Received: 29 June 2026 | Revised: 20 September 2026 | Accepted: 29 September 2026 | Published online: 20 October 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–11
DOI: 10.47852/bonviewAIA620211142

BON VIEW PUBLISHING

Toward Trustworthy EEG-Based Epileptic Seizure Classification: A Hybrid Explainable AI Framework with Statistical Validation

Mohamed Al Saidi[1], Noushath Shaffi[1,*], Vimbi Viswan[2], Abdelhamid Abdesselam[1] and Rachid Hedjam[3,*]

[1] *Department of Computer Science, College of Science, Sultan Qaboos University, Oman*

[2] *Department of Computing and Information Science, University of Technology and Applied Sciences - Sohar, Oman*

[3] *Department of Computer Science, Bishop's University, Canada*

Abstract: Automated classification of epileptic seizure activity from electroencephalography (EEG) signals remains challenging because of the nonstationary nature of EEG and the need for reliable and interpretable clinical decision-support systems. Although recent machine learning and deep learning approaches have reported promising performance, many provide limited statistical validation and model explainability. This study presents a methodologically integrated hybrid artificial intelligence framework for binary EEG seizure classification using 591 ictal and non-seizure segments from the patient-nonspecific Meta-EEG derivative of the CHB-MIT dataset. The framework combines a one-dimensional convolutional neural network (CNN), trained on standardized multichannel EEG signals, with an XGBoost classifier trained on statistical features extracted using level-4 db4 wavelet packet decomposition (WPD). The branch probabilities were combined using validation-selected equal-weight fusion, with a weight of 0.5 assigned to each branch. SHapley Additive exPlanations and Integrated Gradients were used to interpret the XGBoost-WPD and CNN branches, respectively. Statistical evaluation included permutation testing, bootstrap confidence intervals, exact binomial testing, and paired McNemar comparisons. On the held-out test set, the hybrid model achieved 94.38% accuracy, 93.31% balanced accuracy, 90.00% sensitivity, 96.61% specificity, 91.53% F1-score, and 98.19% Receiver Operator Characteristics- Area Under Curve (ROC-AUC). McNemar's test showed a significant difference from the standalone CNN ($p = 0.0117$), but not from XGBoost-WPD or InceptionTimeMini. These findings support the potential value of complementary temporal and time–frequency representations while highlighting the need for cross-validation and external evaluation.

Keywords: EEG, epileptic seizure classification, wavelet packet decomposition, explainable artificial intelligence, statistical validation

1. Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent seizures resulting from abnormal electrical activity in the brain [1]. Electroencephalography (EEG) is widely used for epilepsy diagnosis and monitoring because it provides a noninvasive measurement of neural activity with high temporal resolution [2]. However, manual interpretation of EEG recordings is time-consuming and requires substantial clinical expertise, motivating the development of automated EEG analysis systems [2].

Recent advances in artificial intelligence (AI) have significantly improved EEG-based seizure classification [1, 3, 4]. Traditional machine learning approaches rely on handcrafted signal features extracted from time, frequency, or time–frequency domains, whereas deep learning models learn discriminative representations directly from raw EEG signals [5, 6]. Although both paradigms have demonstrated promising results, individual representations may not fully capture the complex and heterogeneous manifestations of seizure activity present in EEG recordings.

Beyond predictive performance, reliability and interpretability have emerged as important challenges in medical AI [7]. Many existing studies primarily focus on classification accuracy while providing limited statistical evidence regarding the robustness of reported results [8]. In addition, the decision-making process of high-performing models often remains difficult to interpret, which can reduce confidence in their adoption for clinical decision support [7, 8].

This study presents a methodologically integrated hybrid AI framework for binary EEG seizure classification. The principal contribution is the systematic integration of complementary raw-signal and wavelet packet decomposition (WPD) representations, branch-specific explanations, and statistical validation under a common experimental protocol, rather than a fundamentally new neural architecture. The contributions of this study are:

***Corresponding authors:** Noushath Shaffi, Department of Computer Science, College of Science, Sultan Qaboos University, Oman and Rachid Hedjam, Department of Computer Science, Bishop's University, Canada. Email: rhedjam@ubishops.ca

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

1) Probability-level integration of a temporal one-dimensional convolutional neural network (1D-CNN) and an XGBoost classifier trained on WPD-derived descriptors.
2) Branch-specific explainability using Integrated Gradients (IG) for the convolutional neural network (CNN) and SHapley Additive exPlanations (SHAP) for XGBoost-WPD.
3) A statistical evaluation protocol combining permutation tests, bootstrap confidence intervals, exact binomial testing, and paired McNemar comparisons.
4) Controlled same-data standalone, ablation, and adapted published-method comparisons.

2. Literature Review

Automated analysis of EEG signals has attracted significant attention for supporting epilepsy diagnosis and monitoring. The availability of publicly available datasets such as the CHB-MIT Scalp EEG Database has enabled rapid development of AI algorithms for seizure classification [9]. While the traditional machine learning algorithms rely on handcrafted features extracted from time, frequency, or time–frequency domains, the deep learning approaches learn hierarchical representations. A review by Wu et al. highlighted the increased trend in the application of deep learning techniques [10].

CNN-based architectures have demonstrated strong performance in EEG seizure classification by learning discriminative temporal representations directly from signal data [11]. Subsequent studies have investigated CNN-Long Short Term Memory (LSTM) architectures, wavelet-based CNNs, and other deep learning configurations to improve seizure-classification performance [12–14]. Despite these advances, deep neural networks (DNNs) can remain difficult to interpret, motivating approaches that combine predictive modeling with complementary signal representations and explainability techniques.

To address these limitations, hybrid approaches combining handcrafted signal descriptors with data-driven learning models have received increasing attention. Time–frequency techniques such as wavelet decomposition provide explicit multiscale representations of EEG activity and can complement representations learned directly from raw signals. Hassan et al. [15], for example, combined CNN-based representations with conventional machine learning classifiers for seizure detection. In parallel, explainability methods such as SHAP [16] and IG [17] have been increasingly used to examine model predictions, while statistical procedures including permutation testing, bootstrap confidence intervals, and paired significance testing [18, 19] provide complementary evidence regarding performance uncertainty and model differences.

Recent research has increasingly investigated hybrid, multi-representation, and explainable approaches to EEG seizure analysis. Zhang et al. [20] combined feature fusion with hybrid deep learning models for seizure detection and prediction, while Sun et al. [21] used a multi-input network to learn complementary EEG representations. Amrani et al. [22] developed an explainable hybrid DNN for seizure classification and localization, and Murugan et al. [23] paired convolutional modeling with explainable AI for EEG seizure detection. Collectively, these studies show growing interest in combining heterogeneous signal representations and improving interpretability. The present work differs by integrating a raw-temporal 1D-CNN with an XGBoost-WPD branch through validation-selected probability fusion, applying branch-specific IG and SHAP, and evaluating the resulting framework with complementary statistical tests under a common held-out protocol.

