https://doi.org/10.31436/iiumej.v27i3.4522

A CatBoost-Based Framework for Autism Spectrum Disorder Classification from EEG Signals with Blackman–Harris FIR Filtering

BULEUN WALADAH[1], MELINDA MELINDA[1*], YUNIDAR YUNIDAR[1], AUFA RAFIKI[1], SITI RUSDIANA[2], WEI KITT WONG[3]

[1]Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia
[2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh, Indonesia
[3]Department of Electrical and Computer Engineering, Faculty of Engineering and Science, Curtin University Malaysia, Miri, Malaysia

**Corresponding author: melinda@usk.ac.id*

(Received: 15 June 2026; Accepted: 24 August 2026; Published online: 11 September 2026)

ABSTRACT: Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder marked by impairments in social interaction, communication, and repetitive behavior. Early identification remains difficult given its clinical heterogeneity and the subjectivity of conventional diagnosis. This study proposes an EEG-based ASD classification framework combining Blackman–Harris FIR filtering, feature extraction and selection, and machine learning. EEG data were collected from 16 participants (8 ASD, 8 neurotypical) using a 16-channel system, band-pass filtered at 4–40 Hz, segmented into 4-second epochs with 50% overlap, and represented by 176 temporal and spectral features per epoch (11 per channel), followed by fold-specific Mutual Information feature selection. SVM-RBF, CatBoost, LDA, and Gaussian Naive Bayes were evaluated using nested subject-wise 8-fold cross-validation. Based on 7,145 pooled out-of-fold epoch predictions, CatBoost achieved the highest accuracy (99.58%), followed by LDA (97.47%), SVM-RBF (91.60%), and Gaussian Naive Bayes (57.49%). An exploratory paired t-test on fold-wise accuracy showed a nominally significant CatBoost advantage over SVM-RBF (mean difference 8.00 percentage points, $p = 0.042$; Cohen's $d_z = 0.88$). At the participant level, majority voting correctly classified 16/16 participants with CatBoost and LDA, 15/16 with SVM-RBF, and 9/16 with Gaussian Naive Bayes; the 95% CI for 100% accuracy was 79.41–100.00%. These findings support preliminary within-cohort feasibility but not clinical or screening readiness.

ABSTRAK: Gangguan Spektrum Autisme (ASD) adalah gangguan perkembangan neuro yang dicirikan oleh kekurangan interaksi sosial, komunikasi, dan tingkah laku berulang. Pengesanan awal masih sukar disebabkan kepelbagaian klinikal dan sifat subjektif diagnosis konvensional. Kajian ini mencadangkan rangka kerja pengelasan ASD berasaskan EEG yang menggabungkan penapisan Blackman–Harris FIR, pengekstrakan dan pemilihan ciri, serta pembelajaran mesin. Data EEG diperoleh daripada 16 peserta (8 ASD, 8 neurotipikal) menggunakan sistem 16 saluran, ditapis jalur 4–40 Hz, dibahagikan kepada epok 4 saat dengan pertindihan 50%, dan diwakili oleh 176 ciri temporal dan spektral bagi setiap epok (11 setiap saluran), diikuti pemilihan ciri Informasi Bersama khusus bagi setiap lipatan. SVM-RBF, CatBoost, LDA, dan Gaussian Naive Bayes dinilai menggunakan validasi silang 8-lipatan berasaskan subjek. Berdasarkan 7,145 ramalan epok terkumpul luar lipatan, CatBoost mencapai ketepatan tertinggi (99.58%), diikuti LDA (97.47%), SVM-RBF (91.60%), dan Gaussian Naive Bayes (57.49%). Kajian pada ketepatan setiap lipatan bagi ujian-t

berpasangan menunjukkan kelebihan nominal yang signifikan bagi CatBoost berbanding SVM-RBF (perbezaan min 8.00 mata peratus, $p = 0.042$; Cohen's $d_z = 0.88$). Pada peringkat peserta, pengelasan undian majoriti adalah betul iaitu 16/16 peserta menggunakan CatBoost dan LDA, 15/16 menggunakan SVM-RBF, dan 9/16 menggunakan Gaussian Naive Bayes; 95% CI bagi ketepatan 100% ialah 79.41–100.00%. Dapatan ini menyokong kebolehlaksanaan awal kohort ini, tetapi bukan kesediaan klinikal atau saringan.

KEYWORDS: *ASD; Electroencephalography; Blackman–Harris FIR; SVM-RBF; CatBoost*

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental disorder characterized by persistent impairments in social communication and restricted or repetitive behaviors [1], [2]. Neurophysiological studies have reported atypical functional connectivity and neuronal synchronization in individuals with ASD, including altered local and long-range interactions during resting conditions [3], [4]. Electroencephalography (EEG) provides a non-invasive means of examining these neural dynamics and has increasingly been used for ASD classification through conventional feature-based and more recent representation-learning approaches [5], [6], [7], [8].

EEG-based classification remains challenging because of signal nonstationarity, artifact sensitivity, and variability in acquisition protocols and participant characteristics [8], [9]. Preprocessing is therefore important for reducing out-of-band components while preserving information within the frequency range of interest. Window-based FIR filtering provides a finite, non-recursive response with controllable spectral characteristics, while the Blackman–Harris window is distinguished by strong sidelobe suppression that can reduce spectral leakage associated with finite-length signal processing [10], [11], [12]. In this study, a 4–40 Hz Blackman–Harris FIR filter was therefore evaluated against raw EEG and a conventional Butterworth preprocessing alternative [13].

Following preprocessing, temporal and spectral EEG features were classified using SVM-RBF and CatBoost as the primary models, with LDA and Gaussian Naive Bayes (GNB) included as additional baselines. SVM-RBF provides nonlinear decision boundaries through the radial basis function kernel [14], whereas CatBoost uses gradient-boosted decision trees to model complex nonlinear relationships [15], [16]. Recent studies have also demonstrated CatBoost's applicability to EEG and other biomedical classification problems [17], [18], supporting its evaluation as a complementary tree-based approach.

Previous EEG-based ASD studies have primarily emphasized feature representation and classifier development, whereas preprocessing choices have received comparatively less systematic evaluation [6], [8]. Moreover, direct comparison of nonlinear kernel-based and gradient-boosting pipelines under participant-disjoint validation remains limited. Accordingly, this study evaluates Blackman–Harris FIR preprocessing by comparing it with raw EEG and Butterworth filtering, and compares SVM-RBF, CatBoost, LDA, and GNB using nested subject-wise 8-fold cross-validation. Performance is assessed at both epoch and participant levels using classification metrics, confidence intervals, aggregation sensitivity analysis, and statistical testing to determine the within-cohort feasibility and stability of the evaluated EEG-based ASD classification pipelines.

https://doi.org/10.31436/iiumej.v27i3.4522

2. MATERIALS AND METHODS

This study developed a machine-learning framework distinguishing individuals with ASD from typically developing (TD) controls using multichannel EEG. The workflow comprised participant recruitment, EEG acquisition, electrode configuration, preprocessing, epoch segmentation, feature extraction and selection, classification, and statistical evaluation. EEG signals were acquired via a 16-channel OpenBCI Cyton system, preprocessed with a Blackman–Harris FIR band-pass filter, then used to extract temporal and spectral features for classification, as illustrated in Fig. 1.

