Automated Epileptic Seizure Detection from EEG Signals Using a Hybrid BiLSTM-RVM Model

Charles Aba Medzo [1,*], Gildorain Mballa Ndongo [1], Georges Olle Olle [1,2], Handy Kenne Evina [1], Diane Tchakonte Tchuani [1], and Halidou Aminou [2]

[1] Department of Computer Engineering, Higher Technical Teachers' Training College, University of Ebolowa, Ebolowa, Cameroon
[2] Department of Computer Science, Faculty of Sciences, University of Yaoundé I, Yaoundé, PO. BOX 812, Cameroon
Email: medzocharles@yahoo.com (C.A.M.); ndongogildorain@gmail.com (G.M.N.); georgesdelortolleolle@gmail.com (G.O.O.); junior.evina@facsciences-uy1.cm (H.K.E.); diane.tchuani@gmail.com (D.T.T.); halidou.aminou@facsciences-uy1.cm (H.A.)
*Corresponding author

Abstract—**Epilepsy is a chronic neurological disorder characterized by recurrent seizures caused by abnormal electrical activity in the brain, for which timely and accurate automated detection is essential for effective clinical management. Manual Electroencephalography (EEG) interpretation is labor-intensive and subject to inter-expert variability, highlighting the need for automated, reliable, and clinically deployable diagnostic systems. This study proposes a hybrid Bidirectional Long Short-Term Memory-Relevance Vector Machine (BiLSTM-RVM) framework that combines a Bidirectional Long Short-Term Memory (BiLSTM) network with a Relevance Vector Machine (RVM) for automated epileptic seizure detection from Electroencephalography (EEG) signals. The proposed framework explicitly addresses four key challenges: inter-subject generalization, model interpretability, cross-database robustness, and resistance to signal degradation. Evaluated on the University of California, Irvine (UCI) Epileptic Seizure Recognition Dataset, the model achieves an accuracy of 99.57% and an AUC of 99.97%. A Leave-One-Subject-Out (LOSO) cross-validation strategy is employed to evaluate inter-subject generalization, while cross-database assessment on the Children's Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) Scalp EEG Database confirms strong robustness across diverse acquisition conditions and patient populations. Furthermore, Gaussian noise sensitivity analysis is conducted to assess model robustness under simulated real-world EEG artifacts. SHapley Additive exPlanations (SHAP) are employed to provide temporal interpretation of model decisions, enabling clinically meaningful identification of EEG segments associated with seizure activity. Overall, the results demonstrate that the proposed BiLSTM-RVM framework provides an accurate, generalizable, interpretable, and noise-resilient solution for computer-assisted epilepsy diagnosis.**

Keywords—**epilepsy, Electroencephalography (EEG), epileptic seizure detection, Bidirectional Long Short-Term Memory (BiLSTM), Relevance Vector Machine (RVM), leave-one-subject-out, shapley additive explanations**

I. INTRODUCTION

Epilepsy is a chronic neurological disorder characterized by recurrent, unpredictable seizures arising from abnormal electrical activity in the brain, manifesting as brief losses of consciousness or severe convulsions [1]. Affecting more than 60 million people worldwide, it ranks among the most prevalent and debilitating neurological conditions [2, 3], with significant consequences for patients' emotional, social, and cognitive well-being. Despite considerable therapeutic advances, approximately 30% of patients remain refractory to treatment, underscoring the urgent need for more effective diagnostic and monitoring approaches [4–6].

Electroencephalography (EEG) remains the primary clinical tool for epileptic seizure detection and monitoring, providing critical information about brain electrical activity and seizure patterns [7]. However, manual analysis of EEG recordings is labor-intensive and heavily dependent on expert neurologists, making it prone to inter-expert variability, particularly during long-term monitoring sessions or when large data volumes are involved [8, 9]. These limitations have motivated the development of automated seizure detection methods aimed at improving both diagnostic efficiency and clinical reliability.

Recent advances in deep learning have significantly improved biomedical signal analysis, particularly for automated seizure detection [10, 11]. Convolutional Neural Networks (CNNs) excel at extracting discriminative spatial features, while Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, effectively capture temporal dependencies inherent in EEG signals [12–14]. Building on these capabilities, a growing body of research has focused on improving automated epileptic seizure detection and continuous patient monitoring.

Manuscript received December 31, 2025; revised February 27, 2026; accepted June 1, 2026; published September 14, 2026.

doi: 10.12720/jait.17.9.1759-1782

In this work, we propose a hybrid BiLSTM-RVM framework combining a Bidirectional Long Short-Term Memory (BiLSTM) network with a Relevance Vector Machine (RVM) for automated epileptic seizure detection from EEG signals. The BiLSTM models temporal dependencies in both forward and backward directions, effectively capturing the complex dynamics of EEG sequences, while the RVM provides a sparse, probabilistic classifier that enhances generalization particularly in imbalanced binary classification settings.