Table 1 summarizes representative EEG seizure-classification studies according to their use of deep learning, hybrid modeling, explainability, and statistical validation. Although these components have individually received considerable attention, their systematic integration within a common experimental framework remains comparatively less explored. Motivated by this gap, the present study combines raw-EEG temporal modeling with WPD-derived statistical representations, branch-specific explainability, and complementary statistical validation under a unified evaluation protocol.

Table 1
A comparison of representative EEG-based seizure-classification studies

Study	DL	Hybrid	XAI	Stat. val.
Acharya et al. [11]	✓	–	–	–
Nemati et al. [12]	✓	–	–	–
Hassan et al. [15]	✓	✓	–	–
Khan et al. [24]	✓	–	✓	–
Zhang et al. [20]	✓	✓	–	–
Sun et al. [21]	✓	–	–	–
Amrani et al. [22]	✓	✓	✓	–
Murugan et al. [23]	✓	–	✓	–
Proposed framework	✓	✓	✓	✓

3. Methodology

3.1. Dataset

This study used the patient-nonspecific subset of Meta-EEG, a preprocessed and structured derivative of the CHB-MIT Scalp EEG Database [9]. Each segment was provided as a CSV file, with the class label stored in the final row. After restricting the classification task to ictal versus non-seizure activity and excluding pre-ictal, post-ictal, and peri-ictal segments, 591 segments were retained: 197 ictal and 394 non-seizure. A stratified sample-wise split with random state 42 assigned 413 segments to training, 89 to validation, and 89 to the held-out test set. The validation set was used for model selection, fusion-weight selection, and decision threshold selection, whereas the test set was reserved exclusively for final evaluation.

The Meta-EEG files were already preprocessed and formatted by the dataset provider; therefore, no additional artifact-removal or noise-filtering step was applied in this study. Before modeling, all CSV files were checked for successful parsing, two-dimensional structure, label-row consistency, missing and non-finite values, and duplicate content. All 591 retained segments passed these integrity checks. After removing the label row, each signal was standardized using channel-wise z-score normalization.

For the CNN branch, temporal length was fixed at 2,048 samples by center cropping longer segments and zero-padding shorter segments. Because the retained segments contained a variable number of EEG channels, with 38 being the maximum, the channel dimension was fixed at 38 by zero-padding segments with fewer channels. Zero-padding was used only to standardize tensor dimensions and did not introduce synthetic physiological activity. This fixed $38 \times 2{,}048$ representation is a computational input

convention for the present Meta-EEG subset and should not be interpreted as indicating that all original CHB-MIT recordings contain 38 simultaneously acquired channels.

3.2. Proposed hybrid framework

The proposed framework combines deep learning and machine learning to capture complementary representations of disease-specific EEG activity. As illustrated in Figure 1, the framework consists of two parallel branches. The temporal branch employs a 1D-CNN to learn representations directly from standardized multichannel EEG signals, while the spectral-statistical branch uses XGBoost to classify statistical descriptors extracted from WPD subbands. The probability outputs of the two branches are subsequently combined through weighted probability-level fusion to obtain the final classification. Branch-specific explainability and statistical validation are subsequently applied to assess model interpretability and performance reliability. Equal-weight probability fusion and a validation-selected threshold of 0.480 determine the final prediction, while IG and SHAP provide branch-specific interpretability.

The hybridization is motivated by representational complementarity rather than architectural complexity. The CNN learns temporal representations directly from standardized multichannel EEG, whereas the WPD-XGBoost branch models explicit multiscale time–frequency descriptors. Because the branches operate on different representations and learning paradigms, their probability estimates may provide complementary evidence for seizure classification. This hypothesis is evaluated through the standalone and ablation analyses reported in Section 4.4.

3.3. CNN branch

In the CNN branch, the model receives a standardized 38 × 2,048 multichannel EEG segment and learns temporal representations for seizure classification. The architecture comprises four 1D convolutional blocks followed by adaptive global average pooling and a two-layer fully connected classifier.

Each of the first three convolutional blocks consists of Conv1D, batch normalization, Rectified Linear Unit (ReLU) activation, max pooling, and dropout, with the number of filters increasing from 64 to 256. The fourth block uses 128 filters followed by batch normalization, ReLU activation, and adaptive global average pooling to obtain a compact representation.

The classifier applies dropout to the pooled representation, followed by a 128-to-64 fully connected layer, ReLU activation, a second dropout layer, and a 64-to-2 output layer with softmax probabilities. All convolutions use a stride of 1, while temporal down-sampling is performed through max pooling. The complete layer configuration, including kernel sizes and dropout settings, is provided in Table 2.

3.4. XGBoost-WPD branch

To capture discriminative time–frequency characteristics, each EEG segment was decomposed using WPD. Unlike the conventional wavelet decomposition, WPD recursively decomposes both approximation and detail coefficients, providing a multiscale representation of EEG frequency content. This representation complements the temporal features learned directly from the EEG signals by the CNN branch.

As illustrated in Figure 2, a level-4 db4 WPD produces 16 terminal subbands for each EEG channel. From each subband, six statistical descriptors were extracted: relative energy, mean, standard deviation, Shannon entropy, skewness, and kurtosis.

Across 38 EEG channels, this yielded 38 × 16 × 6 = 3,648 features per segment. The resulting feature vectors were used to train the XGBoost classifier [25]. The XGBoost hyperparameters were selected based on validation-set performance, with 300 estimators, a maximum tree depth of 6, and a learning rate of 0.05.

Figure 1
Overview of the proposed hybrid EEG seizure-classification framework combining temporal CNN and WPD-XGBoost branches