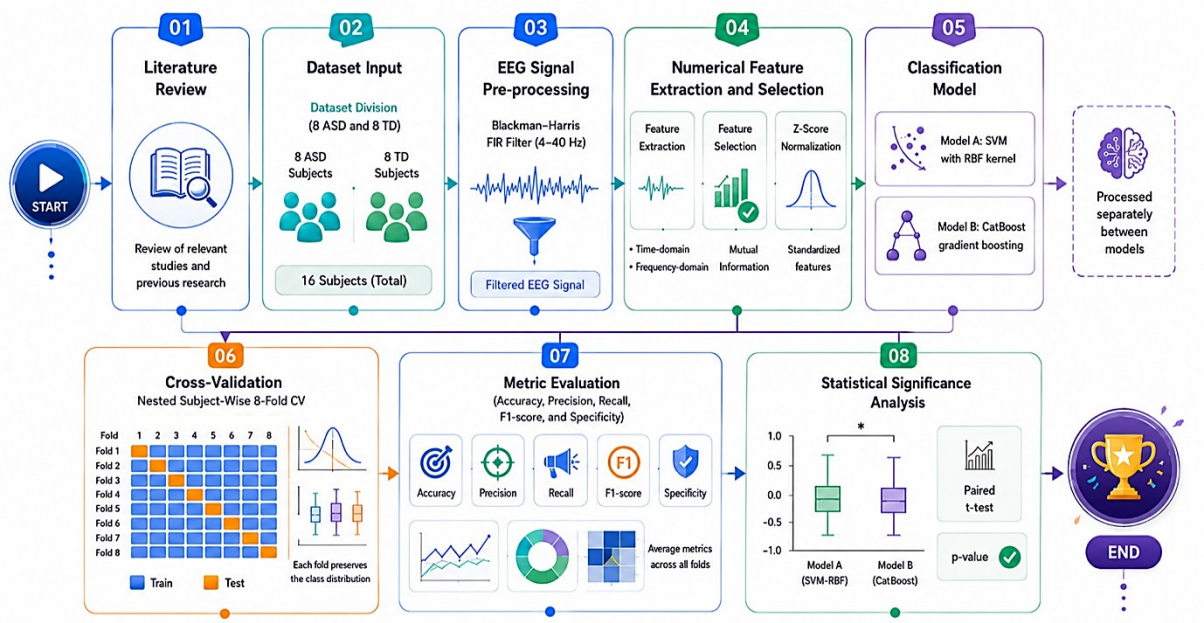

Figure 1. Workflow of the proposed EEG classification system.

2.1. EEG Dataset and Participants

The EEG dataset was collected as a primary dataset at Universitas Syiah Kuala, Indonesia, under a controlled acquisition protocol and environment. The dataset comprised recordings from 16 participants: eight individuals with ASD aged 15–25 years and eight TD controls aged 17–25 years. ASD participants were recruited from a special education school in Banda Aceh, whereas TD controls were recruited from the surrounding university community. The acquisition and participant-group framework are consistent with previous EEG-based ASD datasets and comparative studies [7], [8].

Participant information reported in this study reflects the demographic and diagnostic-group metadata available in the original acquisition records, including group designation, age range, and recruitment source. Additional structured clinical variables, such as standardized ASD severity scores, medication status, detailed comorbidity profiles, and formal group-matching variables, were not part of the analytical metadata and were therefore not retrospectively inferred or used during model development.

Following participant preparation and electrode placement, each EEG recording lasted approximately 15 minutes. The Ethics Committee of the Faculty of Medicine, Universitas Syiah Kuala approved the study protocol (approval no. 117/EA/FK/2024).

https://doi.org/10.31436/iiumej.v27i3.4522

2.2. EEG Acquisition Device and Channel Configuration

EEG signals were acquired using an OpenBCI Cyton-based biosensing system at a sampling frequency of 250 Hz. Sixteen EEG channels were positioned according to the international 10–20 system at Fp1, Fp2, F3, F4, F7, F8, C3, C4, T3, T4, T5, T6, P3, P4, O1, and O2, as illustrated in Fig. 2. This configuration is consistent with the acquisition protocol reported in previous studies using the same EEG setup [8]. The ground electrode was positioned at AFz and the common reference electrode at CPz; these auxiliary electrodes were used only for signal acquisition and were not treated as EEG analysis channels.

Recordings were performed under an eyes-open resting condition in a quiet, low-distraction environment between 09:00 and 12:00. Participants were seated comfortably and instructed to remain calm, minimize head and body movements, and face a blank wall. Conductive gel was applied to improve scalp-electrode contact, and signals were monitored in real time using the OpenBCI software. The acquisition hardware incorporated an ADS1299 analog front-end with 24-bit resolution. No dedicated ocular, myogenic, or movement-artifact correction was applied; subsequent preprocessing was limited to the filtering procedures described in Section 2.3.

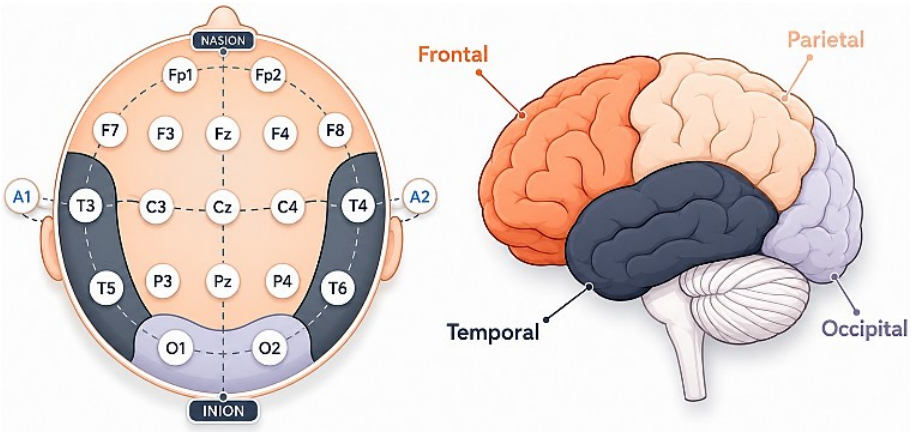

Figure 2. Electrode placement of the 16-channel EEG recording.

2.3. EEG Preprocessing Using Blackman–Harris FIR Filter

Raw continuous EEG recordings were preprocessed before feature extraction to attenuate baseline drift and out-of-band frequency components that could interfere with subsequent analysis. No dedicated ocular, myogenic, or movement-artifact correction procedure was applied; therefore, the filtering stage was not interpreted as removing these artifacts.

Preprocessing was performed independently on each EEG channel using a finite impulse response (FIR) band-pass filter designed with a four-term Blackman–Harris window. This window was selected because its low sidelobe levels and strong stopband attenuation reduce spectral leakage associated with truncation of the ideal impulse response [10], [11], [12]. Filtering was applied to the continuous EEG recordings before epoch segmentation.

The FIR filter was implemented using forward–backward filtering (filtfilt), producing a zero-phase response and preventing phase shifts in the filtered signals. Signal boundaries were handled using the default odd-symmetric extension implemented by SciPy's filtfilt function. For the 101-coefficient FIR filter, a padding length of 303 samples, corresponding to approximately 1.21 s at 250 Hz, was applied at each end of the continuous recording. Filtering the continuous signals before segmentation also avoided introducing separate filtering transients at the boundaries of individual 4-second epochs.

https://doi.org/10.31436/iiumej.v27i3.4522

An FIR filter produces each output sample from a finite weighted combination of the current and preceding input samples without recursive feedback. The filtering operation is expressed as:

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k) \tag{1}$$

where $y(n)$ is the filtered EEG signal, $x(n-k)$ denotes the input signal delayed by k samples, $h(k)$ represents the FIR coefficient, and N is the number of filter coefficients.

In a window-based FIR design, the final coefficients are obtained by multiplying the ideal impulse response by the selected window function:

$$h(n) = h_d(n)w(n) \tag{2}$$

where $h(n)$ is the resulting FIR coefficient, $h_d(n)$ is the ideal impulse response, and $w(n)$ is the window function. Applying a tapered window smooths the truncated impulse response and reduces spectral leakage associated with abrupt truncation [11].

The Blackman–Harris window has a smooth and symmetric profile with strong sidelobe suppression, as illustrated in Fig. 3.

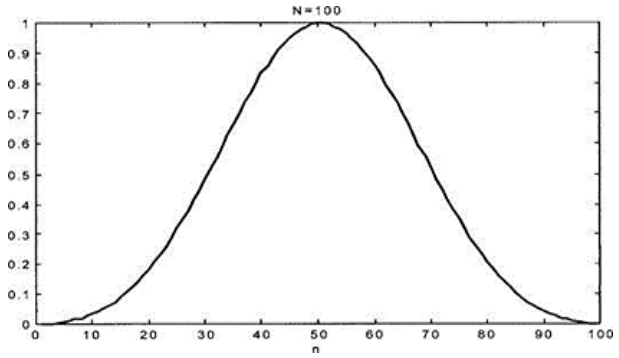

Figure 3. Blackman-Harris window.

The four-term Blackman–Harris window is defined as [12]:

$$w(n) = a_0 - a_1\cos\left(\frac{2\pi n}{L-1}\right) + a_2\cos\left(\frac{4\pi n}{L-1}\right) - a_3\cos\left(\frac{6\pi n}{L-1}\right), 0 \le n \le L-1 \tag{3}$$

where L is the window length, n is the sample index, and a_0, a_1, a_2, and a_3 are the Blackman–Harris coefficients. The coefficients used were $a_0 = 0.35875$, $a_1 = 0.48829$, $a_2 = 0.14128$, and $a_3 = 0.01168$. Table 1 summarizes the complete FIR configuration.

