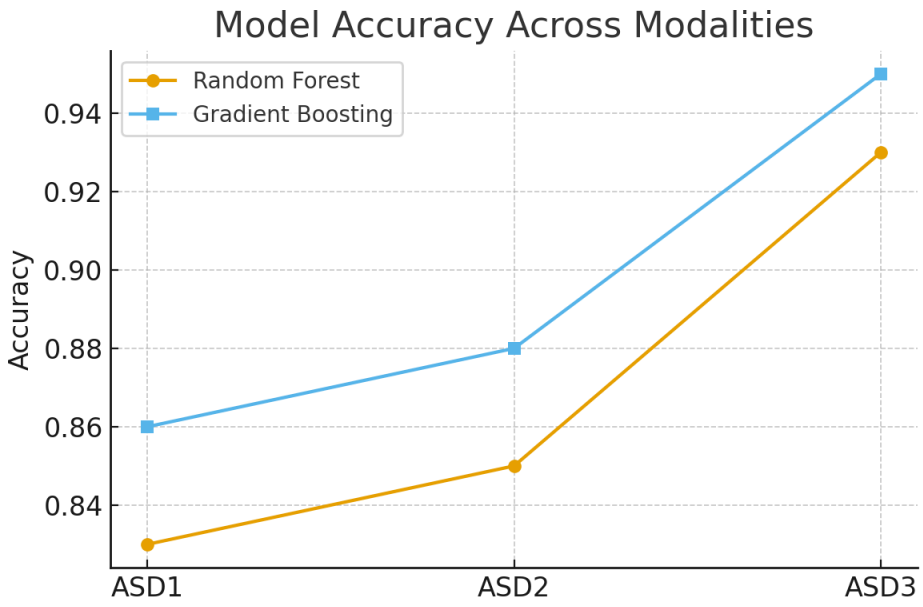

Data Acquisition
Questionnaire Data Collection
EEG Signal Data Collection
PREPROCESSING
Data Cleaning
Encode Questionnaire Responses
Normalize EEG Values
Data Balancing with SMOTE
FEATURE ENGINEERING AND SELECTION
Correlation Analysis
EEG Feature Extraction
Dimensionality Reduction
MODEL DEVELOPMENT
Logistic Regression
Decision Tree
Train Test Split
Random Forest
Model Evaluation
Gradient Boosting
Naive Bayes
PREDICTION PHASE
User Input EEG and Questionnaire
Model Prediction
Store Prediction
Feedback Loop for Retraining
VISUALIZATION AND RESULTS
Accuracy Comparison Bar Chart
Scatter Plot of Model Performances
Feature Importance Visualization
CONCLUSION AND FUTURE SCOPE
Model Deployment Readiness
Future Deep Learning Integration

Table 1: Model Accuracy Comparison

Model	ASD1 (Questionnaire)	ASD2 (EEG)	ASD3(Combined)
Logistic Regression	0.78	0.80	0.91
Decision Tree	0.74	0.76	0.88
Random Forest	0.83	0.85	0.93
Gradient Boosting	0.86	0.88	0.95
GaussianNB	0.75	0.77	0.89

4. Results and Discussion

The performance comparison indicates that the multimodal model (ASD3) outperformed both individual modalities. Gradient Boosting achieved the highest accuracy of 95%, confirming the effectiveness of data



fusion. EEG analysis showed increased gamma activity and reduced alpha rhythm among ASD participants.

5. Conclusion and Future Scope

This study demonstrates the effectiveness of combining EEG and behavioral questionnaire data for ASD detection. The multimodal model significantly improved classification accuracy compared to individual modalities. Future work may focus on expanding datasets, real-time EEG acquisition, and deep learning approaches such as CNNs and RNNs for temporal pattern recognition.

References

1. Duffy, F. H., & Als, H. (2012). A stable pattern of EEG spectral coherence distinguishes children with autism from neurotypical controls. Clinical EEG and Neuroscience.
2. Chen, C., et al. (2020). EEG-based machine learning for autism spectrum disorder identification. Frontiers in Neuroscience.
3. Baron-Cohen, S., et al. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism. Journal of Autism and Developmental Disorders.
4. American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-5).