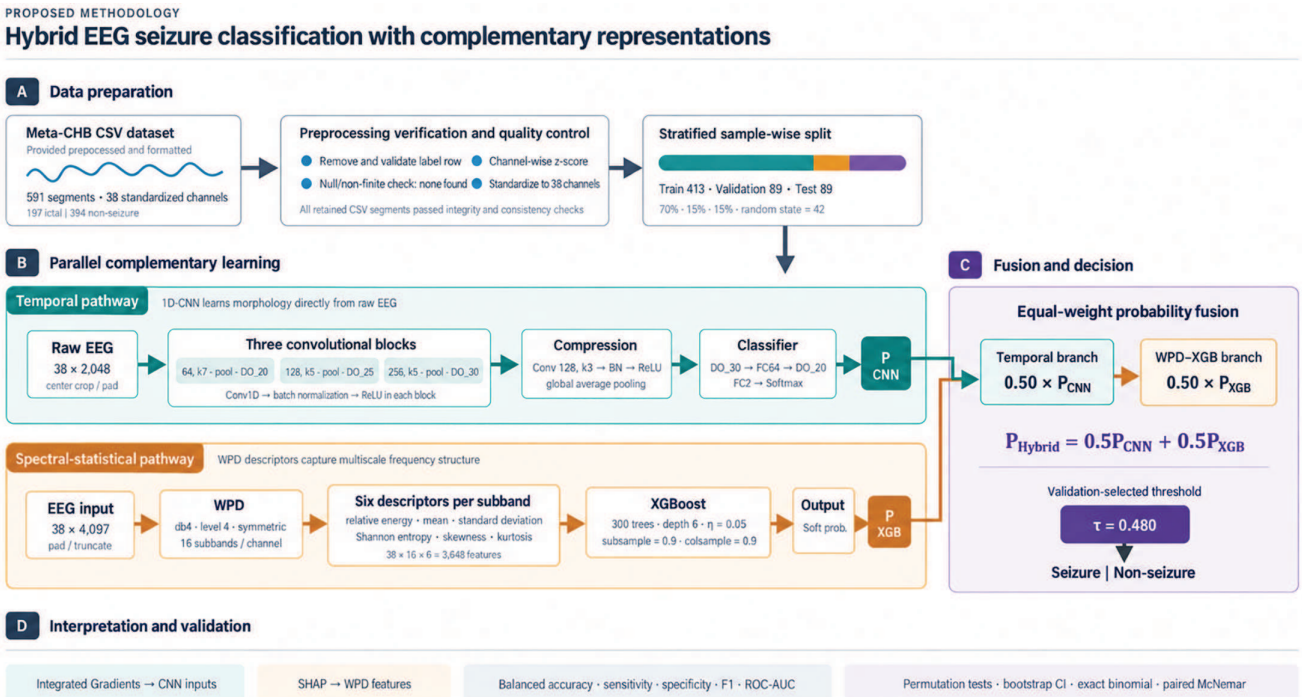
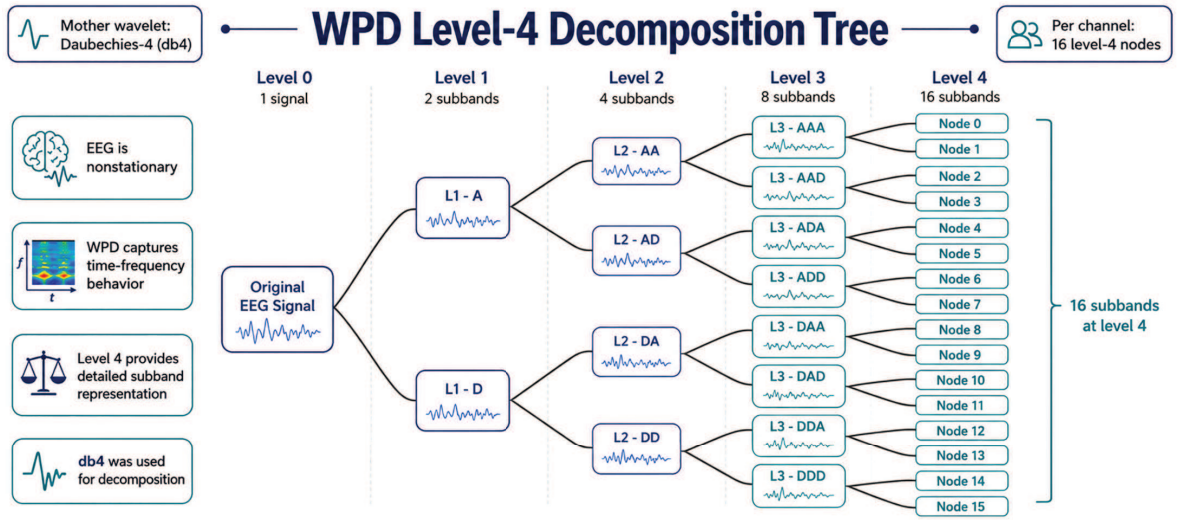


Table 2
CNN architecture and exact dropout locations in the deep learning branch

Stage	Operations/configuration	Dropout
Input	38 channels × 2048 samples	—
Block 1	Conv1D 38→ 64, kernel 7, stride 1, padding 3; BN; ReLU; MaxPool(2)	0.20 after pooling
Block 2	Conv1D 64→ 128, kernel 5, stride 1, padding 2; BN; ReLU; MaxPool(2)	0.25 after pooling
Block 3	Conv1D 128→ 256, kernel 5, stride 1, padding 2; BN; ReLU; MaxPool(2)	0.30 after pooling
Block 4	Conv1D 256→ 128, kernel 3, stride 1, padding 1; BN; ReLU	None
Compression	Adaptive global average pooling to length 1; flatten	—
Classifier	Dropout; FC 128→ 64; ReLU; dropout; FC 64→ 2; softmax	0.30 before FC; 0.20 after ReLU

Figure 2
Level-4 db4 wavelet packet decomposition used to generate time–frequency features from EEG signals



3.5. Probability fusion

The final seizure probability was obtained through probability-level fusion of the CNN and XGBoost-WPD branch outputs. Let P_{CNN} and P_{XGB} denote the seizure probabilities predicted by the CNN and XGBoost-WPD branches, respectively. The fused probability was computed as

$$P_{Hybrid} = w_{CNN}P_{CNN} + w_{XGB}P_{XGB} \quad (1)$$

where $W_{CNN} + W_{XGB} = 1$. Candidate fusion weights were evaluated on the validation set, resulting in equal weights of $w_{CNN} = 0.5$ and $w_{XGB} = 0.5$. Thus, $P_{Hybrid} = 0.5P_{CNN} + 0.5P_{XGB}$. A decision threshold of 0.480, selected on the validation set by maximizing balanced accuracy, was applied to P_{Hybrid}. Neither the fusion weights nor the decision threshold was selected using the held-out test set.

The standalone XGBoost-WPD branch used its independently selected validation-derived threshold of 0.280, whereas the hybrid classifier required a separate threshold because classification was based on the fused probability PHybrid.

3.6. Explainability and statistical validation

To improve model transparency, branch-specific explainability analyses were performed for the two complementary classifiers. SHAP [16] was used to quantify the contribution of WPD-derived features to predictions from the XGBoost-WPD branch, while IG [17] was used to identify channel–time regions contributing to CNN predictions.

Performance was evaluated using accuracy, precision, recall (sensitivity), specificity, F1-score, balanced accuracy, and Receiver Operator Characteristics- Area Under Curve (ROC-AUC). Balanced accuracy was treated as the primary metric because it accounts for both seizure sensitivity and non-seizure specificity.