Table 1. Configuration of the Blackman–Harris FIR band-pass filter

Parameter	Description / Value
Sampling frequency	250 Hz
Filter type	FIR band-pass filter
Window function	Blackman–Harris window
Passband frequency range	4–40 Hz
Filter order/length	100th-order FIR filter / 101 coefficients
Filtering mode	Zero-phase forward–backward filtering

For the preprocessing ablation reported in Section 3.1, a conventional Butterworth IIR band-pass filter was included as a comparator. The comparator used the same 4–40 Hz frequency range and 250 Hz sampling frequency as the conventional Butterworth

https://doi.org/10.31436/iiumej.v27i3.4522

preprocessing approach reported by Muliyadi *et al.* [13]. In the present study, the Butterworth filter was implemented as a fourth-order IIR band-pass filter with lower and upper cut-off frequencies of 4 and 40 Hz, respectively. The filter was applied independently to each EEG channel before epoch segmentation using forward–backward filtering to obtain a zero-phase response. Thus, the Butterworth and Blackman–Harris FIR pipelines used the same 4–40 Hz passband, sampling frequency, filtering direction, and subsequent epoch-generation procedure, while differing in filter structure and design order. Accordingly, the ablation was intended to compare two consistently implemented preprocessing alternatives rather than filters matched by numerical order.

The 4–40 Hz passband retained theta, alpha, beta, and low-gamma activity while attenuating lower-frequency drift and higher-frequency out-of-band components. The effects of the preprocessing alternatives were subsequently assessed using power spectral density analysis and the classification ablation reported in Section 3.1.

The filtered recordings were segmented into 4-second epochs, corresponding to 1,000 samples at a sampling frequency of 250 Hz. Consecutive epochs overlapped by 50%, resulting in a 500-sample (2-second) step; successive epochs therefore covered 0–4 s, 2–6 s, 4–8 s, and so forth.

A total of 7,145 epochs were included across the 16 participants. Fifteen participants contributed 449 epochs each from complete 15-minute recordings, whereas one ASD participant contributed 410 epochs because the recording ended before the planned duration, corresponding to approximately 13.7 minutes of usable EEG data. Each epoch contained all 16 EEG channels and was treated as one sample for epoch-level feature extraction. All epochs from the same participant remained grouped within the same outer cross-validation fold, as described in Section 2.5. The resulting epochs were subsequently processed using the feature-extraction and feature-selection procedures described in Section 2.4.

2.4. Feature Extraction and Feature Selection

After preprocessing, temporal and spectral features were extracted independently from each channel of every 4-second EEG epoch. Eleven features were calculated per channel: log-variance, root mean square (RMS), Hjorth activity, mobility, and complexity, spectral centroid, spectral bandwidth, and absolute band power in the theta, alpha, beta, and gamma bands. These features were selected to represent complementary amplitude, temporal-dynamic, and spectral characteristics of the EEG signals, consistent with previous EEG-based ASD studies [7]. With 16 EEG channels, the resulting representation comprised 176 features per epoch.

For an EEG epoch $x[n]$, $n = 0, \dots, N-1$, log-variance and RMS were calculated as

$$LogVar = ln[Var(x) + \varepsilon], \qquad RMS = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2\,[n]}, \tag{4}$$

where N is the number of samples and $\varepsilon = 10^{-12}$ is a small constant introduced for numerical stability.

The Hjorth parameters were calculated as

$$\text{Activity} = \text{Var}(x) + \varepsilon, \tag{5}$$

$$\text{Mobility} = \sqrt{\frac{\text{Var}(\Delta x)+\varepsilon}{\text{Var}(x)+\varepsilon}}, \tag{6}$$

https://doi.org/10.31436/iiumej.v27i3.4522

$$\text{Complexity} = \frac{\sqrt{\frac{\text{Var}(\Delta^2 x)+\varepsilon}{\text{Var}(\Delta x)+\varepsilon}}}{\text{Mobility}+\varepsilon}, \tag{7}$$

where Δx and $\Delta^2 x$ denote the first and second discrete differences, respectively. Hjorth parameters were included because they characterize EEG signal activity and temporal complexity and have previously been applied to the analysis of ASD-related EEG patterns [7].

Spectral features were derived from a Hann-windowed one-sided real Fourier transform. The power spectral density (PSD) was estimated as

$$X[k] = rFFTx[n]w[n], \qquad P[k] = \frac{|X[k]|^2}{f_s \sum_{n=0}^{N-1} w^2[n]}, \tag{8}$$

where $w[n]$ denotes the Hann window and f_s is the sampling frequency. Each 4-second epoch contained $N = 1{,}000$ samples at $f_s = 250$ Hz; therefore, a 1,000-point rFFT was applied, providing a frequency resolution of 0.25 Hz. No Welch segmentation or spectral averaging was used for feature extraction.

Using the resulting PSD, the spectral centroid f_c and spectral bandwidth B_w were calculated as

$$f_c = \frac{\sum_k f_k P[k]}{\sum_k P[k]}, \qquad B_w = \sqrt{\frac{\sum_k P k^2}{\sum_k P[k]}}, \tag{9}$$

where f_k represents the frequency associated with the k-th spectral bin.

Absolute band power was calculated by integrating the PSD within each predefined frequency range:

$$P_B = \int_{f_l}^{f_h} P\,(f), df, \tag{10}$$

where f_l and f_h denote the lower and upper limits of the corresponding frequency band. Numerical integration was performed using the trapezoidal rule. The frequency ranges were theta, 4–8 Hz; alpha, 8–13 Hz; beta, 13–30 Hz; and gamma, 30–40 Hz. The same feature definitions and spectral parameters were applied consistently to all channels, epochs, and participants.

Mutual Information (MI) was subsequently used for model-specific feature selection within the nested subject-wise cross-validation procedure described in Section 2.5. MI ranking and feature selection were performed exclusively within the corresponding training partitions, without access to the outer-test participants. Any subsequent feature scaling was performed according to the classifier-specific pipeline described in Section 2.6.

2.5. Nested Subject-Wise 8-Fold Cross-Validation

A nested subject-wise cross-validation procedure was used to minimize participant leakage during model development and performance estimation. In the outer loop, the 16 participants (8 ASD and 8 TD) were divided into eight stratified folds, each containing one ASD and one TD participant. In each outer iteration, one fold comprising two participants was reserved exclusively for testing, while the remaining 14 participants were used for model development. All epochs belonging to the same participant remained within the same partition throughout the procedure.

Model selection was performed exclusively within the 14 outer-training participants using an inner subject-wise cross-validation loop. The inner folds were participant-disjoint, such that epochs from an inner-validation participant were never included in the corresponding inner-

training set. Classifier-specific preprocessing, MI estimation, feature ranking and selection, model hyperparameter selection, and, where applicable, CatBoost early stopping were performed using only the inner-training and inner-validation data. Z-score normalization was applied only to the SVM-RBF pipeline and was fitted exclusively on the corresponding training partition.

After the optimal configuration was determined from the inner cross-validation results, the selected preprocessing and feature-selection parameters were refitted using all 14 outer-training participants, and the resulting model was evaluated once on the two previously unseen outer-test participants. This procedure was repeated for all eight outer folds so that each participant contributed exactly once to the out-of-fold test predictions.

For model selection, candidate MI feature-retention fractions of 10%, 20%, ..., 100% were evaluated within the inner cross-validation loop. For SVM-RBF, the candidate search space comprised $C \in \{0.001, 0.01, 0.1, 1, 10, 100\}$ and $\gamma \in \{\text{scale}, 0.01, 0.1\}$. For CatBoost, candidate configurations combined learning rates of $\{0.01, 0.03\}$, tree depths of $\{2, 4\}$, and L2 leaf-regularization values of $\{10, 50\}$. Candidate configurations were ranked primarily using mean inner-validation balanced accuracy, with lower inner-fold standard deviation and the predefined tie-breaking criteria used when necessary.

2.6. Classification Models

The 176-dimensional EEG feature vectors were classified into ASD and TD classes using SVM-RBF and CatBoost as the primary classifiers, with Linear Discriminant Analysis (LDA) and Gaussian Naive Bayes (GNB) included as additional baseline models. All classifiers received the same initial feature representation. Model-specific feature selection and model configuration were performed exclusively within the nested subject-wise validation procedure described in Section 2.5, with the outer-test participants remaining completely unseen during feature selection and model development.

