ISSN 2007-9737

XGBoost-Based Classification of Epileptic versus Non-Epileptic EEG Signals

Said Azael Durán Contreras, Valeria Maeda-Gutiérrez*, Luis C. Reveles-Gómez

Universidad Autónoma de Zacatecas,
Unidad Académica de Ingeniería Eléctrica, Zacatecas,
Mexico

{39205498, valeria.maeda, luiscarlosreveles}@uaz.edu.mx

Abstract. Epilepsy is a chronic neurological disorder that affects more than 50 million people worldwide. The electroencephalogram (EEG) is the clinical gold standard for its assessment; however, manual interpretation is time-intensive and subject to inter-observer variability. This work presents an automatic binary classification system for distinguishing epileptic from normal EEG signals using the Extreme Gradient Boosting (XGBoost) algorithm. The Epileptic Seizure Recognition dataset from the University Hospital of Bonn was used, comprising 11,500 one-second EEG segments, each described by 178 temporal amplitude samples, with a 4:1 class imbalance. The pipeline integrates exploratory analysis in the time and frequency domains, class-imbalance compensation through the `scale_pos_weight` parameter, and 5-fold stratified cross-validation. The XGBoost classifier achieved an AUC-ROC of 0.9953 on the held-out test set and 0.9951 ± 0.0009 in cross-validation, with 97.57% accuracy and an F1-score of 0.9389 on the epileptic class. Under the same protocol, it obtains a discriminative ability statistically equivalent to that of Random Forest and Support Vector Machine, while providing the most balanced sensitivity–specificity trade-off and requiring $2.3\times$ to $3.5\times$ less training time. These results support XGBoost as an accurate and computationally efficient candidate for automated EEG screening in resource-limited clinical settings.

Keywords. Epilepsy, EEG, seizure detection, machine learning, XGBoost.

1 Introduction

Epilepsy is one of the most prevalent neurological disorders worldwide, affecting more than 50 million people across all age groups and socioeconomic conditions [21]. It is characterized by recurrent unprovoked seizures caused by abnormal and hypersynchronous neuronal activity in the brain [8]. Clinical manifestations range from brief lapses of awareness to prolonged tonic-clonic seizures, frequently producing cognitive, physical, and psychosocial consequences.

According to the International League Against Epilepsy (ILAE), approximately 70% of patients can achieve adequate seizure control when diagnosis and pharmacological treatment are established early and accurately [9].

However, delayed diagnosis and limited access to neurological specialists remain major obstacles, particularly in developing countries and rural regions. Untreated epilepsy is associated with severe long-term consequences, including cognitive deterioration, traumatic injury, psychosocial impairment, and sudden unexpected death in epilepsy (SUDEP) [6].

In Latin America, the prevalence of epilepsy has been estimated between 10 and 20 cases per $1{,}000$ inhabitants [4], with treatment gaps that may exceed 50% in rural communities. In Mexico, the limited number of certified clinical neurophysiologists per capita contributes to long waiting times and restricted access to electroencephalographic evaluation. Moreover, manual EEG interpretation is highly dependent on clinician experience and often leads to inter-observer variability and diagnostic delay in high-demand clinical environments.

These limitations motivate the development of automated and scalable computational tools to support electroencephalogram (EEG) interpretation in resource-constrained clinical settings [15, 10].

The EEG is a non-invasive neurophysiological technique that records cortical electrical activity using scalp electrodes distributed according to the international 10–20 system [12]. EEG signals typically exhibit amplitudes between 10 and $200\ \mu$V within a frequency range of approximately 0.5–100 Hz, conventionally divided into the delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–100 Hz) bands. During epileptic activity, recordings commonly present rhythmic high-amplitude discharges, spike–wave complexes, and increased spectral energy in the low-frequency delta and theta bands [16, 14], which provide a rich substrate for automated discrimination between ictal and non-ictal segments.

Over the last two decades, automated EEG classification using machine learning has been extensively investigated on the Bonn EEG dataset. Early approaches relied on handcrafted feature extraction combined with conventional classifiers: Subasi [18] employed discrete wavelet transform (DWT) coefficients together with a Support Vector Machine (SVM), achieving accuracies between 88% and 95%, though requiring substantial domain-specific feature engineering. Acharya et al. [1] incorporated nonlinear descriptors, such as entropy and detrended fluctuation analysis, into artificial neural networks, achieving approximately 95% accuracy at the expense of increased computational complexity.

Ensemble learning methods have also shown competitive performance; Kumar and Kanhangad [11] reported 97.2% accuracy using Random Forest (RF) models with wavelet and statistical features. More recently, deep-learning architectures such as the pyramidal 1D convolutional neural network (P-1D-CNN) proposed by Ullah et al. [19] reached up to 99.6% accuracy directly from raw EEG samples, albeit at the cost of substantially higher computational demands and limited interpretability.

Table 1. Quantitative properties of the dataset used for classification

Property	Value
Source	UCI / Kaggle derived from Bonn
Total segments	11,500
Original subjects/recordings	500, 5 subsets $\times$ 100
Segment duration	1 s
Samples per segment	178
Sampling frequency	$\approx$ 178 Hz
Original classes	5; 1: ictal; 2–5: non-ictal
Binary mapping	Class 1 $\rightarrow$ 1 epileptic; classes 2–5 $\rightarrow$ 0 normal
Class 0 – normal	9,200, 80%
Class 1 – epileptic	2,300, 20%
Class imbalance ratio	4:1
`scale_pos_weight`	4.0
Train / test split	80% / 20%, stratified