The main contributions of this work are summarized as follows:

- We propose a hybrid BiLSTM-RVM framework for automated epileptic seizure detection from EEG signals, combining deep temporal feature learning with probabilistic classification.
- The model leverages bidirectional recurrent architectures to extract discriminative temporal features from EEG sequences, capturing long-range dependencies and complex non-stationary dynamics.
- A sparse Bayesian Relevance Vector Machine (RVM) is integrated to enhance robustness, probabilistic interpretability, and generalization, while preserving a compact model structure.
- The proposed framework explicitly addresses four key challenges in EEG-based learning systems: inter-subject generalization (assessed via LOSO cross-validation), cross-database robustness, clinical interpretability the latter supported by SHapley Additive exPlanations (SHAP) and robustness to signal degradation, evaluated through Gaussian noise sensitivity analysis.
- Extensive experiments on publicly available EEG benchmarks demonstrate that the proposed model is robust, highly generalizable, and competitive with state-of-the-art methods across intra-database, inter-subject, and cross-database evaluation settings.

The remainder of this paper is organized as follows. Section II reviews related work on epileptic seizure detection. Section III describes the materials and methods. Section IV presents the experimental results. Section V discusses the strengths and limitations of the proposed system. Finally, Section VI concludes and outlines directions for future research.

II. Literature Review

Automated epileptic seizure detection from EEG signals has attracted significant research interest in recent decades. Existing approaches can be categorized into three groups: traditional machine learning methods, deep learning techniques, and hybrid frameworks. Although these methods have shown promising results, they exhibit notable limitations in terms of feature representation, computational complexity, interpretability, and cross-subject generalization. This section critically reviews representative works from each category, with emphasis on their key limitations and the challenges that motivate the proposed BiLSTM-RVM framework.

A. Traditional Automatic Detection Methods

Traditional approaches rely on manual extraction of time- and frequency-domain features from EEG signals, followed by classification using conventional algorithms such as Support Vector Machines (SVM) or k-Nearest Neighbors (kNN). However, these methods are sensitive to noise and often fail to generalize effectively, owing to the complex and non-stationary nature of EEG signals.

An epileptic seizure detection method based on sparse multiscale Radial Basis Function (RBF) networks combined with Fisher vector encoding was proposed in Ref. [15]. EEG signals were transformed into time-frequency representations, and dimensionality reduction techniques were applied to extract Gray-Level Co-occurrence Matrix (GLCM) and Fisher Vector (FV) descriptors, subsequently classified using an SVM. However, its high computational cost and multi-stage processing pipeline rendered this approach unsuitable for real-time clinical applications.

Several studies have investigated wavelet-based approaches for epileptic seizure detection. In 2019, nonlinear features were combined with Discrete Wavelet Transform (DWT) decomposition to improve epileptic seizure detection performance [16]. However, reliance on manually selected features limited adaptability and neglected long-term temporal dependencies between EEG segments. In 2020, a DWT-based multiscale framework was introduced [17], in which a genetic algorithm was employed to select the most relevant features from 54 wavelet coefficients for epileptic seizure classification. Despite improved performance, this approach remained sensitive to parameter tuning and computationally costly.

In 2021, SMOTE and ADASYN oversampling techniques were applied to DWT-extracted features to address class imbalance [18]. Although detection performance improved, synthetic oversampling introduced potential statistical bias and raised concerns regarding generalization to real clinical data. In 2022, combinations of DWT-derived features with multiple classifiers including KNN, SVM, Naive Bayes, and decision trees were evaluated [19], revealing a strong dependence on manual feature engineering and an inability to capture complex temporal dependencies. In the same year, classification accuracy was further improved using SVMs and multilayer neural networks [20]. However, evaluation on controlled datasets limited the clinical robustness and scalability of these methods.

In addition to wavelet-based approaches, several studies have investigated advanced mathematical tools to further enhance epileptic seizure detection. In 2021, the Hilbert-Huang transform was employed in conjunction with an extreme learning machine to capture the nonlinear characteristics of EEG signals, leveraging instantaneous frequency and amplitude information at each time point [21]. However, the effectiveness of this approach was contingent on the stability of empirical mode decomposition, which remains highly sensitive to noise.

In 2022, an epileptic seizure detection method based on sparse representations on Riemannian manifolds was introduced [22], in which EEG signals were modeled using covariance matrices projected into a Hilbert space and reconstructed through a Stein kernel. Although theoretically robust, this approach is hindered by high computational and mathematical complexity, limiting its practical applicability in real-time processing scenarios.

In 2024, an enhanced DWT-based feature extraction framework was proposed to improve discriminative capability [23]; however, reliance on handcrafted features and the absence of an explicit mechanism for modeling long-term temporal dynamics fundamentally constrained its applicability. In the same year, an SVM-based method achieved 92% classification accuracy [24], though limitations in robustness and computational efficiency remained unresolved.

In 2025, a random forest-based epileptic seizure detection system was proposed, combining statistical and spectral features extracted from EEG signals [25]. The method achieved 95% accuracy with high precision and recall. Although robust and interpretable, it still relies on manual feature engineering, fails to capture long-term temporal dependencies, and remains limited to epileptic seizure detection tasks.

These limitations particularly the dependence on handcrafted features and the inability to model long-term temporal dependencies have motivated the shift toward deep learning-based methods, which enable automatic extraction of discriminative EEG representations without relying on manual feature engineering.