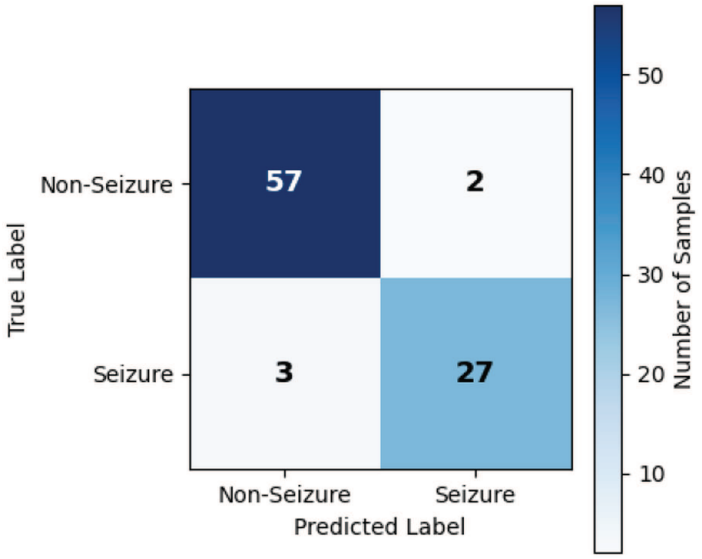
Statistical reliability was assessed using exact permutation testing for balanced accuracy [18], nonparametric bootstrap confidence intervals, exact binomial testing against the majority-class baseline, ROC-AUC permutation testing, and exact paired McNemar tests [19]. Bootstrap 95% confidence intervals were estimated using 5,000 resamples with replacement. Statistical significance was assessed at $\alpha = 0.05$.

4. Results and Discussion

4.1. Classification performance

The final CNN + XGBoost-WPD hybrid model was evaluated on the held-out test set using the fusion weights and decision threshold selected during validation. Figure 3 presents the resulting confusion matrix. The model correctly classified 84 of 89

Figure 3
Confusion matrix of the proposed CNN + XGBoost-WPD hybrid model on the held-out test set

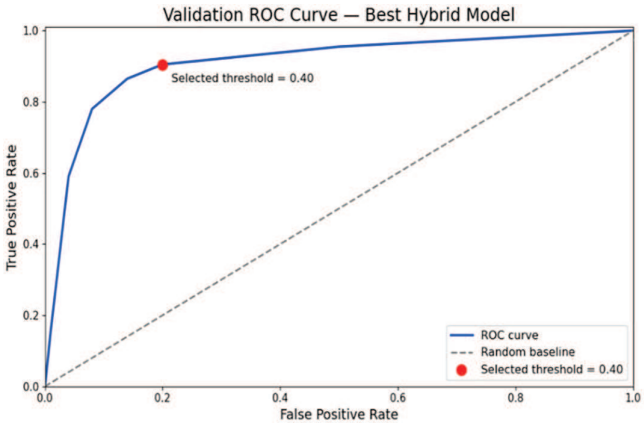


test segments, comprising 57 true negatives and 27 true positives, with 2 false positives and 3 false negatives.

The quantitative performance metrics are summarized in Table 3. The proposed framework achieved an accuracy of 0.9438, balanced accuracy of 0.9331, precision of 0.9310, sensitivity of 0.9000, specificity of 0.9661, F1-score of 0.9153, and ROC-AUC of 0.9819. The balanced accuracy of 0.9331 indicates that performance remained strong across both seizure and non-seizure classes despite the unequal class distribution.

Specificity of 0.9661 was slightly higher than sensitivity of 0.9000, reflecting fewer false-positive than false-negative classifications relative to the corresponding class sizes. The F1-score of 0.9153 indicates a strong balance between precision and sensitivity. Figure 4 presents the receiver operating characteristic (ROC) curve of the final hybrid model. The ROC-AUC of 0.9819 indicates strong discrimination between seizure and non-seizure segments across the evaluated decision thresholds.

Figure 4
Receiver operating characteristic (ROC) curve of the proposed CNN + XGBoost-WPD hybrid model on the held-out test set (ROC-AUC = 0.9819)



4.2. Statistical reliability analysis

Beyond conventional classification metrics, statistical tests were conducted to assess whether the observed performance exceeded chance-level expectations. These included permutation tests for balanced accuracy and ROC-AUC and an exact binomial test against the majority-class baseline.

Table 3
Quantitative performance metrics of the proposed CNN + XGBoost-WPD hybrid model on the held-out test set

Metric	Value
Balanced accuracy	0.9331
ROC-AUC	0.9819
Accuracy	0.9438
Precision	0.9310
Recall (sensitivity)	0.9000
Specificity	0.9661
F1-score	0.9153

Table 4 summarizes the statistical reliability results for the final CNN + XGBoost-WPD hybrid model. The balanced-accuracy permutation test yielded $p = 3.11 \times 10^{-17}$, providing evidence that the observed balanced accuracy of 0.9331 was inconsistent with the chance-level null hypothesis.

The exact binomial test against the majority-class baseline yielded $p = 2.05 \times 10^{-10}$, indicating that the observed classification accuracy was significantly higher than that obtained by always predicting the majority non-seizure class. The ROC-AUC permutation test yielded $p = 5.00 \times 10^{-5}$, providing evidence that the observed ROC-AUC of 0.9819 exceeded that expected under random label association.

4.3. Bootstrap confidence intervals

To quantify uncertainty in the reported performance estimates, 95% confidence intervals were computed using 5,000

Table 4
Statistical reliability analysis of the proposed CNN + XGBoost-WPD hybrid model

Statistical test	Result	Interpretation
Balanced-accuracy permutation test	$p = 3.11 \times 10^{-17}$	Significant above chance
Exact binomial test vs majority baseline	$p = 2.05 \times 10^{-10}$	Significant above majority baseline
ROC-AUC permutation test	$p = 5.00 \times 10^{-5}$	Significant above chance

nonparametric bootstrap resamples of the held-out test set with replacement.

Table 5 summarizes the resulting 95% confidence intervals. The bootstrap intervals quantify uncertainty in performance on the present held-out partition. As shown in Table 5, balanced accuracy was 0.9331 (95% CI: 0.8664–0.9833). The confidence intervals for sensitivity, specificity, and ROC-AUC were 0.7826–1.0000, 0.9174–1.0000, and 0.9460–1.0000, respectively. Although the intervals are compatible with strong discrimination, their width—particularly for sensitivity, which was estimated from only 30 seizure segments—indicates non-negligible uncertainty. These intervals should therefore not be interpreted as evidence of reproducibility beyond the present test partition.

4.4. Ablation and standalone model comparison

To evaluate the contribution of probability-level fusion, the final framework was compared with its standalone CNN and XGBoost-WPD branches and with InceptionTimeMini under the same held-out test protocol. Model-specific decision thresholds were selected using the validation set only. As summarized in Table 6, the observed balanced accuracies were 93.31% for the hybrid framework, 89.12% for the CNN branch, 86.58% for XGBoost-WPD, and 92.40% for InceptionTimeMini.

