Integrating Diverse EEG Datasets: A Novel Approach for Robust and Generalizable ASD Detection

Anshuman Sharma
anshuman.sharma@plaksha.edu.in

Dhruv Srivastava
dhruv.srivastava@plaksha.edu.in

Jiya Agrawal
jiya.agrawal@plaksha.edu.in

Yash Sangtani
yash.sangtani@plaksha.edu.in

***Abstract*—This project aims to identify the optimal combination of preprocessing techniques, feature sets, and model architectures for diagnosing Autism Spectrum Disorder (ASD) using EEG data. By comparing existing models on a combined dataset and performing a grid search, we aim to develop a robust pipeline that works effectively with both single-channel and multi-channel EEG data. The resulting model seeks to assist healthcare professionals in detecting early-onset ASD in toddlers, overcoming the limitations of traditional behavioral assessments.**

I. Introduction

The increasing prevalence of Autism Spectrum Disorder (ASD) necessitates innovative diagnostic approaches that go beyond subjective behavioral assessments. Electroencephalography (EEG), with its non-invasive nature and ability to capture neural dynamics, has emerged as a promising tool for ASD detection. This study aims to leverage EEG data for early ASD diagnosis by integrating datasets, extracting meaningful features, and employing advanced machine learning techniques.

A significant challenge addressed in this research is the harmonization of EEG datasets with varying experimental setups and formats. By systematically preprocessing the data and utilizing robust feature extraction techniques, this study ensures compatibility and preserves critical signal characteristics. The analysis explores single-channel and multi-channel EEG data, highlighting the trade-offs and advantages associated with each approach.

Through the evaluation of diverse machine learning models, including ensemble methods, the study identifies optimal pipelines for EEG-based ASD detection. This work not only advances diagnostic accuracy but also emphasizes the importance of dataset size, feature selection, and model consistency in developing reliable and generalizable diagnostic tools.

II. Literature Review

A. *EEG Analytics for Early Detection of Autism Spectrum Disorder by Bosl et al. (2018) [1]*

The study explores the use of nonlinear analysis of EEG signals combined with pattern classification techniques to predict whether an infant will develop Autism Spectrum Disorder (ASD) as early as three months. EEG data were collected from 19 sensors, arranged according to the standard 10–20 montage, while infants were seated on their mothers' laps in a controlled, soundproof, and electrically shielded environment. The signals were decomposed into six frequency bands using a wavelet transform, and features were extracted from Recurrence Quantitative Analysis (RQA), Sample Entropy (SampE), and Detrended Fluctuation Analysis (DFA). A leave-one-out cross-validation procedure was employed to predict whether an infant had ASD.

The results indicated that classification methods could effectively distinguish ASD infants from those in low-risk families, with significant differences in nonlinear measures between the two groups. These findings suggest that EEG signals can be used for early detection of ASD.

For our implementation, we adopted the Daubechies (DB4) wavelet decomposition for preprocessing EEG data and utilized nine nonlinear features as described in the study.

B. *Exploring Machine Learning Approaches for Autism Diagnosis through EEG Signal by Emara (2024) [2]*

The objective of the paper is to investigate the use of machine learning techniques for analyzing EEG signals to classify Autism Spectrum Disorder (ASD), with a focus on achieving high accuracy while maintaining computational efficiency. The goal is to develop an accessible and cost-effective diagnostic tool for ASD based on EEG data.

The dataset used in the study was publicly available from the University of Sheffield, linked to the publication "Electrophysiological Signatures of Brain Aging in Autism Spectrum Disorder." It included data from 28 individuals with ASD and 28 neurotypical controls. EEG signals were recorded at a sampling rate of 512 Hz from multiple channels (PO4, Oz, Fp1, P8, O2, C3, F8, P1, Fp2) during a 2.5-minute eyes-closed resting state.

Preprocessing involved artifact removal using the Automatic and Tunable Artifact Removal Algorithm (ATAR) with Wavelet Packet Decomposition. The study extracted both time-domain and frequency-domain features, including Root Mean Square (RMS), Hjorth parameters (Activity, Mobility, Complexity), variance, and skewness, as well as band powers (delta, theta, alpha, beta) calculated using Welch's method.