B. Deep Learning-Based Methods

To overcome the inherent limitations of traditional methods, deep learning-based approaches have emerged as a powerful alternative, enabling the automatic learning of hierarchical and discriminative representations directly from raw or minimally preprocessed EEG data. Among these approaches, Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and their bidirectional variants (BiLSTM) have been widely adopted for epileptic seizure detection owing to their ability to capture long-range temporal dependencies while improving both detection accuracy and model robustness.

In 2023, the Deep-EEG model (DCAE-ESD-BiLSTM), which combines a deep convolutional autoencoder with a BiLSTM network, was proposed [26], achieving an accuracy of 99.8% and an F1-score of 99.6% on the CHB-MIT and internal datasets, respectively. However, its high architectural complexity and substantial data requirements raised concerns regarding computational efficiency and practical deployment. In the same year, several RNN variants, including LSTM, BiLSTM, and stacked LSTM architectures, were evaluated on the CHB-MIT dataset for seizure prediction [13]. The results confirmed the superiority of BiLSTM, which achieved an accuracy of 89.78% and a specificity of 92.20%. Nevertheless, the performance remained inferior to that of more complex hybrid architectures, particularly for preictal epileptic seizure detection, highlighting the need for further optimization and complementary classification strategies.

In 2024, a CNN-based model leveraging both spatial and temporal characteristics of EEG signals was proposed for epileptic seizure detection [8], achieving an accuracy of 98.52% on a public benchmark dataset. Despite this strong performance, the approach remained computationally intensive and required large amounts of training data, while its robustness across subjects and under noisy EEG conditions was not fully established. Furthermore, its limited interpretability may hinder clinical adoption.

In 2025, a micro-capsule network was proposed for automated seizure detection from single-channel EEG signals [7], employing dimensional transformation and feature encoding to capture hierarchical spatial relationships. The method outperformed traditional CNNs and classical machine learning models; however, it did not explicitly capture long-term temporal dependencies and remained confined to epileptic seizure detection, without early prediction capability. A Graph Convolutional Neural Network (GCNN) was explored for epileptic seizure detection, explicitly encoding spatial dependencies between EEG electrodes while requiring minimal preprocessing [27]. This approach achieved 98.64% accuracy on the CHB-MIT dataset; nevertheless, its inter-subject generalization capability remained insufficiently validated. In addition, a Neural Ordinary Differential Equations (NODE)-based approach was proposed to model EEG signals in continuous time, integrating a feature extraction module and an attention-based fusion mechanism to improve classification performance [28]. However, this method entailed high computational cost and strong reliance on solver selection and associated hyperparameter tuning.

An end-to-end multi-view deep learning model was designed to detect brain anomalies from multichannel EEG signals by combining unsupervised reconstruction with supervised detection [29]. This framework leveraged an autoencoder to capture inter-channel correlations and incorporated a channel selection mechanism to retain the most informative inputs. Nevertheless, this approach was constrained by high computational complexity, substantial annotated data requirements, limited interpretability, noise sensitivity, and poor generalization to unseen populations or clinical conditions.

Overall, despite their strong predictive performance, deep learning-based approaches continue to suffer from high computational complexity, substantial data requirements, and insufficient robustness to EEG noise and inter-subject variability with clinical interpretability remaining a persistent challenge.

C. Hybrid Approaches

The growing complexity of EEG signals, combined with the limitations of conventional machine learning models, has motivated the development of hybrid approaches. These methods integrate convolutional, recurrent, and signal-processing components to jointly improve epileptic seizure detection accuracy and robustness.

In 2023, several hybrid frameworks were proposed. A fuzzy logic-based model was introduced to handle EEG nonlinearities [30]. Its shallow architecture, however, restricted its ability to capture long-term temporal dependencies. A 1D-CNN–BiLSTM model combined with *K*-means Synthetic Minority Over Sampling Technique (SMOTE) for class balancing was also proposed [31]. Although detection performance improved, the method remained computationally expensive and sensitive to residual class imbalance, limiting its scalability.

Multiple architectures exploiting Continuous Morlet Wavelet Transform (CMWT)-based scalograms were integrated into CNN, CNN-RNN, CNN-LSTM, and CNN-LSTM-LSTM frameworks [32]. Their high complexity, however, constrained applicability in real-time settings. A multichannel CNN-BiLSTM model augmented with an attention mechanism was proposed for classifying normal, preictal, and ictal EEG states [9]. Evaluated on CHB-MIT and UCI datasets, it achieved 94.83% accuracy. However, extensive preprocessing and high computational cost limited its practical deployment.

A multidimensional CNN-BiLSTM framework achieved 99.61% accuracy and 97.42% sensitivity on the UCI Epilepsy dataset [33]. Despite strong performance, it relied heavily on high-quality preprocessing, raising robustness concerns in clinical settings. An optimized Conv1D-LSTM model with dropout regularization and chi-square feature selection reached 99.3% accuracy [1]. Yet its dependence on careful preprocessing and feature engineering constrained generalizability across datasets.