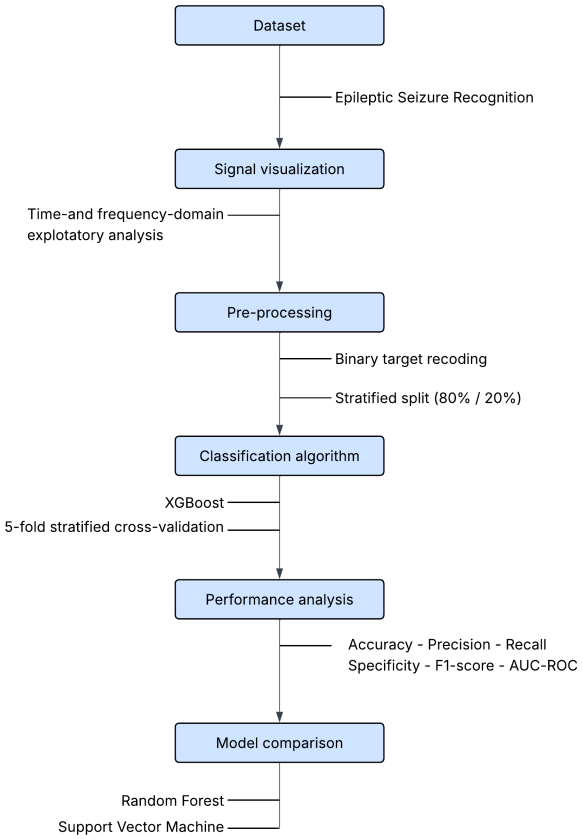

Fig. 1. Methodological pipeline of the proposed epileptic EEG classification system

In a complementary line of work, Shimanto [17] showed that classical machine-learning algorithms

operating on raw temporal EEG samples can achieve high classification performance without explicit feature engineering.

Despite these advances, comparatively few studies simultaneously address raw temporal EEG representation, explicit class-imbalance handling, cross-validation (CV) stability, computational efficiency, and a clinically grounded analysis of false negatives within a single framework.

Unlike deep-learning approaches that often require substantial computational resources and yield limited interpretability, tree-based gradient-boosting methods such as XGBoost [5] provide built-in regularization, efficient parallel computation, and interpretable feature-importance estimates, which make them particularly attractive for clinical deployment.

In this context, the present study investigates whether an XGBoost-based pipeline operating directly on the raw temporal samples of the Bonn EEG dataset, combined with explicit compensation of the natural class imbalance through the `scale_pos_weight` parameter, can deliver highly accurate, statistically stable, and clinically interpretable classification of epileptic versus normal EEG segments. The proposed classifier is benchmarked against RF and SVM under identical experimental conditions, and its residual errors are characterized statistically to assess their clinical implications.

Together, these analyses aim to position XGBoost as a robust, computationally affordable baseline for automated EEG screening in resource-limited clinical environments.

2 Materials and Methods

The methodology was structured into six sequential stages, as illustrated in Fig. 1 (1) dataset acquisition, (2) signal visualization in the time and frequency domains, (3) preprocessing, (4) classification with XGBoost, (5) performance analysis, and (6) benchmarking against RF and SVM. Each stage is detailed in the following subsections.

2.1 Dataset Acquisition

The Epileptic Seizure Recognition dataset [17] used in this study is a publicly available reformulation of the EEG dataset originally collected by Andrzejak et al. [2] at the Department of Epileptology of the University of Bonn, Germany. The Bonn dataset has become one of the most widely used benchmarks for automated seizure detection, as it provides expert-curated recordings, well-defined clinical states, and balanced acquisition conditions, enabling direct comparison with an extensive body of prior work. The original Bonn dataset comprises five subsets, denoted A–E, each containing 100 single-channel EEG segments of 23.6 seconds sampled at 173.61 Hz and bandpass-filtered between 0.53 and 40 Hz. Subsets A and B correspond to surface EEG recordings from healthy volunteers with eyes open and eyes closed, respectively; subsets C and D contain intracranial recordings from the hippocampal formation of the hemisphere opposite to the epileptogenic zone and from the epileptogenic zone itself, both during interictal (seizure-free) periods; finally, subset E comprises intracranial recordings obtained during ictal activity. For the present study, the version of the dataset hosted on Kaggle and originally derived from the UCI ML Repository [22] was adopted. In this version, each 23.6-second recording is segmented into 23 non-overlapping one-second windows, yielding a total of $500 \times 23 = 11{,}500$ samples, each described by 178 amplitude values sampled at approximately 178 Hz.

The Kaggle distribution provides five class labels (1–5): class 1 corresponds to recordings of seizure activity, while classes 2, 3, 4, and 5 correspond to four non-seizure conditions—EEG recorded over a tumoral region, EEG recorded over a healthy region distant from the tumor, EEG with eyes closed, and EEG with eyes open, respectively. Following the binary classification protocol most commonly adopted in the literature [17], [19] class 1 was retained as the positive class (epileptic), and classes 2–5 were merged into a single negative class (normal). This formulation reproduces the canonical seizure-versus-non-seizure task and induces a natural class imbalance of 4:1 between normal

and epileptic segments. The main characteristics of the dataset are summarized in Table 1.

2.2 Signal Visualization

Prior to modeling, an exploratory characterization of the dataset was performed through four complementary visualizations that compare the normal and epileptic classes in both the time and frequency domains. The time-domain plot, the spectrogram, and the frequency spectrum were generated from two representative one-second segments—one randomly drawn from each class—while the descriptive statistics were computed as class averages over the full dataset (9,200 normal and 2,300 epileptic segments).

All analyses were implemented in Python using the `SciPy` [20] signal-processing module:

(a) Time-domain analysis. Each representative segment was plotted as a one-second time series. Local maxima were detected automatically using the `scipy.signal.find_peaks` function. The detection threshold was set at twice the mean absolute amplitude of the corresponding segment. This procedure provide a qualitative indicator of the high-amplitude transient discharges typically associated with ictal activity.