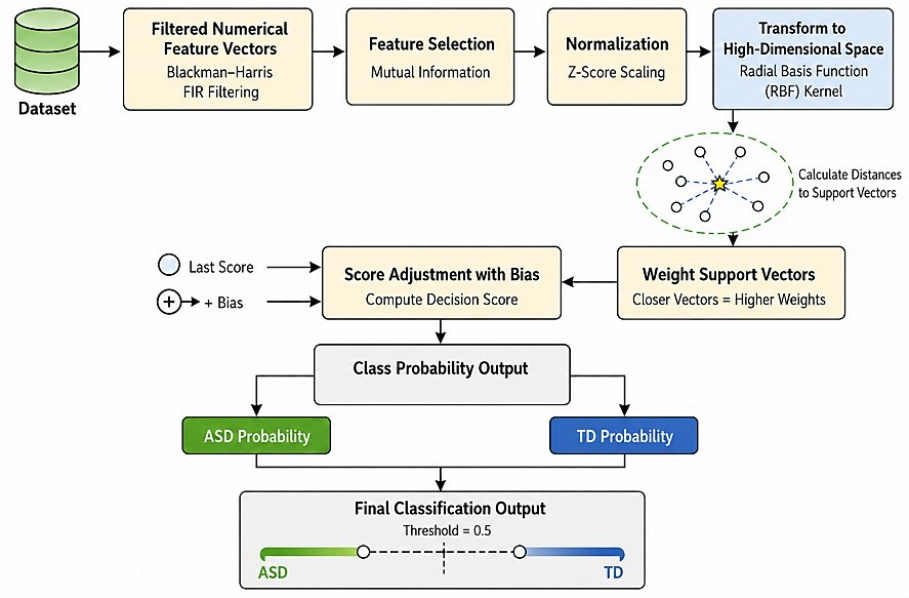

Figure 4. SVM-RBF classification workflow.

2.6.1. Support Vector Machine with Radial Basis Function Kernel

SVM-RBF classified the selected EEG feature vectors after Blackman–Harris FIR preprocessing, feature extraction, MI-based feature selection, and Z-score normalization, as

illustrated in Fig. 4. The RBF kernel enables nonlinear separation of EEG feature patterns by mapping pairwise similarity according to [14]:

$$K(x, x') = \exp\left(-\gamma \parallel x - x' \parallel^2\right) \tag{11}$$

where x and x' are EEG feature vectors and γ controls the kernel width. The SVM decision function is expressed as:

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i \, y_i K(x_i, x) + b) \tag{12}$$

where α_i denotes the support-vector coefficient, y_i is the corresponding class label, $K(x_i, x)$ is the kernel similarity, and b is the bias term.

SVM-RBF model selection was performed exclusively within the inner subject-wise cross-validation loop. For each inner split, Z-score normalization and MI ranking were fitted using only the inner-training participants and subsequently applied to the participant-disjoint inner-validation set. Candidate feature-retention fractions and SVM-RBF configurations were evaluated using mean inner-validation balanced accuracy. After the optimal configuration was selected, MI ranking and normalization were refitted using all participants in the corresponding outer-training set, and the final SVM-RBF model was trained without access to the outer-test participants. The resulting configuration used $C = 0.01$, $\gamma = \text{scale}$, and the top 30% of MI-ranked features. The trained model was then evaluated once on the previously unseen outer-test participants.

2.6.2. CatBoost

CatBoost is a gradient-boosted decision-tree classifier that iteratively combines weak learners while using ordered boosting to reduce prediction shift [16]. As illustrated in Fig. 5, the selected EEG features were used to estimate class probabilities for ASD and TD classification.

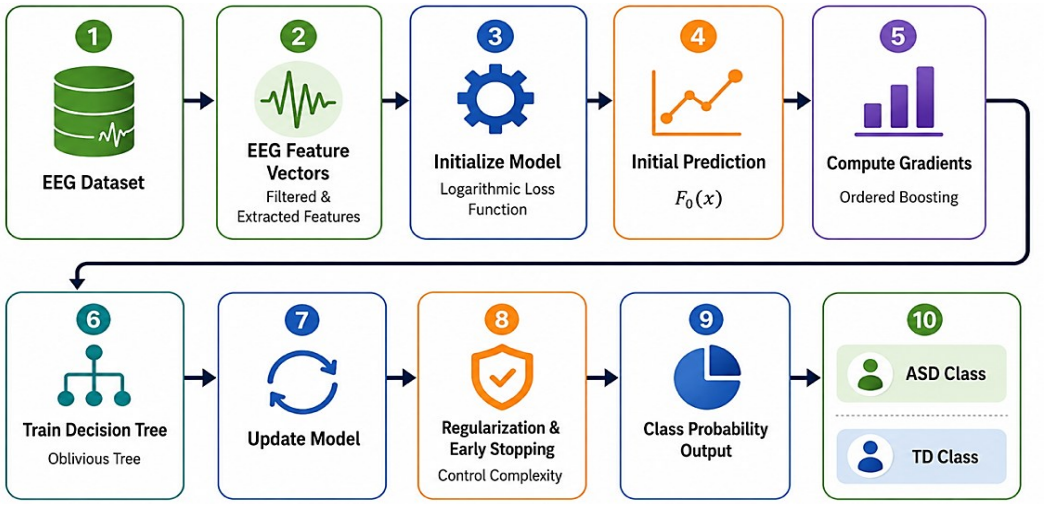

Figure 5. CatBoost classification workflow.

The binary classification objective can be represented as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{13}$$

where $L(y_i, \hat{y}_i)$ denotes the classification loss, $\Omega(f_k)$ represents the regularization term for the k-th tree, and $\hat{y}_i$ is the predicted output. The ensemble is updated iteratively according to:

$$F_{t+1}(x) = F_t(x) + \eta h_t(x) \tag{14}$$

where $F_t(x)$ is the ensemble at iteration t, η is the learning rate, and $h_t(x)$ is the newly added decision tree.

CatBoost feature selection, hyperparameter selection, and early-stopping assessment were performed exclusively within the inner subject-wise cross-validation loop. For each inner split, MI ranking was fitted using only the inner-training participants and subsequently applied to the participant-disjoint inner-validation set. No feature scaling was applied to CatBoost because of its tree-based structure. Candidate CatBoost configurations were evaluated using mean inner-validation balanced accuracy, while early stopping was monitored exclusively on the inner-validation participants. No epochs, labels, or predictions from the outer-test participants were used for feature selection, hyperparameter selection, or early stopping. After the optimal configuration was selected, MI ranking was refitted using all participants in the corresponding outer-training set, and the final CatBoost model was retrained using the selected hyperparameters and boosting-iteration setting determined from the inner-validation procedure. The model was then evaluated once on the previously unseen outer-test participants. The final CatBoost configuration used a maximum of 800 boosting iterations, a learning rate of 0.01, tree depth of 2, L2 leaf regularization of 50, bagging temperature of 2.0, an early-stopping patience of 30 iterations, and the top 40% of MI-ranked features.

2.6.3. Additional Baseline Models

LDA and GNB were included as additional baseline classifiers to determine whether the observed class separation was specific to the primary nonlinear models. Both baselines were evaluated using the same outer subject-wise folds and the same initial 176-dimensional feature representation. All data-dependent preprocessing and feature-selection operations were restricted to the corresponding training partitions.

LDA used the singular-value decomposition (SVD) solver without shrinkage, whereas GNB used a variance-smoothing parameter of 1×10^{-9}. Both baseline models used the top 40% of MI-ranked features within the nested validation procedure. These models were reported as baseline pipelines rather than as evidence of direct classifier superiority because SVM-RBF and the other models did not necessarily retain identical final feature subsets.

2.6.4. Participant-Level Aggregation

Participant-level predictions were generated exclusively from the out-of-fold epoch predictions of each held-out participant. Majority voting was used as the primary aggregation rule. For participant s, the proportion of epochs predicted as ASD was calculated as:

$$V_s = \frac{1}{N_s}\sum_{j=1}^{N_s} \mathbb{I}\left(\hat{y}_{sj} = 1\right), \tag{15}$$

where N_s is the number of epochs belonging to participant s, $\hat{y}_{sj}$ is the predicted label for epoch j, with ASD coded as 1 and TD as 0, and $\mathbb{I}(\cdot)$ is the indicator function. The final participant-level prediction was defined as:

$$\hat{Y}_s = \begin{cases} 1\ (\text{ASD}), & V_s > 0.5, \\ 0\ (\text{TD}), & V_s < 0.5, \\ \mathbb{I}(\bar{p}_s \geq 0.5), & V_s = 0.5, \end{cases} \tag{16}$$

where $\bar{p}_s$ denotes the mean out-of-fold ASD probability across all epochs of participant s. Thus, when ASD and TD votes were equal, the participant was classified using a prespecified probability threshold of 0.5; no voting ties occurred in the present analysis.