In 2024, hybrid approaches further refined preprocessing pipelines and model architectures. A ResBiLSTM architecture combining residual CNN blocks with BiLSTM layers was proposed for joint spatial feature extraction and temporal modeling [13], achieving accuracies between 98.88% and 100% on the Bonn dataset. Nevertheless, the increased architectural complexity of this model remained a notable limitation. A DeepRNN-SGCN framework was investigated for the processing of non-stationary, low-amplitude EEG signals through time-frequency representations [34], reaching 99% accuracy on both the CHB-MIT and TUH datasets. Despite this performance, the framework remained computationally demanding and highly sensitive to data quality.

In 2025, a hybrid CNN-LSTM model combining spatial feature extraction and temporal recognition achieved 98.5% accuracy with a false positive rate of 1.06% on the CHB-MIT dataset [35]. Despite this effectiveness, computational cost continued to hinder real-time deployment. A complementary method based on Archimedean spiral coding (ASC) and a Swin Transformer-based CNN was proposed for EEG-based epileptic seizure detection [36], encoding EEG signals into 3D representations prior to vision transformer-based classification. In a separate study, CNNs and vision transformers were applied to time-frequency images generated by the tunable Q-Wavelet Transform (TQWT) [37], achieving 98.85% accuracy on CHB-MIT. The approach, however, distorted temporal EEG dynamics and lacked seizure prediction capability.

A hybrid framework integrating adaptive preprocessing, nonlinear entropic feature extraction, and dimensionality reduction was also reported [38]. The heavy reliance on manual feature engineering and hyperparameter tuning, however, limited generalization across EEG datasets. An STFT-based framework was proposed for time-frequency feature extraction, followed by classification using pretrained deep learning models [39]. The fixed time-frequency resolution inherent to STFT, however, restricted the framework's capacity to represent the nonstationary dynamics of EEG signals. Finally, an optimized deep residual network was proposed for epileptic seizure detection [40], although high architectural complexity and limited cross-dataset generalization remained key obstacles to real-world applicability.

Overall, hybrid approaches have demonstrated strong effectiveness in epileptic seizure detection. They nonetheless share persistent limitations, namely high computational complexity, strong dependence on preprocessing quality, and potential distortion of EEG temporal dynamics. These factors continue to challenge their deployment in real-time clinical environments, particularly for seizure prediction tasks.

D. Motivation for the Proposed BiLSTM-RVM Architecture

The epileptic seizure detection methods reviewed in Subsections A, B, and C present several important limitations that collectively motivate the present study.

Traditional approaches (Subsection A) primarily rely on the manual extraction of temporal and spectral features from EEG signals. These methods commonly employ classical classifiers such as Support Vector Machines (SVMs), *k*-Nearest Neighbors (*k*-NN), and random forests, as well as signal transformation techniques including the Discrete Wavelet Transform (DWT), Continuous Morlet Wavelet Transform (CMWT), Hilbert-Huang transform, and Riemannian manifold-based representations. Despite their effectiveness under controlled experimental conditions, these approaches exhibit four major limitations: strong dependence on handcrafted features, limited ability to capture long-range temporal dependencies, sensitivity to noise and inter-subject variability, and, in more advanced implementations, increased computational and parametric complexity. Moreover, most traditional methods are restricted to epileptic seizure detection and provide limited capability for reliable seizure prediction.

Deep learning-based approaches (Section II-B) enable the automatic extraction of hierarchical representations directly from raw or minimally preprocessed EEG signals. Architectures such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Bidirectional LSTM (BiLSTM), Convolutional Neural Networks (CNNs), capsule networks, and Graph Convolutional Neural Networks (GCNNs) have been extensively investigated. Although these models improve temporal dependency modeling and generally outperform conventional machine learning techniques, they remain computationally intensive, data-dependent, and sensitive to noise and inter-subject variability. In addition, their clinical interpretability remains limited, and several architectures still struggle to effectively model long-range temporal dependencies or provide reliable seizure prediction.

Hybrid approaches (Section II-C) combine multiple architectural components to further enhance epileptic seizure detection performance. However, these frameworks generally rely on complex multi-stage pipelines, extensive preprocessing procedures, and substantial computational resources. Their performance is highly dependent on signal quality, while certain transformation techniques may alter the intrinsic temporal dynamics of EEG signals. Furthermore, most hybrid frameworks remain focused primarily on epileptic seizure detection, with limited predictive capability and insufficient clinical generalizability. The frequent reliance on softmax-based classifiers may also reduce probabilistic interpretability and increase the risk of overfitting in high-dimensional feature spaces.

Overall, despite significant progress in EEG-based epileptic seizure analysis, existing approaches still face important challenges related to computational efficiency, robustness, interpretability, generalization capability, and reliable seizure prediction. These limitations highlight the need for more efficient and clinically adaptable frameworks capable of simultaneously improving detection accuracy, temporal modeling, and predictive performance.

To address these limitations, a BiLSTM-RVM architecture is proposed, combining the strong temporal modeling capability of Bidirectional Long Short-Term Memory (BiLSTM) networks with the sparsity and probabilistic interpretability of Relevance Vector Machines (RVM). This hybrid model aims to enhance robustness, reduce dependency on intensive preprocessing, and better capture long-term temporal dependencies, while simultaneously lowering computational complexity and enabling real-time seizure detection in clinical environments.