(b) Spectrogram. The time–frequency representation of each representative segment was computed via the Short-Time Fourier Transform with a Hann window of 32 samples ($\approx$ 180 ms) and 75% overlap. The resulting power spectral density $S_{xx}(t, f)$ was expressed in decibels as $10 \log_{10}(S_{xx}+10^{-12})$, where the small additive constant ensured numerical stability at low-energy bins. The spectrogram was visualized in the 0–60 Hz range, which encompasses the canonical EEG bands of clinical interest.

(c) Frequency spectrum. The magnitude of the Fast Fourier Transform (FFT) was computed for each representative segment and displayed on a semi-logarithmic scale in the 0–60 Hz range, with the canonical EEG bands—delta (δ, 0.5–4 Hz), theta (θ, 4–8 Hz), alpha (α, 8–13 Hz), and beta (β, 13–30 Hz)—highlighted as shaded regions to facilitate the comparison of spectral energy distribution between classes.

(d) Descriptive statistics. In contrast to the previous visualizations, the mean amplitude, standard deviation, Fisher kurtosis, and skewness were computed segment-by-segment across the entire dataset and then averaged within each class, yielding four class-level descriptors that summarize the overall statistical behavior of the two populations. These descriptors were min–max normalized to the interval $[0, 1]$ for comparative visualization only.

2.3 Preprocessing

The target variable was recoded as a binary label following the protocol described in Section 2.1 segments belonging to class 1 of the original dataset (ictal activity) were assigned to the positive class (*epileptic*, $y = 1$), while the remaining four classes (2–5), corresponding to non-ictal conditions, were merged into a single negative class (*normal*, $y = 0$). No feature scaling or normalization was applied to the 178 amplitude samples, since gradient-boosted decision trees are invariant to monotonic transformations of the input features and therefore do not benefit from such preprocessing.

The dataset was subsequently partitioned into a training subset (80%, 9,200 segments) and a held-out test subset (20%, 2,300 segments) by means of stratified random sampling implemented with the `train_test_split` function of `scikit-learn` [13]. Stratification ensured that the original 4:1 class proportion was preserved in both subsets.

2.4 Classification Algorithm

The XGBoost algorithm [5] was selected as the core classifier. It builds an additive ensemble of decision trees in a stage-wise manner: at each iteration t, a new tree f_t is added to correct the residual error of the current ensemble.

The parameters of f_t are obtained by minimizing the following regularized objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell\Big(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\Big) + \gamma T + \tfrac{\lambda}{2} \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} |w_j|, \quad (1)$$

where n is the number of training samples, ℓ denotes the binary logistic (cross-entropy) loss, $\hat{y}_i^{(t-1)}$ is the prediction of the previous ensemble for sample i, and $f_t(x_i)$ is the contribution of the new tree. The regularization term penalizes model complexity through three coefficients: γ controls the cost of adding a new leaf (where T is the total number of leaves and w_j their associated output weights), while λ and α act as L2 and L1 regularization coefficients on the leaf weights, respectively.

The natural 4:1 class imbalance of the training set was compensated through the `scale_pos_weight` parameter of the XGBoost classifier, defined as the ratio between the number of negative and positive instances:

$$\texttt{scale_pos_weight} = \frac{N_{\text{neg}}}{N_{\text{pos}}} = \frac{9{,}200}{2{,}300} = 4.0. \quad (2)$$

This adjustment increases the minority (epileptic) class's contribution to gradient updates, allowing the model to learn discriminative patterns despite the imbalance without altering the original sample distribution.

The hyperparameter configuration was initially explored using a Grid Search over the main XGBoost parameters, including the number of estimators, learning rate, maximum tree depth, subsampling ratios, and regularization coefficients. The evaluated combinations produced only marginal differences in validation performance, indicating that the classifier was relatively robust to moderate variations in these parameters. However, the class imbalance in the dataset (4:1 ratio of normal to epileptic segments) had a more substantial impact on model behavior and classification performance.

For this reason, particular attention was given to the `scale_pos_weight` parameter, which was set to 4.0 based on the ratio of negative to positive samples.

This adjustment increases the contribution of epileptic segments during training, reducing the model's tendency to favor the majority class and improving the balance between sensitivity and specificity. The remaining hyperparameters were selected to provide an adequate trade-off between predictive performance, model complexity, and computational efficiency. Specifically, a moderate tree depth (*max_depth* = 4) and L1/L2 regularization (*reg_alpha* = 0.1, *reg_lambda* = 1.0) were adopted to limit overfitting, while *subsample* = 0.8 and *colsample_bytree* = 0.8 introduce additional randomness that improves generalization. Likewise, the combination of *learning_rate* = 0.05 and *n_estimators* = 300 enables gradual ensemble optimization while maintaining stable performance across CV folds.

The complete configuration, together with the rationale behind each value, is summarized in Table 2.

Table 2. Hyperparameters used to train the XGBoost classifier

Hyperparameter	Value	Rationale
`n_estimators`	300	Gradual convergence of the ensemble
`learning_rate`	0.05	Conservative shrinkage for better generalization
`max_depth`	4	Limits individual tree complexity
`min_child_weight`	5	Minimum sample weight per leaf
`subsample`	0.8	Row subsampling to reduce variance
`colsample_bytree`	0.8	Feature subsampling per tree
`reg_alpha`	0.1	L1 regularization on leaf weights
`reg_lambda`	1.0	L2 regularization on leaf weights
`scale_pos_weight`	4.0	Compensation for the 4:1 class imbalance

2.5 Performance Analysis

A common set of complementary metrics was computed in order to characterize different aspects of classifier behavior, since a single aggregate score is rarely sufficient to describe performance under class imbalance.