Relative to the standalone branches, the hybrid model showed observed balanced-accuracy differences of 4.19 percentage points versus CNN and 6.73 percentage points versus XGBoost-WPD. The corresponding difference relative to InceptionTimeMini was 0.91 percentage points. Paired McNemar comparisons of predictions from the same held-out test samples are presented in Section 4.5.

4.5. Comparative evaluation with published methods

To evaluate the proposed framework beyond comparisons with its constituent and conventional baseline models, three representative EEG seizure-classification methods from the published literature were adapted and reimplemented under the same experimental protocol. These included the raw-EEG CNN architecture of Acharya et al. [11], the medium-weight Discrete Wavelet Transform (DWT)-based D-CNN of Nemati and Meshgini [12], and the hybrid CNN–machine learning approach of Hassan et al. [15].

Table 5
Bootstrap 95% confidence intervals for the performance metrics of the proposed CNN + XGBoost-WPD hybrid model

Metric	Estimate	95% CI
Balanced accuracy	0.9331	[0.8664, 0.9833]
Recall (sensitivity)	0.9000	[0.7826, 1.0000]
Specificity	0.9661	[0.9174, 1.0000]
ROC-AUC	0.9819	[0.9460, 1.0000]

Figure 5
ROC curves of the proposed framework and adapted published EEG seizure-classification methods on the common held-out test set

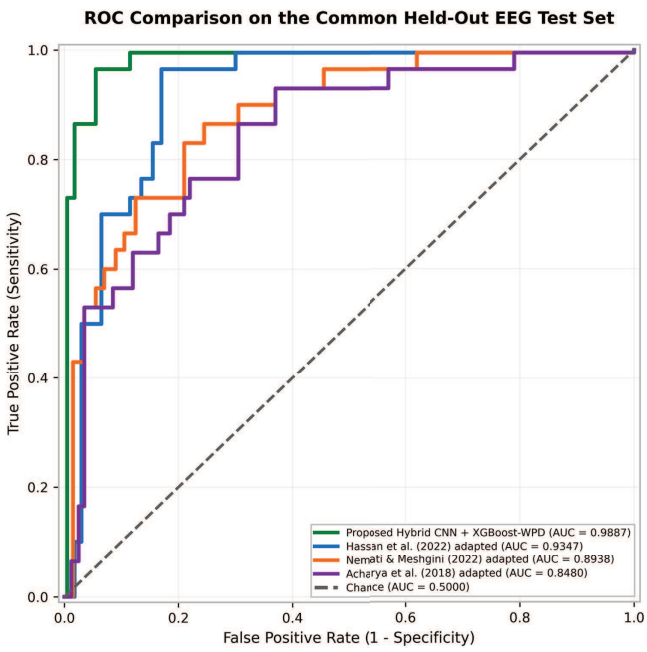


Table 6
Ablation and standalone model performance on the common held-out test set

Configuration	Representation/role	Balanced accuracy (%)
XGBoost-WPD	Standalone WPD-derived statistical features	86.58
CNN	Standalone raw-EEG temporal branch	89.12
Inception-TimeMini	Standalone deep learning comparator	92.40
CNN + XGBoost-WPD	Equal-weight probability fusion (proposed)	93.31

The term "adapted" is used because the original architectures were modified where necessary to accommodate the input representation and binary classification setting employed in this study.

The published architectures were adapted only where necessary to match the present input and binary target. For [11], the 13-layer raw-EEG CNN structure—five convolutional layers, five pooling layers, and three fully connected layers—was retained, while the input was changed from one to 38 channels

and the output from three to two classes; signals were represented with 4,097 temporal samples and trained with a batch size of 12. For reference [12], level-5 db4 decomposition generated the A5, D5, D4, and D3 bands for every channel, producing 152 input maps that were resampled to length 1,024 and processed by the reported four-stage convolutional design with RMSprop and a two-class output. For reference [15], signals were filtered from 0.5 to 50 Hz using a second-order zero-phase Butterworth filter, transformed using level-4 db4 coefficients, processed by the reported two-convolution CNN, reduced to the top 500 features using mutual information, and classified using the reported Support Vector Machine - kNearest Neighbor (SVM-kNN) stacking strategy; the target was adapted to ictal versus non-seizure and the input to the common 38 × 2048 representation. All comparators used the identical 413/89/89 partitions, validation-only threshold selection, and common held-out test set. The results are therefore same-data adapted reimplementations, not direct comparisons with accuracy values reported under the original studies' different targets or validation protocols.

All methods were trained and evaluated using the same 591 EEG segments and common stratified sample-wise partitions. Model selection and decision-threshold optimization were performed using the validation set only, while the held-out test set was reserved for final evaluation. Table 7 summarizes the resulting comparative performance. All performance values in Table 7 are reported as percentages, and bold values indicate the best result for each metric.

As shown in Table 7, the proposed framework yielded an observed balanced accuracy of 93.31% and ROC-AUC of 98.19%. Among the adapted published methods, Hassan et al. [15] yielded a balanced accuracy of 87.37% and ROC-AUC of 93.47%, corresponding to observed differences of 5.94 and 4.72 percentage points, respectively.

Both the proposed framework and the adapted Hassan et al. [15] implementation achieved a sensitivity of 90.00%, correctly identifying 27 of 30 seizure segments. However, the proposed framework produced two false-positive segment classifications compared with 9 for the adapted Hassan et al. [15] implementation, while maintaining the same number of true-positive classifications. This difference was reflected in higher observed precision and specificity on the present test partition.

The adapted Nemati and Meshgini [12] implementation yielded a balanced accuracy of 81.53% and ROC-AUC of 90.73%. The adapted Acharya et al. [11] implementation showed the highest sensitivity (93.33%) among the published comparators, but lower precision (54.90%) and specificity (61.02%), illustrating the trade-off between sensitivity and false-positive classifications at the selected operating threshold.

Figure 5 presents the ROC curves of the proposed framework and the adapted published methods on the common held-out test set, providing a threshold-independent comparison of their discriminative performance. The proposed framework yielded the largest observed ROC-AUC (98.19%) among the evaluated adapted methods. However, this comparison is descriptive because no paired statistical test of ROC-AUC differences was performed. Furthermore, although fewer false-positive segments were observed for the proposed framework than for the adapted Hassan et al. [15], continuous-recording false alarms per hour were not evaluated.

Because paired predictions were available for the internal baseline models evaluated under the same experimental protocol, exact McNemar tests were performed to compare their predictions with those of the proposed framework. The results are summarized in Table 8.

McNemar's exact test identified a statistically significant difference between the hybrid model and the standalone CNN ($p = 0.0117$). Differences relative to XGBoost-WPD ($p = 0.1094$) and InceptionTimeMini ($p = 0.1185$) were not statistically significant at $\alpha = 0.05$. Accordingly, the higher point estimates for these latter comparisons should be interpreted descriptively. The modest test-set size and bootstrap uncertainty further warrant cautious interpretation. Overall, the comparative results show higher observed performance for the proposed framework under the present experimental protocol, while the paired analysis provides statistical evidence for a difference relative to the standalone CNN.