The proposed BiLSTM-RVM framework offers the following key advantages:

- Effective capture of temporal dependencies BiLSTM: networks model temporal relationships in both forward and backward directions, thereby addressing a fundamental limitation of conventional and hybrid approaches that fail to fully exploit long-range temporal information. This capability enhances the identification of preictal states and supports epileptic seizure prediction rather than mere detection.
- Reduced decision complexity and improved interpretability: The RVM provides a sparse probabilistic framework that yields more interpretable decisions than softmax-based classifiers. Furthermore, it reduces the number of relevant basis functions compared to Support Vector Machines (SVMs), thereby improving overall computational efficiency.
- Improved robustness to noisy and imbalanced data: By integrating BiLSTM-based feature learning with RVM probabilistic classification, the proposed model mitigates overfitting and enhances robustness to class imbalance and signal noise, while substantially reducing reliance on manual feature engineering.
- Balanced trade-off between performance and computational cost: Unlike computationally intensive architectures such as CNN-Transformer or DeepRNN-GCN models, the proposed approach constitutes a lightweight and modular framework that sustains high predictive performance while reducing computational overhead.
- Facilitation of real-time clinical deployment: Owing to its reduced architectural complexity and efficient feature learning strategy, the proposed model is more suitable for real-time implementation in clinical or embedded systems compared to deeper hybrid architectures.

III. MATERIALS AND METHODS

This section describes the proposed BiLSTM-RVM hybrid model for automatic epileptic seizure detection from Electroencephalogram (EEG) signals. The methodology encompasses the dataset description, the preprocessing pipeline, the BiLSTM-based feature extraction process, and the RVM-based classification strategy. The hyperparameter configuration of the proposed model is further detailed, along with the evaluation protocols employed to assess its performance.

A. Dataset and Participants

The experiments are conducted using the publicly available Epileptic Seizure Recognition dataset from the UCI Machine Learning Repository, a widely adopted benchmark for EEG-based epileptic seizure detection. The dataset comprises 500 EEG recordings organized into five classes, each containing 100 signals corresponding to distinct brain states. Each recording spans 23.6 s and consists of 4097 samples acquired from a single EEG channel. Since the dataset is fully anonymized and publicly accessible, no demographic or clinical information is available; accordingly, no ethical approval is required. The CHB-MIT Scalp EEG Database is further employed to evaluate the generalization capability of the proposed model across subjects and recording conditions.

B. Data Preparation and Pre-processing

1) Understanding of the data

This study employs raw EEG recordings, each comprising 4097 samples. Each recording captures brain activity from a single subject and is assigned to one of five categories corresponding to distinct brain states. The primary objective is epileptic seizure detection. To this end, the classification task is reformulated as a binary problem by retaining the epileptic class and merging the four remaining non-epileptic classes into a single category. This formulation allows the model to focus on seizure episode identification while reducing decision complexity.

2) Acquisition of the data

The EEG recordings are imported into the Python environment using the pandas and NumPy libraries. Exploratory Data Analysis (EDA) is then performed to examine the dataset structure, including its dimensions and variable types, identify irrelevant columns such as automatic indices and record identifiers, and verify the numerical consistency and integrity of the data. The cleaning stage involves removing non-informative columns, retaining only the numerical variables representing EEG measurements corresponding to 178 principal components and screening for missing values, inconsistencies, and duplicate entries.

3) Pre-processing

The EEG signal preprocessing involves the following steps:

a) Segmentation

Each EEG recording is divided into non-overlapping segments of 178 samples, as described in the previous section.

b) Normalization

Each segment is normalized using min-max scaling to map all values to the [0, 1] range, ensuring consistent feature scaling and stable convergence during model training.

c) Class relabeling

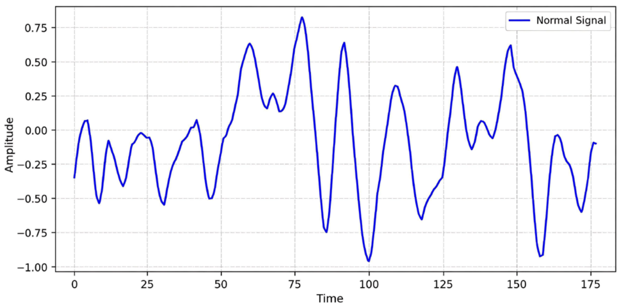
EEG segments are assigned to one of two classes:

- Class 0—Non-epileptic: aggregating the four non-epileptic categories
- Class 1—Epileptic: segments corresponding to ictal episodes

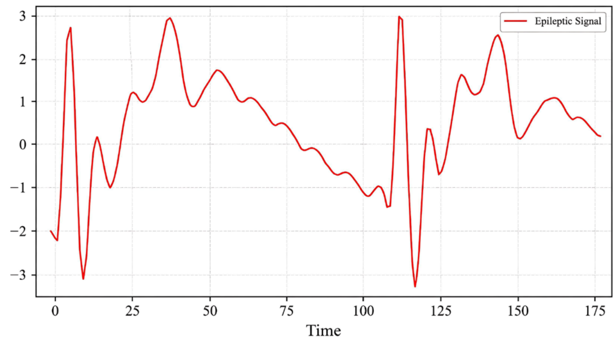
d) Data cleaning and quality control

Non-informative columns are removed, missing values are verified, and duplicate or inconsistent entries are discarded.