The starting point of the evaluation is the confusion matrix, which summarizes the test-set predictions in terms of four quantities: true positives (TP, epileptic segments correctly classified), true negatives (TN, normal segments correctly classified), false positives (FP, normal segments labeled as epileptic), and false negatives (FN, epileptic segments missed by the model). All the metrics described below are derived from these four entries.

Sensitivity, also known as recall or True Positive Rate (TPR), measures the fraction of epileptic segments correctly detected by the model, whereas specificity, or True Negative Rate (TNR), measures the fraction of normal segments correctly retained:

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \qquad \text{Specificity} = \frac{TN}{TN+FP}. \quad (3)$$

These two metrics are inherently complementary: raising the decision threshold typically increases specificity at the expense of sensitivity, and vice versa. Analyzing them jointly therefore provides a more faithful picture of the trade-off between the two types of classification errors than either metric alone.

Accuracy measures the overall proportion of correctly classified segments, while precision quantifies the reliability of the positive predictions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \qquad \text{Precision} = \frac{TP}{TP+FP}. \quad (4)$$

Although accuracy is straightforward to interpret, its informativeness is limited under class imbalance: a trivial classifier that always predicts the majority class would achieve 80% accuracy on this dataset without any discriminative ability. Accuracy was therefore interpreted jointly with the remaining metrics rather than in isolation.

The F1-score is the harmonic mean of Precision and Sensitivity:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} = \frac{2TP}{2TP+FP+FN}. \quad (5)$$

Because the harmonic mean is dominated by the lower of its two arguments, a high F1-score requires both Precision and Sensitivity to be high, making it especially informative under class imbalance.

Finally, the Receiver Operating Characteristic (ROC) curve is obtained by plotting sensitivity against the False Positive Rate (FPR $= FP/(FP+TN) = 1 -$ Specificity) across all possible classification thresholds, and the area under this curve (AUC-ROC) condenses the overall discriminative ability of the classifier into a single scalar between 0.5 (random) and 1.0 (perfect) [7]:

$$\text{AUC-ROC} = \int_0^1 \text{Sensitivity}\big(\text{FPR}\big)\, d(\text{FPR}). \quad (6)$$

AUC-ROC was adopted as the primary evaluation metric because it is threshold-independent and assesses the model's ability to rank positive instances above negative ones, making it particularly robust to class imbalance.

2.6 Model Comparison

To contextualize the performance of XGBoost, two widely used reference algorithms for EEG-based classification were trained and evaluated under the same experimental protocol: RF [3] and SVM with a radial basis function (RBF) kernel.

The RF model was configured with 300 trees and a maximum depth of 10, while the SVM was trained with a regularization parameter $C = 1.0$ and the default scaled bandwidth γ provided by `scikit-learn` [13].

Both reference models were trained with class-weight balancing to compensate for the 4:1 class imbalance.

Feature standardization was applied exclusively to the SVM via a pipeline that coupled a standard scaler with the classifier, since gradient-boosted decision trees and RF are invariant to monotonic transformations of the input features.

Finally, to gain further insight into the nature of the misclassifications produced by XGBoost, a Welch's t-test ($\alpha = 0.05$) was conducted on the test set to determine whether the mean absolute amplitude of false-negative segments differed significantly from that of correctly classified epileptic segments (TP).

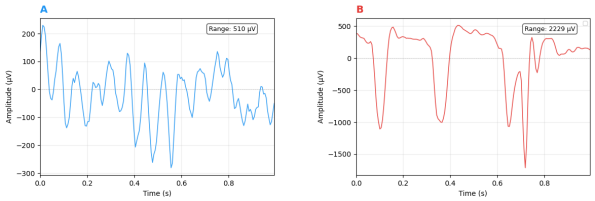

Fig. 2. Representative one-second EEG segments in the time domain. (A) Normal segment, peak-to-peak range $510\ \mu$V. (B) Epileptic segment, peak-to-peak range $2{,}229\ \mu$V

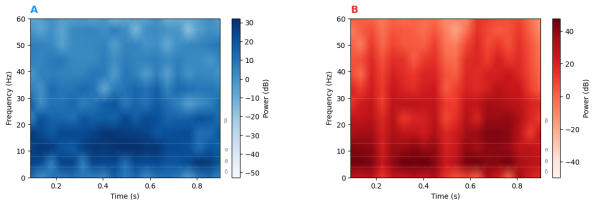

Fig. 3. Spectrograms of the same representative EEG segments shown in Fig. 2, computed via the Short-Time Fourier Transform with a Hann window of 32 samples and 75% overlap. (A) Normal segment. (B) Epileptic segment. Color encodes power in dB; horizontal dotted lines delimit the canonical EEG bands (δ, θ, α, β)

3 Results

The experimental evaluation of the proposed methodology is reported below. The exploratory characterization of normal and epileptic EEG segments is presented first, followed by the predictive performance of the XGBoost classifier on the held-out test partition and across the 5-fold stratified CV. A comparative evaluation against RF and SVM, the model's learning behavior, and a clinical interpretation of the misclassified segments are subsequently presented.

3.1 Signal Visualization and Characterization

An exploratory characterization of the dataset was performed to compare the behavior of normal and epileptic EEG signals in both the time and frequency domains. This analysis provides a qualitative and quantitative understanding of the discriminative information present in the raw temporal representations used as input to the classification stage.