Each participant contributed exactly one final classification decision, whether 449 or 410 epochs were available, preventing participants with more epochs from receiving greater weight

in the participant-level metrics. Accuracy, sensitivity, specificity, precision, and F1-score were calculated from the 16 participant-level decisions. Exact two-sided Clopper–Pearson 95% confidence intervals were calculated for accuracy, sensitivity, and specificity.

As a sensitivity analysis, participant-level predictions were also obtained by averaging the out-of-fold ASD probabilities across all epochs of each participant and applying the same fixed threshold of 0.5. Neither the majority-voting rule nor the probability threshold was optimized using the outer-test predictions.

2.6.5. Performance Evaluation

All four classifiers were evaluated using the outer subject-wise folds of the nested cross-validation procedure described in Section 2.5. Performance was assessed using accuracy, precision, sensitivity (recall), specificity, F1-score, ROC-AUC, and confusion matrices, with ASD treated as the positive class and TD as the negative class. Participant-level accuracy, sensitivity, and specificity were additionally reported with exact two-sided Clopper–Pearson 95% confidence intervals [19].

Accuracy was designated as the primary overall metric for the inferential comparison between SVM-RBF and CatBoost, whereas precision, sensitivity, specificity, and F1-score were retained as complementary descriptive measures. An exploratory paired-samples t-test was performed using the accuracy values obtained from the same eight outer subject-wise test folds [20], [21]. For fold i, the paired difference was defined as:

$$d_i = A_{\mathrm{CatBoost},i} - A_{\mathrm{SVM-RBF},i}, \tag{17}$$

such that a positive d_i indicated higher accuracy for CatBoost. The paired t-statistic was calculated as

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \tag{18}$$

where $\bar{d}$ is the mean paired difference, s_d is the standard deviation of the paired differences, and $n = 8$ is the number of outer folds. Statistical significance was assessed at a nominal level of $\alpha = 0.05$. The two-sided 95% confidence interval for the mean difference was calculated as

$$\bar{d} \pm t_{0.975,n-1}\frac{s_d}{\sqrt{n}}, \tag{19}$$

and the standardized paired effect size was calculated using Cohen's d_z:

$$d_z = \frac{\bar{d}}{s_d}. \tag{20}$$

Because the eight outer folds contained substantially overlapping training sets, their accuracy estimates were not fully independent [22]. Therefore, the paired t-test, confidence interval, and effect size were interpreted as exploratory within-cohort summaries rather than confirmatory participant- or population-level inference. No multiplicity correction was applied because accuracy was the only metric subjected to this inferential comparison.

To complement the exploratory fold-wise analysis, an exact participant-level permutation test was performed using the final out-of-fold CatBoost participant predictions. Participant identity was treated as the exchangeable unit, while the original class composition of eight ASD and eight TD participants was preserved. Because the cohort contained only 16 participants, all possible class-balanced diagnostic-label assignments were evaluated exhaustively rather than approximated using Monte Carlo sampling. The total number of unique assignments was

https://doi.org/10.31436/iiumej.v27i3.4522

$$M = \binom{16}{8} = 12{,}870. \tag{21}$$

For each assignment, the diagnostic labels were reassigned to the fixed out-of-fold participant predictions, and participant-level accuracy was recalculated. The exact one-sided permutation p-value was defined as

$$p_{\text{perm}} \frac{1}{M} \sum_{b=1}^{M} I\,(A_b \geq A_{\text{obs}}), \tag{22}$$

where A_{obs} denotes the observed participant-level accuracy, A_b is the accuracy obtained under the b-th class-balanced label assignment, $I(\cdot)$ is the indicator function, and M is the total number of unique class-balanced assignments defined in Eq. (21).

This analysis tested whether the association between the held-out participant predictions and the diagnostic labels exceeded that expected under random class-balanced label assignment. Because all possible assignments were enumerated, no Monte Carlo approximation or correction was required. The test was conditional on the observed out-of-fold predictions and was therefore interpreted as a within-cohort robustness analysis rather than as an assessment of external generalizability [23].

3. Results and Discussion

This section presents the preprocessing ablation results, out-of-fold epoch-level classifier performance, participant-level classification using majority voting, exploratory fold-wise statistical comparisons, and the study's principal findings and limitations.

3.1. EEG Signal Preprocessing Results

Blackman–Harris FIR filtering attenuated baseline drift and frequency components outside the 4–40 Hz passband and was not interpreted as an artifact-removal procedure, as no dedicated ocular, myogenic, or movement-artifact correction was applied. Representative ASD and TD signals before and after filtering are shown in Figs. 6 and 7, while the corresponding PSD comparison in Fig. 8 confirms attenuation below 4 Hz and above 40 Hz with preservation of activity within the selected theta, alpha, beta, and low-gamma ranges.

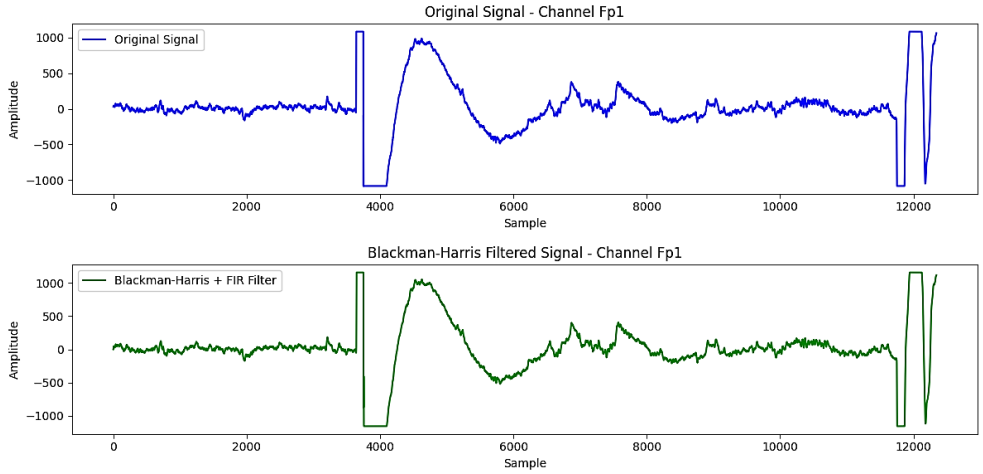

Figure 6. ASD EEG signals before and after Blackman–Harris FIR filtering.

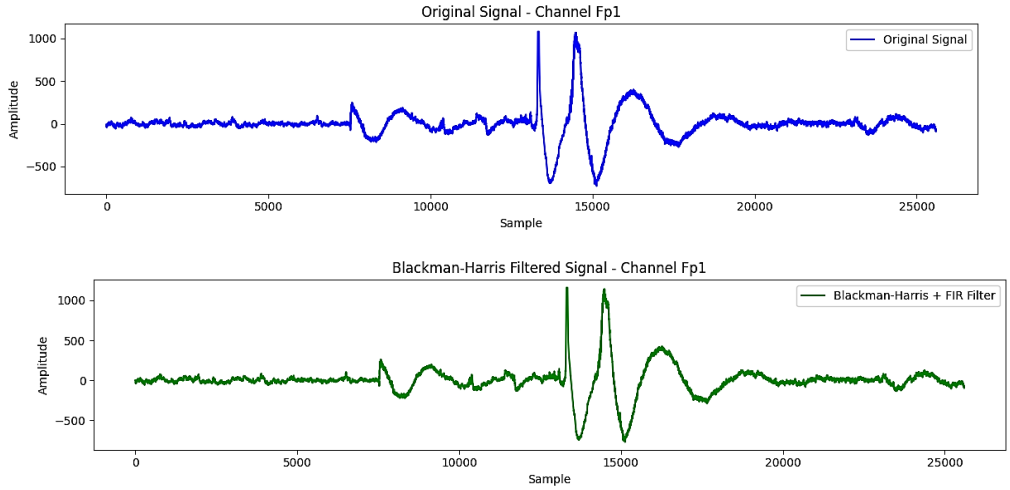

Figure 7. TD EEG signals before and after Blackman–Harris FIR filtering.