Statistical measures such as mean and kurtosis from the Power Spectral Density (PSD) were also included. Multiple machine learning models were evaluated, including VM, KNN, Gradient Boosting, Random Forest, and Logistic Regression.

The best performance was achieved by the Diverse Feature Bagging Classifier, which reached an accuracy of 98.85% on 80% of the training data and a nearly perfect AUC (0.99) on the ROC curve.

For our implementation, we used time and frequency domain features, and developed a grid search backbone based on the machine learning models discussed in the study.

C. *Slower Binocular Rivalry in the Autistic Brain by Spiegel et al. (2019) [3]*

This study investigates whether neural dynamics of binocular rivalry differ between autistic and non-autistic individuals, reflecting potential excitation-inhibition (E/I) imbalances in the visual cortex. Using EEG, steady-state visually evoked potentials (SSVEPs) were recorded as 37 participants (18 autistic, 19 controls) viewed true and simulated frequency-tagged binocular rivalry displays.

A key innovation was the Neural Rivalry Index (NRI), a metric distinguishing autistic from control participants. Derived from amplitude traces at 5.7 Hz and 8.5 Hz, the NRI involves subtracting and demeaning these traces to compute a "difference time course." The power spectrum is estimated using the multitaper method, averaged across trials, and normalized. The NRI is defined as the frequency where the normalized cumulative power spectrum reaches its half-maximum. Participants with NRIs beyond two standard deviations from the group mean were excluded to ensure data reliability.

Figures 1 and 2 represent EEG data collected during a binocular rivalry simulation task. The first figure shows oscillatory patterns for participants with autism, while the second represents control participants. The red and blue lines depict neural transitions between "Suppressed to Dominant" and "Dominant to Suppressed" perceptual states, emphasizing group differences in neural processing during perceptual rivalry.

Findings revealed significantly slower neural and behavioral binocular rivalry alternations in autistic individuals. This supports the NRI's potential as a non-verbal marker for autism and its relevance for developmental and cross-species studies.

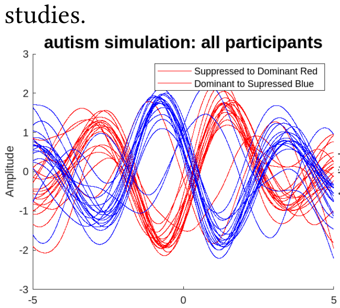

Fig. 1: Simulation of rivalry in the autistic brain

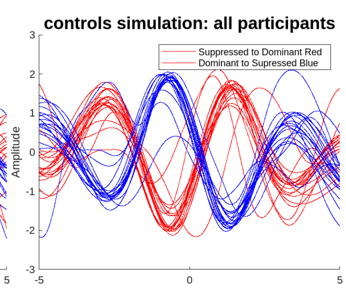

Fig. 2: Simulation of rivalry in the control brain

D. *Electroencephalography (EEG) signal processing for epilepsy and autism spectrum disorder diagnosis by Ibrahim et al. (2018) [4]*

The study investigates EEG signal processing techniques for the diagnosis of Autism Spectrum Disorder (ASD) and epilepsy, focusing on feature extraction and classification methods. EEG datasets from various sources, including the University of Bonn, MIT, King Abdulaziz University (KAU), and the authors' own recordings, were utilized. The analysis included two key classification problems: Autistic vs. Normal (single-channel) and Autistic vs. Normal (multi-channel).

The EEG signals were decomposed into frequency subbands (delta, theta, alpha, beta, and gamma) using Discrete Wavelet Transform (DWT). Features were extracted using statistical measures such as Standard Deviation, Band Power, Shannon Entropy, and Largest Lyapunov Exponent. The classification was performed using four machine learning models: Artificial Neural Networks (ANNs), k-Nearest Neighbor (KNN), Support Vector Machines (SVM), and Linear Discriminant Analysis (LDA).

The results highlighted that the combination of DWT, Shannon Entropy, and KNN yielded the highest classification accuracy, outperforming existing methods for ASD and epilepsy diagnosis, particularly on larger datasets.

In our implementation, we employed the DWT pipeline, integrated with Shannon Entropy for feature extraction, and used KNN for classification, replicating the most effective approach from the study.