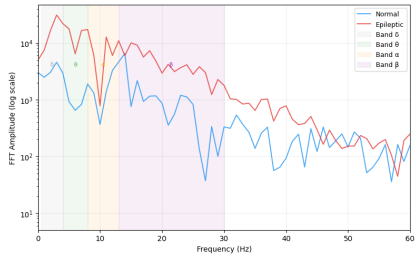

Fig. 4. FFT magnitude spectrum (logarithmic scale) of the same representative segments shown in Fig. 2, in the 0–60 Hz range. The canonical EEG bands (δ, θ, α, β) are highlighted as shaded regions. The epileptic spectrum (red) lies consistently above the normal spectrum (blue) across all bands, with the largest separation observed in the low-frequency δ–θ range

The contrast between the two classes is already evident in the raw time-domain waveforms (Fig. 1). The epileptic segment exhibits amplitude excursions more than four times those of the normal segment and contains several high-amplitude transient discharges that have no counterpart in the normal recording, consistent with the hypersynchronous neuronal activity typically observed during ictal events.

The time–frequency analysis confirms and refines these observations (Fig. 3). The normal segment concentrates most of its energy in the alpha–beta range (8–30 Hz), with a stable and uniform power distribution along the one-second window.

In contrast, the epileptic segment exhibits substantially higher spectral power across the entire 0–60 Hz band, with a particularly strong contribution in the low-frequency delta–theta range (0.5–8 Hz) and a markedly more transient temporal structure, in agreement with the neurophysiological signature of ictal activity.

The frequency-domain analysis (Fig. 4) provides a complementary view of the spectral differences between the two classes. Across the entire 0–60 Hz range, the epileptic spectrum lies consistently above the normal one, reflecting a generalized increase in spectral amplitude rather than a band-specific phenomenon.

The largest separation between classes is observed in the delta (δ, 0.5–4 Hz) and theta (θ, 4–8 Hz) bands, where the epileptic amplitude reaches values up to one order of magnitude greater than the normal one.

The alpha (α, 8–13 Hz) and beta (β, 13–30 Hz) bands also show a clear separation, although less pronounced than at low frequencies. At frequencies above 30 Hz, both spectra decay progressively and become noisier, while still preserving the relative ordering between classes.

Table 3. Performance of the XGBoost classifier on the training and test sets

Metric	Training	Test
Accuracy	0.9952	0.9757
Precision	0.9777	0.9430
Recall	0.9989	0.9348
Specificity	0.9993	0.9859
F1-score	0.9882	0.9389
AUC-ROC	0.9999	0.9953

Table 4. 5-fold stratified CV of the XGBoost classifier on the training set. Values are reported as the mean $\pm$ standard deviation across folds for both the training and test partitions of each fold

Metric	Train ($\mu \pm \sigma$)	Test ($\mu \pm \sigma$)	Δ
Accuracy	0.9953 ± 0.0003	0.9740 ± 0.0019	0.0213
Precision	0.9781 ± 0.0015	0.9364 ± 0.0071	0.0417
Sensitivity	0.9987 ± 0.0007	0.9335 ± 0.0076	0.0652
Specificity	0.9944 ± 0.0004	0.9841 ± 0.0019	0.0103
F1-score	0.9883 ± 0.0008	0.9349 ± 0.0047	0.0534
AUC-ROC	0.9999 ± 0.0000	0.9951 ± 0.0009	0.0048

3.2 XGBoost Model Performance

Table 3 reports the performance of the XGBoost classifier on the 80%/20% training–test partition. The model achieves an AUC-ROC of 0.9953 on the held-out test set, with a train–test gap of only 0.0046, indicating excellent generalization and no evidence of overfitting. Test accuracy reaches 97.57%, and the remaining metrics confirm a strong predictive performance on both classes: precision, recall, and F1-score for the normal class are 0.9853, 0.9843, and 0.9848, respectively, while for the epileptic class they are 0.9430, 0.9348, and 0.9389. The slightly lower scores on the epileptic class reflect the natural difficulty of detecting the minority class under a 4:1 imbalance, although the high recall (93.48%) indicates that the vast majority of seizure segments are correctly identified.

The discriminative behavior of the classifier is summarized by the ROC curve shown in Fig. 5. The curve approaches the upper-left corner over almost its entire range, reflecting an extremely high separation between the two classes across all possible decision thresholds. The optimal operating point, defined as the point on the test curve closest to $(\text{FPR}, \text{TPR}) = (0, 1)$, is located at $\text{FPR} = 0.037$ and $\text{TPR} = 0.974$. At this operating point, the classifier detects 97.4% of the epileptic segments at the cost of only 3.7% false alarms, which constitutes a clinically reasonable trade-off for continuous EEG screening.

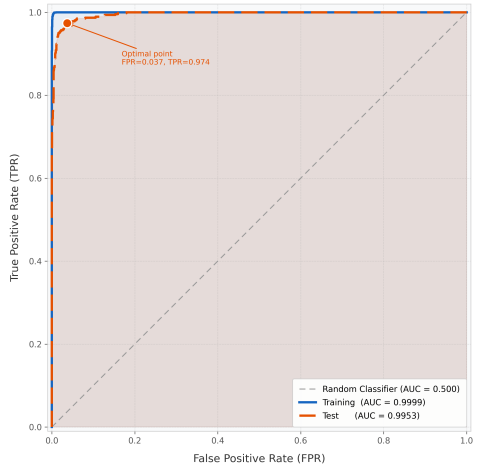

Fig. 5. ROC curves of the XGBoost classifier on the training (blue, $\text{AUC} = 0.9999$) and test (orange, $\text{AUC} = 0.9953$) sets, compared to a random classifier (gray dashed line, $\text{AUC} = 0.500$). The orange marker indicates the optimal operating point on the test curve, located at $\text{FPR} = 0.037$ and $\text{TPR} = 0.974$

3.3 Cross-Validation

To assess the stability and generalization capability of the XGBoost classifier, a 5-fold stratified CV

was performed on the training set, preserving the original class proportion in each fold. The model achieved a mean test AUC-ROC of 0.9951 ± 0.0009 and a mean test accuracy of 0.9740 ± 0.0019, with all the remaining metrics summarized in Table 4. The standard deviations remain below 0.008 for every metric, confirming that the classifier behaves consistently across all folds. The train–test gaps are small for the threshold-independent AUC-ROC (0.0048) and for accuracy and specificity (≤ 0.022), and somewhat larger for threshold-dependent metrics such as sensitivity (0.0652) and F1-score (0.0534), which is the expected behavior of an ensemble model that fits the training folds nearly perfectly while still generalizing well to unseen segments.