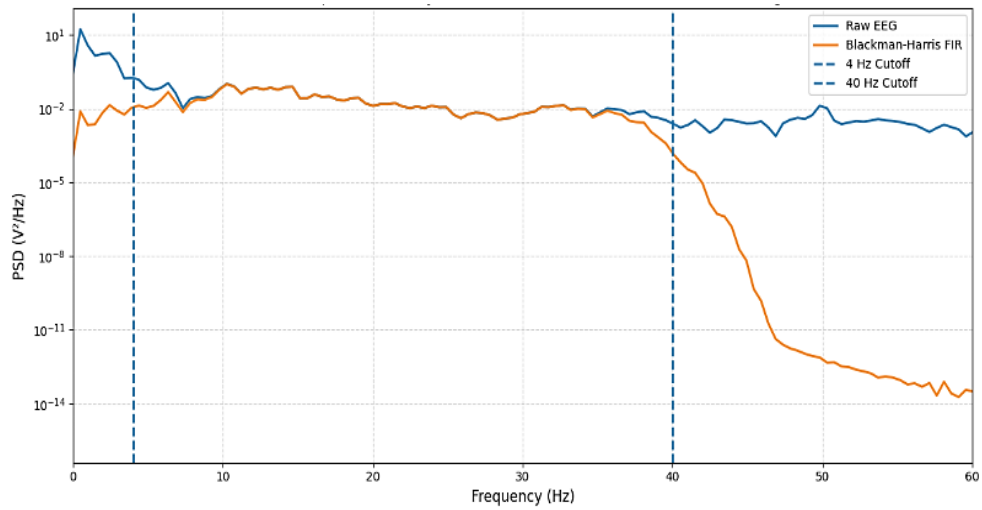

Figure 8. PSD of EEG signals before and after Blackman–Harris FIR filtering.

To quantify the effect of preprocessing on classification, raw EEG, Butterworth band-pass filtering, and Blackman–Harris FIR filtering were compared, as summarized in Table 2.

Table 2. Performance comparison of EEG preprocessing approaches

Preprocessing pipeline	Mean accuracy (%)	SD (%)
Raw EEG + SVM-RBF	80.01	1.14
Raw EEG + CatBoost	88.92	3.88
Butterworth band-pass + SVM-RBF	88.53	1.16
Butterworth band-pass + CatBoost	93.97	0.86
Blackman–Harris FIR + SVM-RBF	91.59	9.15
Blackman–Harris FIR + CatBoost	**99.58**	**0.23**

Both filtering approaches improved accuracy relative to raw EEG, with Blackman–Harris FIR producing the highest accuracy for both classifiers. The Blackman–Harris FIR–CatBoost pipeline achieved the best performance at 99.58%, indicating that preprocessing influenced feature discriminability within the present cohort.

https://doi.org/10.31436/iiumej.v27i3.4522

3.2. Epoch-Level Performance of SVM-RBF

SVM-RBF was evaluated using out-of-fold predictions from the outer folds of the nested subject-wise eight-fold cross-validation. In each fold, one ASD and one TD participant were held out for testing, with all epochs from each participant remaining exclusively within the corresponding partition. Table 3 summarizes the fold-wise epoch-level results.

Table 3. Out-of-fold epoch-level performance of SVM-RBF across eight outer folds

Fold	Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)	Specificity (%)
Fold 1	93.21	100.00	86.41	92.71	100.00
Fold 2	71.71	64.03	99.11	77.80	44.32
Fold 3	94.88	100.00	89.76	94.60	100.00
Fold 4	99.44	100.00	98.89	99.44	100.00
Fold 5	88.31	81.73	98.66	89.40	77.95
Fold 6	88.47	88.40	87.32	87.85	89.53
Fold 7	99.89	100.00	99.78	99.89	100.00
Fold 8	96.77	100.00	93.54	96.66	100.00
Mean	**91.59**	**91.77**	**94.18**	**92.29**	**88.98**
SD	**9.15**	**13.20**	**5.67**	**7.29**	**19.73**

SVM-RBF achieved a mean fold-wise accuracy of 91.59%, with notable variability across held-out participants. Fold 2 showed the lowest performance, where specificity decreased to 44.32% because 250 of 449 TD epochs were classified as ASD, whereas specificity reached 100% in five folds. Despite this variability, sensitivity remained generally high across folds. The corresponding out-of-fold confusion matrices are shown in Fig. 9.

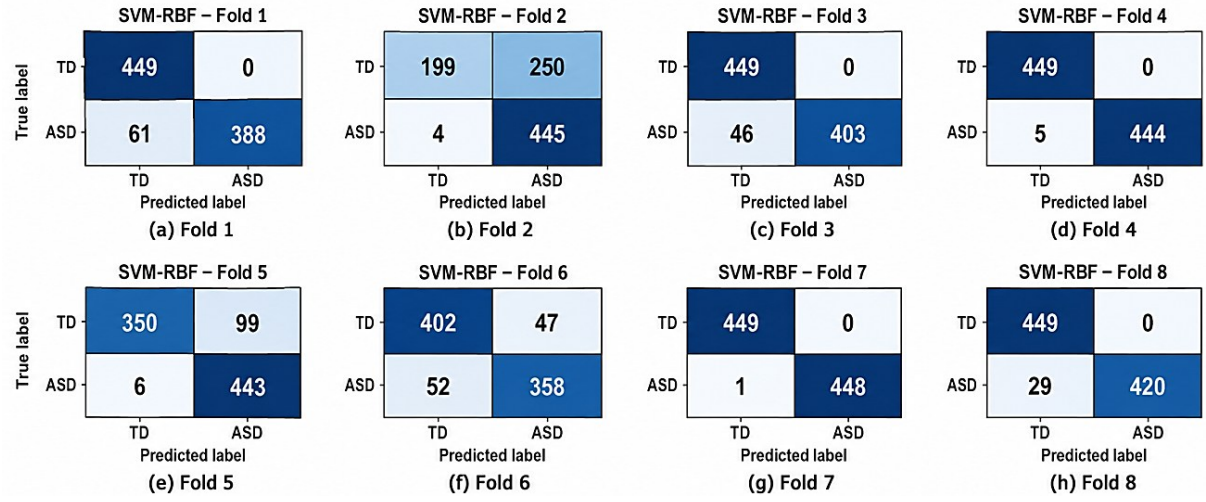

Figure 9. Confusion matrices of SVM-RBF across eight outer folds.

3.3. Epoch-Level Performance of CatBoost

CatBoost was evaluated using the same eight outer subject-wise folds, with fold-wise out-of-fold epoch-level performance summarized in Table 4. CatBoost achieved consistently high performance across all outer folds, with a mean accuracy of 99.58% and low fold-to-fold variability (SD = 0.23%). Accuracy ranged from 99.33% to 99.88%. Across the 7,145 out-of-fold epoch predictions, CatBoost produced eight false-positive and 22 false-negative classifications. Fig. 10 shows the corresponding confusion matrices.

Table 4. Out-of-fold epoch-level performance of CatBoost across eight outer folds

Fold	Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)	Specificity (%)
Fold 1	99.33	100.00	98.66	99.33	100.00
Fold 2	99.33	98.90	99.78	99.33	98.89
Fold 3	99.78	99.78	99.78	99.78	99.78
Fold 4	99.67	100.00	99.33	99.66	100.00
Fold 5	99.78	100.00	99.55	99.78	100.00
Fold 6	99.88	99.76	100.00	99.88	99.78
Fold 7	99.55	99.78	99.33	99.55	99.78
Fold 8	99.33	100.00	98.66	99.33	100.00
Mean	**99.58**	**99.78**	**99.39**	**99.58**	**99.78**
SD	**0.23**	**0.37**	**0.50**	**0.23**	**0.38**

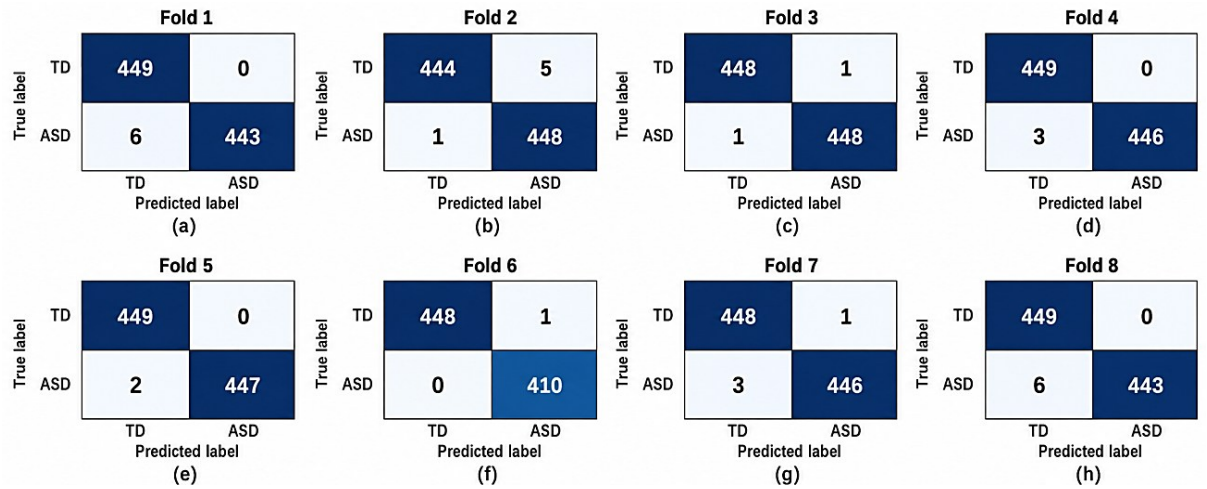

Figure 10. Confusion matrices of CatBoost across eight outer folds.