Table 7
Performance of the proposed framework and adapted published EEG seizure-classification methods on the common held-out test set

Method	Accuracy	Balanced acc.	Precision	Sensitivity	Specificity	F1-score	ROC-AUC
Acharya et al. (adapted) [11]	71.91	77.18	54.90	93.33	61.02	69.14	84.80
Nemati and Meshgini (adapted) [12]	82.02	81.53	70.59	80.00	83.05	75.00	90.73
Hassan et al. (adapted) [15]	86.52	87.37	75.00	90.00	84.75	81.82	93.47
CNN + XGBoost-WPD (proposed)	94.38	93.31	93.10	90.00	96.61	91.53	98.19

Table 8
Exact paired McNemar comparisons of the hybrid model and internal baseline models on the common held-out test set

Comparison	Exact McNemar p-value	Significant at $\alpha = 0.05$
Hybrid vs CNN	0.0117	Yes
Hybrid vs XGBoost-WPD	0.1094	No
Hybrid vs InceptionTimeMini	0.1185	No

4.6. Explainability analysis

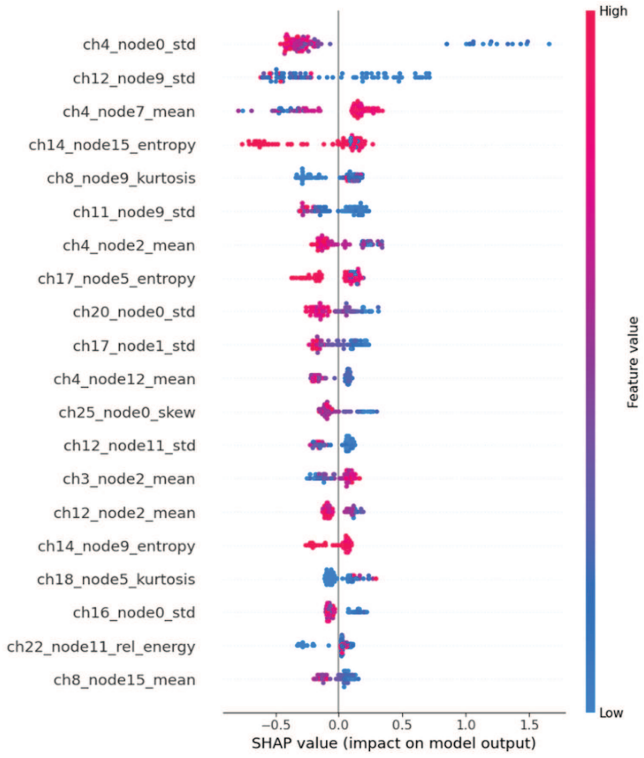
To examine the basis of predictions from the two complementary branches, SHAP was applied to the XGBoost-WPD branch, while IG was used to interpret the CNN branch.

Figure 6 presents the SHAP summary plot for the XGBoost-WPD branch. The influential features include statistical descriptors derived from the WPD subbands, including standard deviation, entropy, kurtosis, skewness, and relative energy. These features are distributed across multiple decomposition subbands rather than being concentrated within a single subband. The SHAP distributions also show that both low and high feature values are associated with positive and negative contributions to the classifier output, reflecting the nonlinear relationships captured by the XGBoost model. Features are ranked by their contribution to model predictions; color represents feature magnitude, and horizontal position indicates the impact on the classifier output.

Figure 7 presents the IG attribution analysis for a correctly classified seizure segment. The attribution heatmap shows a non-uniform distribution of attribution across EEG channels and temporal locations, with higher attribution concentrated in specific channel–time regions. The channel-wise and temporal attribution profiles further indicate that the CNN prediction for this example was associated with a subset of channels and localized temporal regions rather than being distributed uniformly across the entire input segment. The top panel shows the channel-time attribution heatmap, the middle panel shows channel-wise attribution scores, and the bottom panel shows temporal attribution scores.

Altogether, the SHAP and IG analyses provide complementary views of the two model branches: SHAP characterizes the contribution of WPD-derived statistical features to XGBoost-WPD predictions, whereas IG identifies channel–time regions associated with CNN predictions. These analyses improve model transparency but do not by themselves establish clinical validity or deployment readiness.

Figure 6
SHAP summary plot showing the contributions of WPD-derived features to XGBoost-WPD predictions



5. Discussion

The experimental results indicate that the proposed CNN + XGBoost-WPD framework achieved strong discrimination between seizure and non-seizure EEG segments on the held-out test partition. The ablation results (Table 6) showed that probability-level fusion yielded higher observed performance than either standalone branch, supporting the rationale for combining complementary representations. Specifically, the CNN learns temporal representations directly from standardized EEG signals, whereas the XGBoost-WPD branch models statistical descriptors derived from multiscale time–frequency decomposition. Their integration, therefore, provides complementary sources of predictive information rather than relying on a single signal representation.

The explainability analyses (Figures 6 and 7) provide additional insight into how the two branches use their respective input representations. SHAP identified influential WPD-derived statistical features across multiple decomposition subbands, whereas IG highlighted localized channel–time regions associated with the CNN prediction in the examined seizure example. These findings are consistent with the intended representational complementarity of the hybrid framework, although the explainability analyses should not be interpreted as causal evidence for the observed performance gains.