3.4 Model Comparison

To contextualize XGBoost's performance, a comparative evaluation was conducted against RF and SVM under identical experimental conditions, using 5-fold stratified CV. The mean and standard deviation of the validation scores across folds for the three classifiers are reported in Table 5 and visually summarized in Fig. 6.

The three classifiers exhibit very similar discriminative ability, with AUC-ROC values within a narrow range (0.9946–0.9951) and overlapping standard deviations, indicating that none of the models dominates the others in pure ranking performance.

Table 5. Model comparison based on 5-fold stratified CV. Values are reported as the mean $\pm$ standard deviation of the validation scores across folds. The last row reports the average training time per fold

Metric	XGBoost	RF	SVM
Accuracy	0.9740 ± 0.0019	0.9671 ± 0.0026	0.9749 ± 0.0021
Precision	0.9364 ± 0.0071	0.9581 ± 0.0069	0.9188 ± 0.0071
Sensitivity	0.9335 ± 0.0076	0.8739 ± 0.0106	0.9591 ± 0.0056
Specificity	0.9841 ± 0.0019	0.9904 ± 0.0016	0.9788 ± 0.0020
F1-score	0.9349 ± 0.0047	0.9140 ± 0.0070	0.9385 ± 0.0051
AUC-ROC	0.9951 ± 0.0009	0.9950 ± 0.0008	0.9946 ± 0.0008
Training time (s)	≈ 11	≈ 25	≈ 38

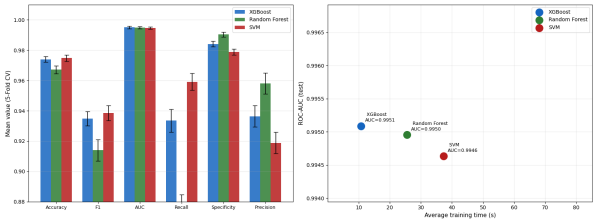

Fig. 6. Comparison of XGBoost, RF, and SVM under 5-fold stratified CV. Left: mean validation scores for the six classification metrics, with error bars representing one standard deviation across folds. Right: performance–efficiency trade-off, where each model is placed according to its mean test AUC-ROC and its average training time per fold; XGBoost occupies the most favorable position, with the highest AUC-ROC and the lowest training time

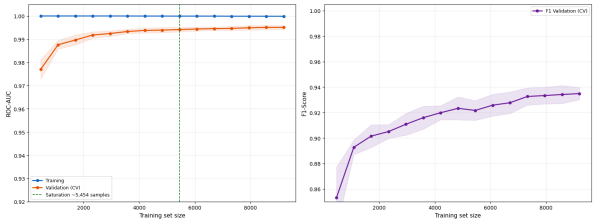

Fig. 7. Learning curves of the XGBoost classifier as a function of the training set size, evaluated through 5-fold stratified CV. Left: ROC-AUC for the training (blue) and test (orange) partitions; the green dashed line marks the saturation point at approximately $5{,}454$ training samples. Right: F1-score on the epileptic class along the validation folds. Shaded bands represent ± 1 standard deviation across folds

Their behavior differs more clearly in the sensitivity–specificity trade-off. RF reaches the highest specificity (0.9904) and precision (0.9581) but the lowest sensitivity (0.8739), behaving as a conservative classifier that minimizes false alarms at the cost of missing a larger fraction of epileptic segments.

SVM exhibits the opposite trend, with the highest sensitivity (0.9591) and accuracy (0.9749) but the lowest specificity and Precision, leading to a higher rate of FP.

XGBoost positions itself between the two extremes, with sensitivity and specificity simultaneously above 0.93 and 0.98, respectively,

and an F1-score (0.9349) only marginally lower than that of SVM (0.9385).

The clearest advantage of XGBoost is observed in the performance–efficiency trade-off shown in the right panel of Fig. 6: with an average training time of approximately 11 seconds per fold, it is roughly $2.3\times$ faster than RF ($\sim$ 25 s) and $3.5\times$ faster than SVM ($\sim$ 38 s), while preserving comparable or superior discriminative performance.

3.5 Learning Curves

To assess the sample efficiency of the XGBoost model, learning curves were computed over a range of training sizes spanning 5%–100% of the available data, with performance evaluated through 5-fold stratified CV at each size. As shown in Fig. 7, the validation AUC-ROC increases rapidly during the first half of the curve and reaches a plateau at approximately $5{,}454$ training samples ($\sim 47\%$ of the dataset), beyond which additional data produces only marginal improvements. The corresponding F1-score on the minority class follows the same trend. This behavior indicates that the proposed classifier extracts most of its discriminative power from a relatively modest amount of training data, a property of practical interest for clinical scenarios in which expert-annotated EEG segments are scarce.

3.6 Error Analysis

The relative weight of these errors can be further modulated by the model's decision threshold (Table 6). At the default threshold, the classifier achieves a sensitivity of 93.5% with an FPR of only 1.4%.