3.4. Pooled Out-of-Fold Epoch-Level Comparison

The pooled out-of-fold predictions from all 7,145 epochs were used to compare the four classification pipelines. SVM-RBF retained the top 30% of MI-ranked features, whereas CatBoost, LDA, and Gaussian Naive Bayes retained the top 40%. Thus, Table 5 compares separately configured pipelines rather than classifiers using identical final feature subsets.

Table 5. Pooled out-of-fold epoch-level performance of the four classification pipelines

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	ROC-AUC (%)
CatBoost	**99.58**	**99.77**	99.38	**99.78**	**99.58**	**99.99**
LDA	97.47	96.65	98.31	96.63	97.47	99.58
SVM-RBF	91.60	89.43	94.26	88.98	91.78	98.21
Gaussian Naive Bayes	57.49	53.94	**99.49**	15.95	69.95	63.41

CatBoost achieved the highest pooled accuracy, precision, specificity, F1-score, and ROC-AUC, followed by LDA and SVM-RBF. Gaussian Naive Bayes showed high sensitivity but very low specificity (15.95%), indicating a strong tendency to classify TD epochs as ASD. These results compare the performance of the configured pipelines and do not, by themselves, exclude potential cohort-related confounding.

https://doi.org/10.31436/iiumej.v27i3.4522

3.5. Participant-Level Classification Using Majority Voting

Participant-level performance was obtained by majority voting across the out-of-fold epoch predictions of each held-out participant. Each participant contributed one final decision regardless of the number of available epochs, thereby preventing participants with more epochs from receiving greater weight. Table 6 summarizes the results.

Table 6. Participant-level performance using majority voting

Model	Accuracy, % (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	F1-score (%)
CatBoost	**100.00 (79.41–100.00)**	**100.00 (63.06–100.00)**	**100.00 (63.06–100.00)**	**100.00**
LDA	**100.00 (79.41–100.00)**	**100.00 (63.06–100.00)**	**100.00 (63.06–100.00)**	**100.00**
SVM-RBF	93.75 (69.77–99.84)	100.00 (63.06–100.00)	87.50 (47.35–99.68)	94.12
Gaussian Naive Bayes	56.25 (29.88–80.25)	100.00 (63.06–100.00)	12.50 (0.32–52.65)	69.57

CatBoost and LDA correctly classified all 16 participants, whereas SVM-RBF classified 15/16 and Gaussian Naive Bayes 9/16. Despite the perfect CatBoost and LDA point estimates, the 95% CI for accuracy was 79.41–100.00%, reflecting uncertainty due to the small participant sample.

As a sensitivity analysis, participant labels were also obtained by averaging out-of-fold ASD probabilities and applying the fixed threshold of 0.5. CatBoost, LDA, and Gaussian Naive Bayes produced unchanged decisions, whereas SVM-RBF improved from 15/16 under majority voting to 16/16 under probability averaging, indicating some sensitivity to the aggregation rule.

For CatBoost, an exact participant-level permutation test evaluated all $M = \binom{16}{8} = 12{,}870$. class-balanced diagnostic-label assignments. Only one assignment achieved an accuracy equal to the observed 100.00%, yielding $p_{perm} = 7.77 \times 10^{-5} (p < 0.001)$. This supports an association between the held-out CatBoost predictions and diagnostic labels beyond random class-balanced assignment, while remaining conditional on the observed out-of-fold predictions and therefore representing within-cohort rather than external or clinical evidence.

3.6. Exploratory Fold-Wise Statistical Comparison of Accuracy

Accuracy was treated as the primary inferential metric, while precision, sensitivity, specificity, and F1-score were reported descriptively. An exploratory paired t-test compared CatBoost and SVM-RBF accuracy across the same eight outer subject-wise folds. Because these folds shared substantially overlapping training sets, the resulting inference was interpreted as within-cohort exploratory evidence rather than confirmatory population-level evidence.

Table 7. Exploratory paired t-test results for CatBoost minus SVM-RBF fold-wise differences

Comparison	Mean difference (pp)	95% CI (pp)	t-statistic	p-value	Cohen's d_z	Conclusion ($\alpha = 0.05$)
CatBoost – SVM-RBF	8.00	0.38–15.61	2.48	0.042	0.88	Nominally significant; exploratory

CatBoost achieved a mean fold-wise accuracy 8.00 percentage points higher than SVM-RBF (95% CI: 0.38–15.61; $t = 2.48$, $p = 0.042$, Cohen's $d_z = 0.88$). Although the difference

was nominally significant with a large standardized effect, it should be interpreted cautiously because the eight folds were not fully independent and the analysis was based on only 16 participants.

3.7. Reproducibility and Implementation Details

All analyses were implemented in Python 3.12.12 using NumPy 2.0.2, Pandas 2.3.3, SciPy 1.16.3, scikit-learn 1.6.1, and CatBoost 1.2.10. Experiments were conducted on Kaggle Notebook under Linux 6.6.122+ x86_64 with glibc 2.35, an Intel Xeon 2.00 GHz CPU (four cores), 31.35 GB RAM, and two NVIDIA Tesla T4 GPUs with 15,360 MiB memory each. A fixed random seed of 1234 was used for data splitting, MI estimation, and algorithms supporting random-state control. All data-dependent preprocessing, feature selection, model tuning, and early-stopping procedures followed the nested subject-wise protocol described in Section 2.5.

SVM-RBF used $C = 0.01$, $\gamma = \text{scale}$, and the top 30% of MI-ranked features. CatBoost used up to 800 iterations, a learning rate of 0.01, depth 2, L2 leaf regularization of 50, bagging temperature of 2.0, early-stopping patience of 30, and the top 40% of MI-ranked features. LDA used the SVD solver without shrinkage, while Gaussian Naive Bayes used a variance-smoothing parameter of 1×10^{-9}; both baselines retained the top 40% of MI-ranked features. All results in Sections 3.2–3.6 were derived from the same 7,145 out-of-fold predictions for each classifier.

3.8. Discussion

The preprocessing ablation showed that filtering influenced the discriminability of the extracted EEG features. Both Butterworth and Blackman–Harris FIR filtering improved performance relative to raw EEG, with Blackman–Harris FIR producing the highest accuracy for the evaluated classifiers. However, the 4–40 Hz filtering stage only attenuated baseline drift and out-of-band components and did not specifically remove ocular, myogenic, or movement artifacts.

At the epoch level, CatBoost achieved the highest pooled performance and the lowest fold-to-fold variability, followed by LDA and SVM-RBF. SVM-RBF was more sensitive to the held-out participants, particularly in Fold 2, where specificity decreased to 44.32%. The exploratory CatBoost–SVM-RBF comparison yielded an 8.00-percentage-point difference in fold-wise accuracy ($p = 0.042$, 95% CI: 0.38–15.61; Cohen's $d_z = 0.88$). Because the outer folds shared substantially overlapping training sets, interpret this result as exploratory within-cohort evidence rather than confirmatory population-level inference.

Participant-level evaluation provides the more clinically relevant assessment because each participant contributes only one final decision. CatBoost and LDA correctly classified all 16 participants, whereas SVM-RBF classified 15/16 under majority voting and 16/16 under probability averaging. Thus, participant-level superiority of CatBoost over LDA was not established, and the SVM-RBF result showed some sensitivity to the aggregation rule. Moreover, despite 100% accuracy for CatBoost and LDA, the 95% confidence interval remained wide, reflecting the uncertainty associated with only eight participants per group.