E. *EEG-Based Autism Detection Using Multi-Input 1D Convolutional Neural Networks by Naaman (2024) [5]*

The study proposes a novel method for early Autism Spectrum Disorder (ASD) detection using EEG signals and a multi-input 1D convolutional neural network (CNN), aiming to improve diagnostic precision and efficiency.

EEG data were collected from 29 participants (20 ASD and 9 healthy controls) using a G-tec EEG cap with 16 channels (e.g., FP1, FP2, F3, Fz) under artifact-free protocols. Data augmentation through an overlapping sliding window technique expanded the dataset to 820 ASD samples and 369 healthy samples.

The model used a multi-input 1D CNN, processing each EEG channel independently with convolutional layers, batch normalization, ReLU activation, and fully connected layers. Training employed the Adam optimizer with a ten-fold cross-validation, mini-batch size of 16, and a learning rate of 0.001.

The approach achieved 99.16% accuracy with frontal channels, 98.32% with central and temporal channels, and 97.65% with occipital channels. Outperforming models like ResNet18 (98.88%) and SVM-based methods (96.44%), this method demonstrates superior accuracy and potential for EEG-based ASD diagnosis.

III. Methodology

A. *Multi Channel Evaluation*

a) ***Features and Preprocessing:*** The consolidated model integrates EEG data from two distinct datasets—Aging and BCIAUT—each characterized by unique experimental setups, frequency ranges, and formats. Harmonizing these datasets while preserving critical signal characteristics posed a significant challenge. A structured workflow was adopted to address these discrepancies, as outlined below:

Data Harmonization:

- ATAR (Advanced Time-Aligned Resampling): Applied to align temporal inconsistencies across datasets.
- Band-Pass Filtering (1-40 Hz): Ensured uniformity in the frequency range of interest.
- Common Channel Selection: Identified eight shared EEG channels to standardize input dimensions.
- Standardization: All features were normalized before combining datasets to ensure compatibility.

Feature Extraction:

Feature extraction was performed using two dedicated scripts:

- Script 1: Focused on both time-domain and frequency-domain features.
- Script 2: Extracted advanced Recurrence Quantification Analysis (RQA) features.

Time-Domain Features:

- *Mean*: Represents the average brain activity, reflecting variability and rapid changes.
- *Standard Deviation*: Measures the variation or dispersion of the EEG signal from the mean.
- *Variance*: Quantifies how far the EEG signal values spread out from the mean.
- *Peak-to-Peak*: Difference between the maximum and minimum values of the signal.
- *Root Mean Squared*: Captures the overall amplitude and energy of the signal.
- *Hjorth Parameters*: Describe the power, frequency characteristics, and complexity of the signal.
- *Skewness*: Measures the asymmetry of the EEG signal distribution.
- *Kurtosis*: Quantifies the "tailedness" of the EEG signal distribution.

Frequency-Domain Features:

- *Welch's Method*: Estimates the power spectral density (PSD) using Fast Fourier Transform (FFT), useful for analyzing non-stationary EEG signals.
- *Spectral Entropy and Statistical Features*: Measure the unpredictability and statistical characteristics of the power distribution across frequency bands.

The feature sets were extracted individually for the Aging dataset and the combined dataset (Aging + BCIAUT). This approach ensured that each dataset's unique characteristics were captured while maintaining consistency across the combined dataset.

b) ***Feature Selection:***

The Featurewiz Python library, using the Minimum Redundancy Maximum Relevance (MRMR) algorithm, was employed for selecting the most relevant features while minimizing redundancy.

These features provide a comprehensive analysis of EEG signals, capturing both time and frequency domain characteristics for effective classification.

c) ***Model Architecture:***

This study utilized ten different machine learning models to classify autism using EEG data, focusing on diverse algo-

TABLE I: Non-linear signal features derived from Recurrence Quantitative Analysis

Feature	Abbreviation	Description	Inference
Entropy	L_entr	Entropy from diagonal lines of the recurrence plot, linked to Shannon entropy.	High L_entr implies chaotic or irregular EEG behavior.
Max line length	L_max	Longest diagonal line, indicating the largest recurrent pattern.	High L_max suggests a dominant pattern in the EEG signal.
Mean line length	L_mean	Average diagonal line length, reflecting signal complexity.	Large L_mean indicates extended periods of signal stability.
Recurrence rate	R_rate	Proportion of recurrent points, indicating signal predictability.	High R_rate implies high predictability in the EEG signal.
Determinism	DET	Proportion of diagonal recurrent points, reflecting signal determinism.	High DET indicates strong predictability in the EEG signal.
Laminarity	LAM	Proportion of vertical recurrent points, indicating signal structure.	High LAM suggests repeated patterns in the signal.
Trapping time	TT	Average length of recurrent points, indicating time spent in similar states.	High TT shows prolonged stability in the system.