Table 6. Prediction distribution of the XGBoost classifier on the held-out test set ($n = 2{,}300$ segments)

Group	Count	% of class	Clinical implication
TP	430	93.5%	Seizures correctly detected
TN	1,815	98.6%	Normal segments correctly retained
FN	30	6.5%	Seizures missed by the model
FP	25	1.4%	Normal segments incorrectly flagged

Shifting the operating point to the optimum identified on the ROC curve (FPR $= 0.037$, TPR $= 0.975$; Fig. 5) raises recall to 97.4%, which reduces the number of undetected seizures from 30 to approximately 12 ($\sim$ 60% reduction) at the cost of an additional ~ 43 false alarms (from 25 to ~ 68 out of $1{,}840$ normal segments).

This trade-off is generally acceptable in supervised inpatient EEG monitoring, where every flagged event is reviewed by a clinical expert before any therapeutic action is taken.

Table 7. Comparison with prior works on the Bonn EEG dataset

Reference	Method	Features	Accuracy (%)
Subasi (2007) [18]	SVM + Wavelets	DWT coefficients	88–95
Acharya et al. (2013) [1]	ANN + nonlinear	Entropy, DFA	95.0
Ullah et al. (2018) [19]	P-1D-CNN	Raw samples	99.6
Kumar & Kanhangad (2019) [11]	RF	Wavelet + stats	97.2
Shimanto (2021) [17]	ML benchmark	Raw temporal	≤ 97.0
This work	**XGBoost**	**Raw (178)**	**97.57**

4 Discussion and Conclusions

The XGBoost model achieved an AUC-ROC of 0.9953 and an accuracy of 97.57% on the held-out test set, with a train–test gap below 0.005 in AUC-ROC and a mean CV AUC-ROC of 0.9951 ± 0.0009. This combination of high discriminative power, stability across folds, and a small generalization gap indicates that the model captures the underlying ictal patterns rather than memorizing the training data, in agreement with the L1/L2 regularization and the ensemble averaging that characterize gradient-boosted trees.

Compared with previous studies on the Bonn EEG dataset (Table 7), the proposed XGBoost classifier achieves higher accuracy than classical pipelines based on hand-crafted features. Subasi [18] reported accuracies between 88% and 95% using SVM combined with wavelet coefficients, while Acharya et al. [1] achieved 95% using ANNs fed with nonlinear descriptors. Kumar and Kanhangad [11] reached 97.2% accuracy with RF and wavelet-derived statistics. The 97.57% accuracy obtained in this work confirms that an ensemble of regularized decision trees, operating directly on the 178 raw amplitude samples, can match or slightly outperform these approaches without requiring any manual feature engineering.

DL architectures still hold the upper bound on this benchmark: the P-1D-CNN model proposed

by Ullah et al. [19] reaches 99.6% accuracy on the same dataset. This approximately 2-percentage-point gap reflects the ability of deep convolutional networks to automatically learn hierarchical representations from raw temporal data. However, the XGBoost classifier remains attractive in resource-constrained scenarios: it requires roughly 11 seconds of training per fold against the substantially longer training times typically reported for deep networks, exposes interpretable feature-importance scores, and can be deployed on standard CPU hardware. In the within-paper comparison, this efficiency advantage is also observed against the classical baselines: XGBoost is approximately $2.3\times$ faster than RF and $3.5\times$ faster than SVM under the same CV protocol, while obtaining a comparable AUC-ROC (0.9946–0.9951 across the three models) and the most balanced sensitivity–specificity profile of the three.

The error analysis provides additional context about the remaining classification errors. The 30 FN observed on the test set exhibit a mean absolute amplitude of 81.02 ± 12.91 μV, approximately $3.2\times$ lower than the 259.29 ± 118.50 μV observed in the correctly classified epileptic segments. The Welch's t-test confirms that this difference is highly significant ($t = -28.74$, $p < 10^{-99}$), suggesting that the model fails systematically on low-energy ictal events—compatible with focal or subclinical seizures—rather than on canonical high-amplitude discharges. Moreover, shifting the operating point to the optimum on the ROC curve (FPR $= 0.037$, TPR $= 0.974$) reduces the number of undetected seizures from 30 to approximately 12, at the cost of ~ 43 additional false alarms.

Taken together, these results support gradient-boosted trees as a robust, computationally efficient, and interpretable baseline for automated EEG-based seizure detection. By operating directly on raw temporal amplitudes and explicitly compensating for the class imbalance inherent to clinical EEG datasets, the proposed pipeline avoids the manual feature-engineering stage that has traditionally limited the reproducibility and portability of classical approaches.

Moreover, the residual errors observed in this study correspond predominantly to low-amplitude ictal events with limited cortical manifestation, which are precisely the cases in which the assistance of an automated screening system is most valuable, since they tend to be the most challenging for visual inspection as well. The combination of high discriminative power, modest computational cost, and the ability to adjust the decision threshold to favor sensitivity makes the proposed classifier well-suited as a first-line decision-support tool that can assist clinical neurophysiologists in screening long EEG recordings, especially in resource-limited settings where specialist availability is scarce. Rather than replacing expert interpretation, such a system can help to prioritize segments of clinical interest and reduce the manual workload associated with continuous EEG monitoring, contributing to faster and more equitable access to epilepsy diagnosis.

Beyond its direct clinical application, the proposed methodology is also relevant to intelligent systems and emerging healthcare robotics applications. Automated EEG classification is a key component in brain–computer interfaces (BCIs), neuroprosthetic devices, and assistive robotic systems that require real-time interpretation of neural activity [14]. In these contexts, the high classification performance and computational efficiency of the proposed XGBoost-based model could facilitate its integration into embedded decision-support systems and future human–robot interaction frameworks for neurological monitoring and adaptive assistance.