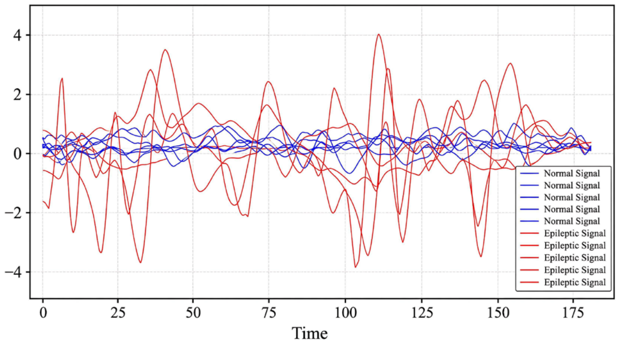
Fig. 1 illustrates the temporal representation of EEG signals across both classes. The non-epileptic signal (Fig. 1(a)) exhibits relatively regular oscillatory activity with moderate amplitude variations. In contrast, the epileptic signal (Fig. 1(b)) is characterized by high-frequency oscillations and markedly elevated amplitudes. The superimposition of both signals (Fig. 1(c)) highlights their temporal divergence, reflecting the non-stationary dynamics of EEG signals during ictal events. These amplitude and frequency variations constitute discriminative temporal features that the BiLSTM model exploits to learn the dependencies between epileptic and non-epileptic states.



(a)



(b)



(c)

Fig. 1. Visualization of the dataset: (a) Non-epileptic EEG signal; (b) Epileptic EEG signal; (c) The superimposition of the two signals.

C. Proposed BiLSTM-RVM Architecture

The proposed BiLSTM-RVM architecture is a hybrid Deep Learning framework designed for automatic epileptic seizure detection from EEG signals. It integrates the temporal feature extraction capability of Bidirectional Long Short-Term Memory (BiLSTM) networks with the Bayesian sparse probabilistic classification of Relevance Vector Machines (RVM).

1) BiLSTM-based feature extraction

The BiLSTM network is employed as a feature extractor to model the temporal dynamics of EEG signals. Unlike standard LSTM networks, a BiLSTM processes the input sequence in both forward and backward directions, thereby capturing past and future temporal dependencies simultaneously. At each time step, the LSTM cell updates its internal state through gating mechanisms governed by sigmoid and hyperbolic tangent (tanh) activation functions. This mechanism ensures the selective preservation of relevant information over extended temporal sequences.

A Long Short-Term Memory (LSTM) cell is a specialized recurrent neural network (RNN) unit designed to model long-term dependencies while mitigating the vanishing gradient problem. Its operation is governed by three principal gates:

a) The forget gate

The forget gate determines which information from the previous hidden state is retained or discarded. The previous hidden state and the current input are concatenated and passed through a sigmoid activation function, producing output values in the range [0, 1], as expressed in Eq. (1):

$$F_t = \sigma\left(W_f[H_{t-1}, X_t] + b_f\right) \tag{1}$$

where F_t denotes the forget gate activation vector, $\sigma(\cdot)$ represents the sigmoid activation function, W_f and b_f are the weight matrix and bias term, respectively, H_{t-1} is the previous hidden state, and X_t is the current input.

b) The input gate

The input gate regulates the flow of new information into the cell state. The current input and the previous hidden state are concatenated and processed through two parallel operations. First, a sigmoid activation function is applied to determine which new information is selectively retained, yielding the input gate activation vector. Second, a hyperbolic tangent (tanh) activation function generates a candidate cell vector representing the new content to be potentially added. The element-wise product of these two vectors is then incorporated into the cell state, thereby updating the long-term memory of the network. The input gate operations are formulated in Eqs. (2)–(4).

$$K_t = \tanh(W_k[H_{t-1}, X_t] + b_k) \tag{2}$$

$$I_t = \sigma(W_i[H_{t-1}, X_t] + b_i) \tag{3}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot K_t \tag{4}$$

where, X_t is the input at time t; H_{t-1} is the previous hidden state; C_{t-1} is the previous cell state; K_t is the candidate vector generated by the *tanh;* I_t is the input gate vector obtained via the sigmoid function; σ denotes the sigmoid function; $\odot$ is hadamard product.

c) The output gate

The output gate determines the hidden state transmitted to the next time step by selectively filtering the internal cell state. The current input and the previous hidden state are first processed through a sigmoid activation function, yielding the output gate activation vector (Eq. (5)). Concurrently, the updated cell state is passed through a tanh activation function to normalize its values. The element-wise product of these two vectors produces the new hidden state h_t, which is subsequently propagated to the next LSTM unit (Eq. (6)). This operation is repeated at each time step across the full input sequence.

$$O_t = \sigma(W_O[H_{t-1}, X_t] + b_O) \tag{5}$$

$$H_t = O_t \odot \tanh(C_t) \tag{6}$$

The LSTM unit is shown in Fig. 2. A BiLSTM architecture is employed to capture temporal dependencies in EEG sequences across both forward and backward directions. By processing each EEG segment from past to future and from future to past simultaneously, the BiLSTM yields richer temporal representations than conventional unidirectional recurrent models.