TABLE II: Non-linear signal features derived from Recurrence Quantitative Analysis

Model	Key Hyperparameters
Support Vector Machine (SVM)	Kernel, C, Gamma
Random Forest	Number of trees, Max depth
Gradient Boosting	Learning rate, Max depth
K-Nearest Neighbors	Number of neighbors
Logistic Regression	Penalty, Solver
Decision Tree	Max depth, Split criteria
AdaBoost	Number of estimators, Learning rate
Extra Trees	Number of estimators, Max depth
Multilayer Perceptron (MLP)	Hidden layers, Activation function
Ensemble	No hyperparameter tuning

rithms to capture varying decision boundaries. The following models were implemented and optimized as in table Table II

Hyperparameter optimization was performed using GridSearchCV with 5-fold cross-validation for robust parameter selection. Each model was evaluated using accuracy, ROC AUC, and F1-score metrics to ensure a comprehensive performance analysis.

d) ***Ensemble Methods**:*

To enhance prediction performance, ensemble techniques were incorporated:

- *Voting Classifier*: It aggregates predictions from the base models (SVM, KNN, RF, GB) using soft voting, which averages their probability outputs. This method leverages the diverse decision-making strategies of individual models to create a balanced and generalized classifier.

- *Stacking Classifier*: The base models provide their predictions as inputs to a meta-learner (Logistic Regression). Logistic Regression combines these predictions to generate the final classification, effectively synthesizing the strengths of the base models while mitigating their individual weaknesses.

Base Models - SVM, Random Forest, Gradient Boosting, and KNN:

Final Estimator: Logistic Regression, trained on predictions from base models to make the final decision. These ensemble methods leveraged the strengths of multiple algorithms, thereby reducing overfitting and improving generalization.

Patient-Level Aggregation:

Since EEG data is recorded at the channel level, individual predictions were aggregated to the patient level. Two methods were employed:

- *Majority Voting*: The class label most frequently predicted by channels.
- *Average Probability*: Averaging the probability outputs across all channels for a patient, which showed superior performance.

This multi-level approach ensured clinical applicability and robust classification of autism using EEG data.

B. *Single Channel Evaluation*

The implementation process began with replicating and enhancing the findings of [3], which demonstrated that the Neural Rivalry Index (NRI) significantly improved classification accuracy by capturing neural dynamics of rivalry alternations independent of behavioral reports. While the original study reported an accuracy of 86.5%, our replication efforts achieved a best accuracy of 76%. This discrepancy was attributed to variations in data preprocessing, feature extraction, or participant demographics.

To address these challenges, alternative methods were implemented based on techniques described in [4]. Specifically, wavelet transformation and Shannon entropy were used for feature extraction, combined with k-Nearest Neighbors (KNN) for classification. These methods resulted in accuracies ranging from 75% to 77%, aligning with initial replication results but still falling short of the original study's performance.

To further refine the approach, additional experiments incorporated a broader range of features and classifiers. Frequency-domain features, such as frequency-specific amplitude differences at 5.7 Hz and 8.5 Hz derived via Fast Fourier Transform (FFT), were combined with the NRI, calculated from rivalry transition data in RLS files. A diverse set of classifiers, including Random Forest, XGBoost, Decision Tree, and KNN, was employed to evaluate the performance of the extracted features. To ensure robust evaluation, a Leave-One-Out cross-validation framework was adopted throughout the process.

***Feature Engineering**:* Feature extraction was performed using the following methods:

- *FFT Features*: Amplitude differences and ratios were extracted at specific frequencies (5.7 Hz and 8.5 Hz) to capture frequency-specific neural activity.
- *Neural Rivalry Index (NRI)*: Calculated from rivalry transition data in RLS files, providing a neural measure of rivalry switch rates.
- *Classifiers used:*
  - Random Forest (n_estimators=10,000): Leveraged for its robustness through ensemble learning.
  - XGBoost (n_estimators=10,000): Utilized for boosted performance with gradient-based optimization.
  - Decision Tree: Employed as a simpler baseline for comparison.
  - KNN (k=3): Evaluated for proximity-based decision-making capabilities.