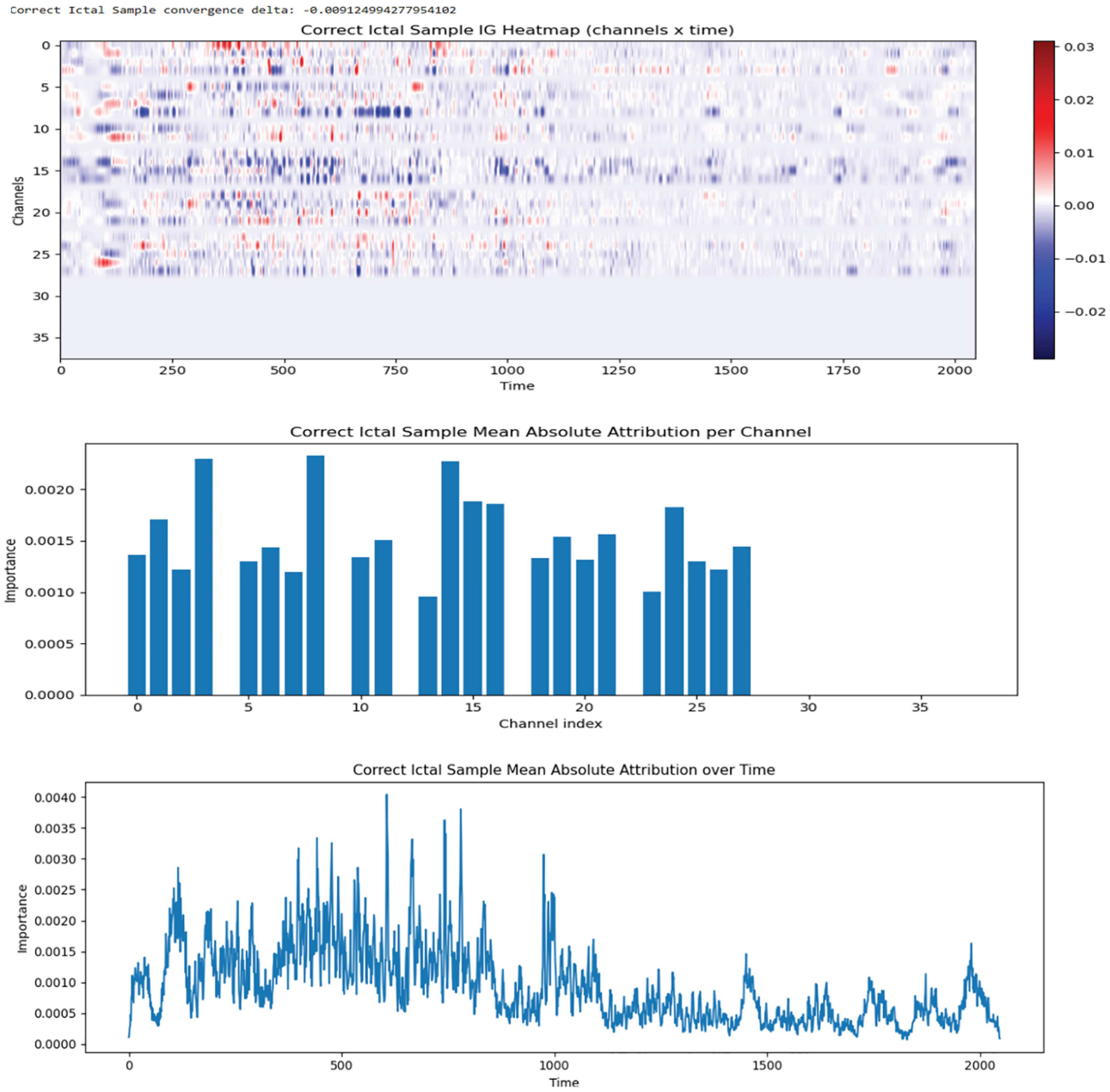
From an application perspective, automated seizure classification could potentially support EEG review and monitoring workflows. However, the present study evaluates offline binary segment classification and does not establish clinical effectiveness or deployment readiness. Extension to multiclass seizure classification, patient-specific modeling, and continuous EEG analysis therefore represents an important direction for future investigation.

The two branches have different computational profiles. The CNN contains approximately 330,242 trainable parameters and processes a 38 × 2048 input tensor. The WPD branch performs level-4 DB4 decomposition, producing 16 terminal subbands per channel; extraction of six descriptors per subband yields 38 × 16 × 6 = 3648 features. The XGBoost classifier uses 300 trees with a maximum depth of 6. Probability fusion and thresholding introduce negligible computation relative to feature extraction and branch inference.

The present study does not report measured end-to-end inference latency and therefore does not establish real-time capability. At 256 Hz, a 2048-sample window corresponds to approximately 8 seconds of EEG acquisition, and the current center-cropping procedure was designed for offline segment analysis rather than causal streaming. A prospective implementation would require sliding causal windows and device-specific evaluation of latency, throughput, memory use, model size, and energy consumption. Continuous EEG evaluation should additionally quantify false alarms per hour and detection delay. Thus, the present findings support offline segment-level classification and motivate, rather than establish, future real-time clinical monitoring. However, the model has not undergone prospective clinical validation, patient-independent external evaluation, or continuous-recording testing. Its clinical utility, false-alarm burden, and operational latency must therefore be established before deployment.

Figure 7

Integrated Gradients attribution heatmap and channel-wise and temporal attribution profiles for a correctly classified seizure segment



Several limitations constrain the interpretation of these findings. First, the analysis used 591 segments from a single public dataset derivative and one fixed stratified sample-wise 70%–15%–15% split. Although permutation tests and bootstrap confidence intervals quantify chance-level evidence and sampling uncertainty on this held-out partition, they do not replace repeated cross-validation, subject-wise validation, or external validation. Because the patient-nonspecific segment formulation may permit correlated segments from the same patient to occur across partitions, the reported results should not be interpreted as evidence of patient-independent generalization. Second, only 30 seizure segments were included in the test set, contributing to uncertainty in sensitivity estimates. Third, the analysis was limited to binary segment classification and did not evaluate seizure subtypes, continuous-recording false-alarm rates, or detection delay. Future work should therefore include repeated subject-wise cross-validation, independent external cohorts, multiclass evaluation, and continuous EEG monitoring protocols.

Overall, the results indicate that combining complementary temporal and time–frequency representations can improve observed performance on the present held-out partition. The principal contribution should therefore be understood as an empirically evaluated methodological integration of complementary representations, branch-specific explainability, and statistical validation, rather than as a fundamentally new CNN or boosting architecture.

6. Conclusion

This study presented a hybrid EEG seizure-classification framework that integrates temporal representations learned by a 1D-CNN with statistical time–frequency features modeled using WPD and XGBoost. On the held-out test partition, the proposed framework achieved a balanced accuracy of 93.31%, sensitivity of 90.00%, specificity of 96.61%, and ROC-AUC of 98.19%. Statistical analyses, including permutation testing, bootstrap confidence intervals, and paired McNemar comparisons, provided complementary evidence regarding performance significance and uncertainty.

Under the common experimental protocol, the proposed framework yielded higher observed performance than the adapted published-method implementations across most evaluation metrics. Relative to the strongest published implementation, it maintained the same sensitivity of 90.00% while producing 2 false-positive segment classifications compared with 9. These comparisons should be interpreted within the present held-out partition because statistical significance of the ROC-AUC differences between the adapted published methods was not evaluated.

Branch-specific explainability further provided complementary views of the two classifiers: SHAP characterized the contribution of WPD-derived statistical features to XGBoost-WPD predictions, while IG identified localized channel-time regions associated with CNN predictions. Together with the ablation results, these findings support the rationale for integrating complementary temporal and time–frequency representations within a common framework. Future work will focus on repeated subject-wise validation and independent external EEG cohorts, extension to multiclass seizure classification, and evaluation under continuous-recording and real-time monitoring settings.