The proposed feature extractor consists of stacked BiLSTM layers, in which the hidden states of the forward and backward LSTM networks are concatenated to form a unified feature representation. This bidirectional formulation enables the model to learn discriminative temporal features that more effectively characterize ictal and non-ictal EEG dynamics.

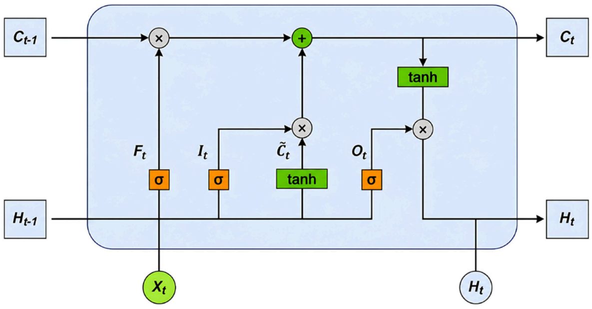


Fig. 2. Architecture of an LSTM cell [19].

Eqs. (7) and (8) express the state updates of the forward and backward LSTM networks, respectively, while Eq. (9) defines the resulting bidirectional output.

$$\overrightarrow{h}_t = W_{x\overrightarrow{h}_t} x_t + W_{\overrightarrow{h}_t\overrightarrow{h}_t}\overrightarrow{h}_{t-1} + \overrightarrow{b}_{\overrightarrow{h}_t} \tag{7}$$

$$\overleftarrow{h}_t = W_{x\overleftarrow{h}_t} x_t + W_{\overleftarrow{h}_t\overleftarrow{h}_t}\overleftarrow{h}_{t-1} + \overleftarrow{b}_{\overleftarrow{h}_t} \tag{8}$$

$$y_t = W_{y\overrightarrow{h}_t} x_t + W_{y\overleftarrow{h}_t}\overleftarrow{h} + b_y \tag{9}$$

where W denotes the weight matrices governing the connections between the input x_t and the previous hidden state h_{t-1}, and b represents the corresponding bias terms. This bidirectional formulation enables richer temporal feature learning from EEG sequences.

The high-level temporal features extracted by the BiLSTM network are subsequently passed to a Relevance Vector Machine (RVM) for classification. The RVM receives the feature vectors produced by the BiLSTM and performs binary classification between ictal and non-ictal EEG segments.

2) Relevance vector machine classifier

The Relevance Vector Machine (RVM) is a supervised Bayesian kernel-based learning model that enforces sparsity by automatically selecting a subset of relevance vectors from the training data. It can be regarded as the probabilistic counterpart of the Support Vector Machine (SVM). Each model weight is associated with an individual variance hyperparameter inferred from the training data, driving most weights toward zero in the posterior distribution. This yields a highly sparse and compact predictive model with performance comparable to, or in some cases surpassing, that of a conventional SVM.

Formally, for a given training dataset, the predictive function is defined in Eq. (10),

$$y(x) = \sum_{j=1}^{N} w_j K(x, x_j) + b \tag{10}$$

where $K(,)$ denotes a kernel function and the weights w_j are estimated within a Bayesian framework.

The relevance vectors correspond to training samples associated with non-zero weights. Unlike standard SVMs, the RVM yields probabilistic outputs, enabling confidence estimation in classification decisions a property of critical relevance in clinical decision-support systems. Fig. 3 illustrates the overall RVM architecture.

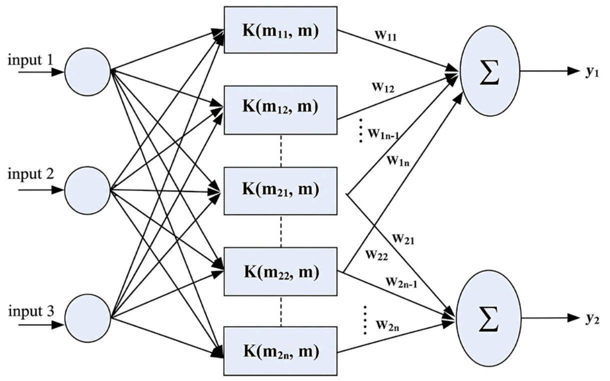


Fig. 3. Architecture of RVM [41].

In the proposed framework, the feature vector extracted by the BiLSTM network is passed to the Relevance Vector Machine (RVM) for binary classification. Since the task is formulated as a binary problem, the RVM estimates the posterior probability of class membership under a Bernoulli likelihood model. The corresponding likelihood function is defined in Eq. (11).

$$p(t|w) = \prod_{i=1}^{N} \sigma\left(y(x_i)\right)^{t_i} [1 - \sigma(y(x_i))]^{1-t_i} \quad (11)$$

where, N denotes the number of samples, $y_i \in \{0,1\}$ represents the class label of the i-th sample, and $\prod(.)$ corresponds to the probability predicted by the RVM that the i-th sample belongs to class 1. The vector θ denotes the model parameters. The sigmoid function $\sigma(.)$is used to map the model output to the posterior probability $p(t_i = 1|x_i, \theta)$.