IV. Results

A. *Single Channel Evaluation*

The study explored the integration of FFT-based features (amplitude differences at 5.7 Hz and 8.5 Hz) with the Neural Rivalry Index (NRI) using ensemble classifiers such

TABLE III: Grid Search Results with highest Accuracy Scores for ML Models

Dataset → Features ↓	Aging	Combined
TimeFreq	MLP - 0.87	RandomForest - 0.84
RQA	Stacking Classifier - 0.79	MLP - 0.81
Features Combined	ExtraTrees - 0.85	MLP - 0.78

TABLE IV: Grid Search Results with highest Accuracy Scores for Neural Networks

Dataset → Features ↓	Aging	Combined
TimeFreq	Single Channel - 0.68	Multi Channel - 0.92
RQA	Single Channel - 0.71	Single Channel - 0.84
Features Combined	Single Channel - 0.75	Single Channel - 0.77

as Random Forest and XGBoost. Hyperparameter optimization was performed through a systematic Grid Search, and a Leave-One-Out cross-validation approach ensured robust evaluation. The optimal configuration, combining NRI and FFT features with XGBoost, achieved a peak accuracy of 91.98%. However, the addition of wavelet transformation and Shannon entropy led to reduced performance, underscoring the importance of selecting complementary features. This approach highlights the value of targeted feature engineering and classifier selection in enhancing EEG-based classification.

B. *Multi Channel Evaluation*

Combining datasets, though broadening perspectives, reduced accuracy due to differences in experimental setups. Among models, MLP achieved the highest accuracy but proved inconsistent, while Gradient Boosting and Extra Trees offered more reliable performance.

Single-channel analysis excelled over multi-channel methods for smaller datasets, as multi-channel processing reduced dataset size and accuracy. However, augmenting multi-channel analysis with simpler time-frequency features achieved the best accuracy of 0.92. Multi-channel approaches showed clear advantages for larger datasets, emphasizing the importance of dataset volume in leveraging their potential.

These insights guide optimal strategies for EEG-based autism diagnostics, balancing dataset characteristics and model capabilities.

V. Conclusion

- **Dataset Challenges and Data Integration:** The study addresses several challenges inherent in EEG research, including variability in signal quality, subject demographics, and recording protocols. To mitigate these limitations, the dataset integrates recordings from both Autism Spectrum Disorder (ASD) and neurotypical controls, sourced from diverse studies such as [2] and [3]. This approach provides a varied and representative dataset for model training and testing, improving generalization across populations.
- **Harmonizing Diverse EEG Datasets:** A significant novelty in this study lies in the ability to harmonize disparate EEG datasets, including both single-channel and multi-channel recordings. This process addresses the complexities of standardizing and merging data from varied sources and configurations. The successful integration of such diverse datasets allows for a unified framework that adapts to different EEG formats and setups, enhancing model robustness and broadening its application.

References

[1] W. J. Bosl, H. Tager-Flusberg, and C. A. Nelson, "EEG analytics for early detection of autism spectrum disorder: a data-driven approach," *Scientific reports*, vol. 8, no. 1, p. 6828, 2018.

[2] Y. Emara, "Exploring Machine Learning Approaches for Autism Diagnosis through EEG Signal Analysis," 2024.

[3] A. Spiegel, J. Mentch, A. J. Haskins, and C. E. Robertson, "Slower binocular rivalry in the autistic brain," *Current Biology*, vol. 29, no. 17, pp. 2948–2953, 2019.

[4] S. Ibrahim, R. Djemal, and A. Alsuwailem, "Electroencephalography (EEG) signal processing for epilepsy and autism spectrum disorder diagnosis," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 1, pp. 16–26, 2018, doi: https://doi.org/10.1016/j.bbe.2017.08.006.

[5] O. Naaman, "EEG-Based Autism Detection Using Multi-Input 1D Convolutional Neural Networks," *Traitement du Signal*, vol. 41, no. 1, p. 333, 2024.