Three main limitations of the present study should be acknowledged. First, the hyperparameters of the XGBoost classifier were selected manually following standard recommendations for imbalanced tabular classification, rather than through a systematic optimization procedure. Second, the experimental evaluation is restricted to the Bonn dataset, which contains a limited number of subjects and acquisition conditions that are more controlled than those encountered in real clinical EEG monitoring. Third, the class-imbalance compensation relied exclusively on the `scale_pos_weight` parameter, and resampling strategies such as SMOTE or

ADASYN were not explored. These limitations directly motivate the following lines of future work: a systematic hyperparameter optimization via Grid Search or Bayesian optimization; the inclusion of frequency-domain features derived from the Discrete Wavelet Transform to complement the raw temporal samples; external validation on larger and more clinically diverse corpora such as CHB-MIT and the Temple University EEG corpus; the comparison with recent deep-learning architectures (1D-CNN and LSTM) under the same evaluation protocol; the evaluation of oversampling techniques (SMOTE, ADASYN) for handling class imbalance; and the integration of the trained model into a real-time clinical interface with adjustable decision thresholds for ambulatory EEG monitoring.

References

1. **Acharya, U. R., Sree, S. V., Swapna, G., Martis, R. J., Suri, J. S. (2013).** Automated EEG analysis of epilepsy: A review. Knowledge-Based Systems, Vol. 45, pp. 147–165.
2. **Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., Elger, C. E. (2001).** Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Physical Review E, Vol. 64, No. 6, pp. 061907.
3. **Breiman, L. (2001).** Random forests. Machine Learning, Vol. 45, No. 1, pp. 5–32.
4. **Burneo, J. G., Tellez-Zenteno, J., Wiebe, S. (2005).** Understanding the burden of epilepsy in Latin America: a systematic review of its prevalence and incidence. Epilepsy Research, Vol. 66, No. 1–3, pp. 63–74.
5. **Chen, T., Guestrin, C. (2016).** XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 785–794.
6. **Devinsky, O., Hesdorffer, D. C., Thurman, D. J., Lhatoo, S., Richerson, G. (2016).** Sudden unexpected death in epilepsy: epidemiology, mechanisms, and prevention. The Lancet Neurology, Vol. 15, No. 10, pp. 1075–1088.
7. **Fawcett, T. (2006).** An introduction to ROC analysis. Pattern Recognition Letters, Vol. 27, No. 8, pp. 861–874.
8. **Fisher, R. S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J. H., Elger, C. E., Engel, J., Forsgren, L., French, J. A., Glynn, M., Hesdorffer, D. C., Lee, B. I., Mathern, G. W., Mosh'e, S. L., Perucca, E., Scheffer, I. E., Tomson, T., Watanabe, M., Wiebe, S. (2014).** ILAE official report: a practical clinical definition of epilepsy. Epilepsia, Vol. 55, No. 4, pp. 475–482.
9. **Fisher, R. S., Cross, J. H., D'Souza, C., French, J. A., Haut, S. R., Higurashi, N., Hirsch, E., Jansen, F. E., Lagae, L., Mosh'e, S. L., Peltola, J., Roulet Perez, E., Scheffer, I. E., Schulze-Bonhage, A., Somerville, E., Sperling, M., Yacubian, E. M., Zuberi, S. M. (2017).** Instruction manual for the ILAE 2017 operational classification of seizure types. Epilepsia, Vol. 58, No. 4, pp. 531–542.
10. **Grant, A. C., Abdel-Baki, S. G., Weedon, J., Arnedo, V., Chari, G., Koziorynska, E., Lushbough, C., Maus, D., McSween, T., Mortati, K. A., et al. (2014).** Eeg interpretation reliability and interpreter confidence: a large single-center study. Epilepsy & Behavior, Vol. 32, pp. 102–107.
11. **Kumar, T. S., Kanhangad, V. (2019).** EEG-based automated detection of epilepsy using machine learning approach. Biocybernetics and Biomedical Engineering, Vol. 39, No. 3, pp. 702–709.
12. **Niedermeyer, E., da Silva, F. L. (2004).** Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. Lippincott Williams & Wilkins, 5th edition.
13. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel,**

M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, Vol. 12, pp. 2825–2830.

14. **Rashid, M., Sulaiman, N., PP Abdul Majeed, A., Musa, R. M., Ab Nasir, A. F., Bari, B. S., Khatun, S. (2020).** Current status, challenges, and possible solutions of eeg-based brain-computer interface: a comprehensive review. Frontiers in neurorobotics, Vol. 14, pp. 515104.
15. **Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., Faubert, J. (2019).** Deep learning-based electroencephalography analysis: a systematic review. Journal of neural engineering, Vol. 16, No. 5, pp. 051001.
16. **Sharma, R., Meena, H. K. (2024).** Emerging trends in eeg signal processing: A systematic review. SN Computer Science, Vol. 5, No. 4, pp. 415.
17. **Shimanto, H. (2021).** Machine learning algorithms for epileptic seizures. `www.kaggle.com/code/harunshimanto/machine-learning-algorithms-for-epileptic-seizures`. Kaggle Notebook. Accessed: March 2025.
18. **Subasi, A. (2007).** EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Systems with Applications, Vol. 32, No. 4, pp. 1084–1093.
19. **Ullah, I., Hussain, M., Qazi, E.-u.-H., Aboalsamh, H. (2018).** An automated system for epilepsy detection using EEG brain signals based on deep learning approach. Expert Systems with Applications, Vol. 107, pp. 61–71.
20. **Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020).** Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods, Vol. 17, No. 3, pp. 261–272.
21. **World Health Organization (2023).** Epilepsia: datos y cifras. `https://www.who.int/es/news-room/fact-sheets/detail/epilepsy`. Accessed: March 2025.
22. **Wu, Q., Fokoue, E. (2017).** Epileptic seizure recognition data set. DOI: `10.24432/C5G308`.

Article received on 13/03/2026; accepted on 18/07/2026.
**Corresponding author is Valeria Maeda-Gutiérrez.*