Using 50% overlapping 4-second epochs increased the number of segments but introduced dependence between adjacent observations. Subject-wise cross-validation prevented participant leakage between outer training and test sets, but epoch-level results still represent correlated repeated observations from the same individuals and should therefore be regarded as secondary measures of segment-level discriminability. In addition, the present ablation did

not separately quantify the effects of MI-based feature selection, temporal versus spectral feature families, or epoch overlap; these component-wise effects require further evaluation under the same nested subject-wise framework.

The single-site design also leaves open potential recruitment, demographic, recording-session, and hardware-related confounding. High classification accuracy across several models does not exclude these alternative explanations. Accordingly, the findings should be regarded as preliminary within-cohort evidence, and larger, demographically matched, independent, and multi-site datasets are required before considering the framework for ASD screening or clinical decision support.

Finally, the exact participant-level permutation test for CatBoost yielded $p_{\text{perm}} = 7.77 \times 10^{-5}$, indicating that the observed association between held-out predictions and diagnostic labels was unlikely under random class-balanced assignment. Nevertheless, this test remained conditional on the observed out-of-fold predictions and does not establish external validity or eliminate potential cohort- and acquisition-related confounding.

4. CONCLUSIONS

This study developed an EEG-based ASD classification framework combining Blackman–Harris FIR preprocessing, temporal and spectral feature extraction, MI-based feature selection, and machine-learning classification. Blackman–Harris FIR preprocessing yielded higher classification accuracy than raw EEG and Butterworth filtering. Among the evaluated pipelines, CatBoost achieved the highest pooled out-of-fold epoch-level accuracy (99.58%), followed by LDA (97.47%), SVM-RBF (91.60%), and Gaussian Naive Bayes (57.49%). The exploratory fold-wise comparison favored CatBoost over SVM-RBF ($p = 0.042$, Cohen's $d_z = 0.88$), although this result should be interpreted cautiously because the outer folds were not fully independent. At the participant level, CatBoost and LDA correctly classified all 16 participants, while SVM-RBF and Gaussian Naive Bayes classified 15/16 and 9/16, respectively. Thus, participant-level superiority of CatBoost over LDA was not established. The exact CatBoost permutation test also exceeded random class-balanced label assignment ($p < 0.001$). Overall, the findings support preliminary within-cohort feasibility, but the small sample size, overlapping epochs, single-site acquisition, and lack of external validation preclude claims of clinical or screening readiness.

ACKNOWLEDGEMENT

The authors thank Universitas Syiah Kuala for institutional and technical support.

REFERENCES

[1] Rogala J, et al. (2023) Enhancing autism spectrum disorder classification in children through the integration of traditional statistics and classical machine learning techniques in EEG analysis. Sci. Rep., 13:21748. https://doi.org/10.1038/s41598-023-49048-7

[2] Romero M, et al. (2016) Psychiatric comorbidities in autism spectrum disorder: A comparative study between DSM-IV-TR and DSM-5 diagnosis. Int. J. Clin. Health Psychol., 16(3):266-275. https://doi.org/10.1016/j.ijchp.2016.03.001

[3] Wantzen P, et al. (2022) EEG resting-state functional connectivity: evidence for an imbalance of external/internal information integration in autism. J. Neurodev. Disord., 14:47. https://doi.org/10.1186/s11689-022-09456-8

[4] Kang J, Xie H, Mao W, Wu J, Li X, Geng X (2023) EEG connectivity diversity differences between children with autism and typically developing children: A comparative study. Bioengineering, 10(9):1030. https://doi.org/10.3390/bioengineering10091030

[5] Dissanayake T, Fernando T, Denman S, Sridharan S, Fookes C (2021) Deep learning for patient-independent epileptic seizure prediction using scalp EEG signals. IEEE Sens. J., 21(7):9377-9388. https://doi.org/10.1109/JSEN.2021.3057076

[6] Melinda M, Gazali S, Away Y, Rafiki A, Wong WK, Muliyadi M, Rusdiana S (2026) Heavy–light soft-vote fusion of EEG heatmaps for autism spectrum disorder detection. J. Electron. Electromed. Eng. Med. Inform., 8(1):409-429. https://doi.org/10.35882/jeeemi.v8i1.1377

[7] Muliyadi M, Melinda M, Away Y, Gazali S (2025) Comparative analysis of normal and ASD EEG data using Hjorth parameters. 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA), pp. 1-6. https://doi.org/10.1109/ICHORA65333.2025.11017066

[8] Melinda M, Purnamasari PD, Fahmi F, Sinulingga EP, Muliyadi M, Away Y, Yunidar Y, Juwono FH (2025) A comprehensive EEG dataset and performance assessment for autism spectrum disorder. Sci. Rep., 15:34981. https://doi.org/10.1038/s41598-025-18934-7

[9] Ding Y, Robinson N, Zhang S, Zeng Q, Guan C (2023) TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. IEEE Trans. Affect. Comput., 14(3):2238-2250. https://doi.org/10.1109/TAFFC.2022.3169001

[10] Pant A, Kumar A (2024) Hanning FIR window filtering analysis for EEG signals. Biomedical Analysis, 1(2):111-123. https://doi.org/10.1016/j.bioana.2024.05.003

[11] Pant A, Kumar A, Verma C, Illés Z (2024) Comparative exploration on EEG signal filtering using window control methods. Results Control Optim., 17:100485. https://doi.org/10.1016/j.rico.2024.100485

[12] Harris FJ (1978) On the use of windows for harmonic analysis with the discrete Fourier transform. Proc. IEEE, 66(1):51-83. https://doi.org/10.1109/PROC.1978.10837

[13] Muliyadi M, Melinda M, Away Y, Gazali S, Rafiki A, Wong WK (2026) Bayesian-optimized Butterworth-PID preprocessing for EEG-based autism spectrum disorder classification. IEEE Access. https://doi.org/10.1109/ACCESS.2026.3709120

[14] Mastropietro A, Feldmann C, Bajorath J (2023) Calculation of exact Shapley values for explaining support vector machine models using the radial basis function kernel. Sci. Rep., 13:19561. https://doi.org/10.1038/s41598-023-46930-2

[15] Geeitha S, Ravishankar K, Cho J, Easwaramoorthy SV (2024) Integrating cat boost algorithm with triangulating feature importance to predict survival outcome in recurrent cervical cancer. Sci. Rep., 14:19828. https://doi.org/10.1038/s41598-024-67562-0

[16] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems 31 (NeurIPS 2018), pp. 6638-6648.

[17] Nassir LM, Ramadhan AJ, Al-Sharify NT, Khalaf MI, Ogaili AAF, Jaber AA, Al-Sharify ZT (2025) Robust multi-state EEG cognitive classification via optimized time-domain features and CatBoost. Int. J. Robot. Control Syst., 5(2):968-989. https://doi.org/10.31763/ijrcs.v5i2.1799

[18] Bhaskar N, Borhade RR, Barekar S, Bachute M, Bairagi V (2024) CNN-CatBoost ensemble deep learning model for enhanced disease detection and classification of kidney disease. Indonesian J. Electr. Eng. Comput. Sci., 34(1):144-151. https://doi.org/10.11591/ijeecs.v34.i1.pp144-151

[19] Garthwaite PH, Moustafa MW, Elfadaly FG (2024) Locally correct confidence intervals for a binomial proportion: A new criteria for an interval estimator. Scand. J. Stat., 51(1):220-244. https://doi.org/10.1111/sjos.12672

[20] Chicco D, Sichenze A, Jurman G (2025) A simple guide to the use of Student's t-test, Mann-Whitney U test, Chi-squared test, and Kruskal-Wallis test in biostatistics. BioData Min., 18:56. https://doi.org/10.1186/s13040-025-00465-6

[21] Nariya MK, Mills CE, Sorger PK, Sokolov A (2023) Paired evaluation of machine-learning models characterizes effects of confounders and outliers. Patterns, 4(8):100791. https://doi.org/10.1016/j.patter.2023.100791

[22] Jafrasteh B, Adeli E, Pohl KM, Kuceyeski A, Sabuncu MR, Zhao Q (2025) Statistical variability in comparing accuracy of neuroimaging based classification models via cross validation. Sci. Rep., 15:28745. https://doi.org/10.1038/s41598-025-12026-2

[23] Steinert S, Ruf V, Dzsotjan D, Großmann N, Schmidt A, Kuhn J, Küchemann S (2024) A refined approach for evaluating small datasets via binary classification using machine learning. PLoS ONE, 19(5):e0301276. https://doi.org/10.1371/journal.pone.0301276