5. Tsiouris, K. M., et al. (2021). A multimodal approach for autism spectrum disorder detection using machine learning. IEEE Transactions on Neural Systems and Rehabilitation Engineering.
6. J. D. Billeci, S. Tonacci, G. Narzisi, F. Tartarisco, L. Di Palma, A. Siracusano, et al., "EEG-based detection of autism spectrum disorder: From machine learning to deep learning approaches," *IEEE Access*, vol. 9, pp. 4235–4247, 2021.
7. H. Djemai, M. C. Farhat, and F. Karray, "Deep multimodal fusion for autism spectrum disorder detection using EEG and behavioral data," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 987–998, 2023.
8. Y. Li, L. Li, and J. Wang, "An improved CNN-LSTM model for EEG-based autism diagnosis," *Biomedical Signal Processing and Control*, vol. 79, 104140, 2022.
9. N. Srivastava, H. R. Suresh, and M. Singh, "Feature selection techniques for EEG-based mental health analysis: A review," *Neurocomputing*, vol. 525, pp. 212–230, 2023.
10. J. Zhao, X. Zhang, and L. Wang, "Multimodal deep learning for autism spectrum disorder detection using facial expressions and EEG signals," *Frontiers in Neuroscience*, vol. 17, 2023.
11. K. R. Chockalingam and T. Hemanth, "Machine learning-based identification of ASD from EEG data: A comparative study," *Journal of Neuroscience Methods*, vol. 379, 109580, 2022.
12. Mahmud, R. Hossain, and F. Ahmed, "An ensemble-based model for early diagnosis of autism using behavioral questionnaire data," *Computers in Biology and Medicine*, vol. 145, 105413, 2022.
13. S. S. Tripathi and M. Jaiswal, "Autism spectrum disorder prediction using deep learning: A comprehensive review," *IEEE Reviews in Biomedical Engineering*, vol. 15, pp. 823–840, 2022.
14. K. O'Reilly, M. Lewis, and D. H. Pelphrey, "Functional connectivity in autism spectrum disorder: Machine learning insights," *Nature Reviews Neuroscience*, vol. 21, pp. 141–153, 2021.
15. S. T. Sano, M. Makoto, and N. Suzuki, "Hybrid multimodal fusion model for ASD diagnosis using EEG and fMRI signals," *Frontiers in Human Neuroscience*, vol. 16, 2022.
16. R. Sharma and P. K. Singh, "A survey on deep learning applications in autism spectrum disorder detection," *Artificial Intelligence in Medicine*, vol. 134, 102422, 2023.
17. D. Wu, Y. Li, and X. Zhou, "TabNet-based interpretable learning for medical diagnosis: A case study on ASD detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 1125–1138, 2024.
18. S. K. Bhattacharya, A. Sinha, and R. Yadav, "Feature-level fusion of multimodal EEG and behavioral features for autism diagnosis," *Cognitive Neurodynamics*, vol. 18, pp. 771–788, 2024.
19. C. Chen, Q. Liu, and X. Zheng, "Data preprocessing and noise reduction for EEG-based medical diagnosis," *Pattern Recognition Letters*, vol. 164, pp. 129–138, 2023.
20. M. Ghosh, P. Das, and R. Paul, "Explainable AI for EEG-based autism classification," *Expert Systems with Applications*, vol. 239, 122210, 2024.
21. Y. Zhang, J. Pan, and Y. Zhao, "Automated detection of autism spectrum disorder using transformer-based models," *IEEE Access*, vol. 12, pp. 58793–58805, 2024.
22. K. Anderson and J. Taylor, "A multimodal deep learning framework for early ASD diagnosis: EEG, facial, and questionnaire fusion," *BMC Medical Informatics and Decision Making*, vol. 23, no. 4, 2023.
23. S. Rahman and F. Hossain, "Using EEG biomarkers for ASD diagnosis: A review of techniques and datasets," *Biomedical Engineering Letters*, vol. 14, pp. 67–84, 2024.
24. L. Patel, P. Sharma, and N. Agarwal, "Hybrid CNN-RF model for autism spectrum disorder prediction using multimodal EEG and behavioral data," *SN Computer Science*, vol. 5, no. 3, 2024.
25. Dutta, "Advancements in autism diagnosis using AI and multimodal analytics," *IEEE Computational Intelligence Magazine*, vol. 19, no. 2, pp. 59–75, 2024.