Ethical Statement

This research work has utilized a publicly available dataset from PhysioNet. This dataset is distributed under Open Data Commons Attribution License v1.0. Since the data are completely anonymized, publicly available, and contains no human subject information, this study did not require ethical approval from our institutional Research Ethics Board.

Conflicts of Interest

Noushath Shaffi is the Associate Editor for *Artificial Intelligence and Applications*, and he was not involved in the editorial review or the decision to publish this article. The authors declare that they have no conflicts of interest in this work.

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.6062372, reference number [9].

Author Contribution Statement

Mohamed Al Saidi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Noushath Shaffi:** Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Vimbi Viswan:** Validation, Writing – review & editing, Supervision. **Abdelhamid Abdesselam:** Methodology, Writing – review & editing, Supervision. **Rachid Hedjam:** Writing – review & editing, Supervision.

References

[1] Han, K., Liu, C., & Friedman, D. (2024). Artificial intelligence/machine learning for epilepsy and seizure diagnosis. *Epilepsy & Behavior*, *155*, 109736. https://doi.org/10.1016/j.yebeh.2024.109736

[2] Saminu, S., Xu, G., Zhang, S., Kader, Ab El., Aliyu, I., A, H., Jabire, A. H., Adamu, M. J., … Adamu, M. J. (2023). Applications of artificial intelligence in automatic detection of epileptic seizures using EEG signals: A review. *Artificial Intelligence and Applications*, *1*(1), 11–25. https://doi.org/10.47852/bonviewAIA2202297

[3] Alkhaldi, M., Joudeh, L. A., Ahmed, Y. B., & Husari, K. S. (2024). Artificial intelligence and telemedicine in epilepsy and EEG: A narrative review. *Seizure: European Journal of Epilepsy*, *121*, 204–210. https://doi.org/10.1016/j.seizure.2024.08.024

[4] Cao, W., Hedjam, R., Abdulrazak, B., & Zhu, J. P. (2025). A hierarchical wavelet bispectrum framework for automated seizure detection. In *2025 International Conference on Big Data and Data Mining*, 1–5. https://doi.org/10.1109/BDDM68348.2025.11442079

[5] Kudale, T., & Pagare, R. (2026). A comprehensive review for seizure prediction based on EEG dataset: From handcrafted features to deep neural models. *IEEE Access*, *14*, 44591–44606. https://doi.org/10.1109/ACCESS.2026.3674570

[6] Kunekar, P., Gupta, M. K., & Gaur, P. (2024). Detection of epileptic seizure in EEG signals using machine learning and deep learning techniques. *Journal of Engineering and Applied Science*, *71*(1), 21. https://doi.org/10.1186/s44147-023-00353-y

[7] Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., …, & Pardalos, P. M. (2024). A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*, *118*, 109370. https://doi.org/10.1016/j.compeleceng.2024.109370

[8] Ahmad, I., Zhu, M., Li, G., Javeed, D., Kumar, P., & Chen, S. (2024). A secure and interpretable AI for smart healthcare system: A case study on epilepsy diagnosis using EEG signals. *IEEE Journal of Biomedical and Health Informatics*, *28*(6), 3236–3247. https://doi.org/10.1109/JBHI.2024.3366341

[9] Handa, P., Gupta, E., Gupta, M., Anand Sachdeva, R., Goel, N., & Woitek, R. (2022). Meta-EEG of CHB-MIT Scalp EEG Database v1.0.0.0 [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.6062372

[10] Wu, Y., Su, E. L. M., Wu, M., Ooi, C. Y., & Holderbaum, W. (2025). A review of machine learning and deep learning trends in EEG-based epileptic seizure prediction. *IEEE Access*, *13*, 159812–159842. https://doi.org/10.1109/ACCESS.2025.3606966

[11] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, *100*, 270–278. https://doi.org/10.1016/j.compbiomed.2017.09.017

[12] Nemati, N., & Meshgini, S. (2022). A medium-weight deep convolutional neural network-based approach for onset

epileptic seizures classification in EEG signals. *Brain and Behavior*, *12*(11), e2763. https://doi.org/10.1002/brb3.2763

[13] Shen, M., Wen, P., Song, B., & Li, Y. (2023). Real-time epilepsy seizure detection based on EEG using tunable-Q wavelet transform and convolutional neural network. *Biomedical Signal Processing and Control*, *82*, 104566. https://doi.org/10.1016/j.bspc.2022.104566

[14] Anita, M., & Kowshalya, A. M. (2024). Automatic epileptic seizure detection using MSA-DCNN and LSTM techniques with EEG signals. *Expert Systems with Applications*, *238*, 121727. https://doi.org/10.1016/j.eswa.2023.121727

[15] Hassan, F., Hussain, S. F., & Qaisar, S. M. (2022). Epileptic seizure detection using a hybrid 1D CNN-machine learning approach from EEG data. *Journal of Healthcare Engineering*, *2022*(1), 9579422. https://doi.org/10.1155/2022/9579422

[16] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

[17] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 3319–3328.

[18] Holt, C. A., & Sullivan, S. P. (2023). Permutation tests for experimental data. *Experimental Economics*, *26*(4), 775–812. https://doi.org/10.1007/s10683-023-09799-6

[19] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

[20] Zhang, J., Zheng, S., Chen, W., Du, G., Fu, Q., & Jiang, H. (2024). A scheme combining feature fusion and hybrid deep learning models for epileptic seizure detection and prediction. *Scientific Reports*, *14*(1), 16916. https://doi.org/10.1038/s41598-024-67855-4

[21] Sun, Q., Liu, Y., & Li, S. (2024). Automatic seizure detection using multi-input deep feature learning networks for EEG signals. *Journal of Sensors*, *2024*(1), 8835396. https://doi.org/10.1155/2024/8835396

[22] Amrani, G., Adadi, A., & Berrada, M. (2024). An explainable hybrid DNN model for seizure vs. non-seizure classification and seizure localization using multi-dimensional EEG signals. *Biomedical Signal Processing and Control*, *95*, 106322. https://doi.org/10.1016/j.bspc.2024.106322

[23] Murugan, T. K., & Kameswaran, A. (2024). Employing convolutional neural networks and explainable artificial intelligence for the detection of seizures from electroencephalogram signal. *Results in Engineering*, *24*, 103378. https://doi.org/10.1016/j.rineng.2024.103378

[24] Khan, F. A., Umar, Z., Jolfaei, A., & Tariq, M. (2024). Explainable fuzzy deep learning for prediction of epileptic seizures using EEG. *IEEE Transactions on Fuzzy Systems*, *32*(10), 5428–5437. https://doi.org/10.1109/TFUZZ.2024.3434709

[25] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

How to Cite: Saidi, M. A., Shaffi, N., Viswan, V., Abdesselam, A., & Hedjam, R. (2026). Toward Trustworthy EEG-Based Epileptic Seizure Classification: A Hybrid Explainable AI Framework with Statistical Validation. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA620211142