To maximize $p(t|w)$, the Maximum A Posteriori (MAP) approach is employed to estimate the weight vector w. Let α_i^* be the maximum a posteriori of the hyperparameter α_i. The MAP estimate of the weights, denoted as w_{MAP}, is obtained by maximizing the posterior distribution of the class labels given the input vectors.

This optimization is equivalent to maximizing the objective function defined in Eq. (12):

$$J(w) = \sum_{i=1}^{N} \log p(t_i|w_i) + \sum_{i=1}^{N} \log p(w_i|\alpha_i^*) \quad (12)$$

where the first term of the sum corresponds to the likelihood of the class labels and the second in the prior distribution on parameters w.

In the resulting solution, only the training samples associated with non-zero weight coefficients w_i, referred to as relevance vectors, contribute to the final decision function. The gradient of the objective function J with respect to the weight vector w is expressed in Eq. (13):

$$\nabla J = -A^T w - \Phi^T (f - t) \quad (13)$$

where $f = \left[\sigma\left(y(x_1)\right), .., \sigma\left(y(x_N)\right)\right]^T$ denotes the vector of sigmoid-transformed model outputs for all training samples. The design matrix $\Phi\,(*)$ has elements $\phi_{ij} = K\left(x_i, x_j\right)$, where $K(*)$ denotes the kernel function.

The Hessian matrix of the objective function J with respect to the weight vector w is defined in Eq. (14):

$$H = \nabla^2 J(J) = -\left(\Phi^T B \Phi + A'\right), \quad (14)$$

where $B = diag(\beta_1, \dots, \beta_N)$ and $\beta_i = \sigma\big(y(x_i)\big)\big[1 - \sigma\big(y(x_i)\big)\big]$.

The posterior distribution is then approximated around w_{MAP} by a Gaussian distribution, whose mean and covariance are given in Eqs. (15) and (16), respectively.

$$\sum = -\left(H|w_{MAP}\right)^{-1} \quad (15)$$

$$\mu = \sum \Phi^T B t \quad (16)$$

Fig. 4 illustrates the overall architecture of the proposed BiLSTM–RVM model.

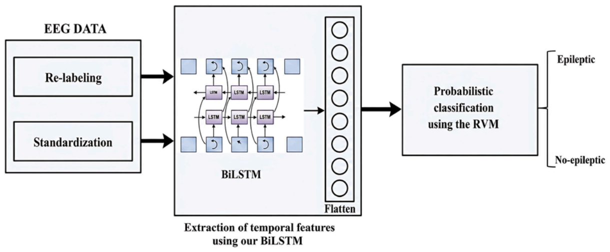


Fig. 4. Architecture of the proposed method.

D. Hyperparameters of the BiLSTM-RVM Model

The hyperparameters of the proposed BiLSTM–RVM model are summarized in Table I. The architecture is designed to capture long-range temporal dependencies while maintaining computational efficiency. Dropout regularization and batch normalization are incorporated to enhance generalization and mitigate overfitting. The Adam optimizer, configured with a learning rate of 0.001, ensures stable convergence, and early stopping is applied as an additional regularization measure. The RVM classifier, equipped with a Radial Basis Function (RBF) kernel, enforces non-linear discriminability while preserving model sparsity and probabilistic interpretability.

TABLE I. HYPERPARAMETER CONFIGURATION OF THE PROPOSED BILSTM-RVM MODEL

Model Component	**Parameter**	**Value**
Architecture BiLSTM	Number of layers	8 (organized into 4 blocks of 2 layers)
	Unit-Block 1 &4	128 per direction (256 bidirectional units)
	Unit-Block 2 &3	64 per direction (128 bidirectional units)
	Recurrent activation	Sigmoid
	Activation	Tanh
	Dropout rate	0.3
	Batch normalization	Applied after each BiLSTM layer
	Dense layer	64 units, ReLU activation
	Feature Vector Dimension	64
	Total trainable parameters	~2.1 million
	Input shape	178 time steps, 1 channel
Training Configuration	Optimizer	Adam (learning rate = 0.001)
	Loss function	Binary cross-entropy
	Batch size	32
	Maximum epochs	100
	Early stopping	patience = 10 epochs
	Best model saving	enabled (save_best_only = True)
	Execution environment	NVIDIA T4 GPU (Google Colab)
RVM Classifier	Kernel	Radial Basis Function (RBF)
	Gamma parameter	“scale” (1 / (n_features × Var(X)))
	Maximum EM iterations	500
	Convergence tolerance	1×10^{-3}
	Training subset size	2.000 samples (stratified sampling)
	Number of relevance vectors	41 (sparsity = 2.05%)

E. Evaluation Protocol

The performance of the proposed BiLSTM-RVM model is evaluated using a comprehensive experimental protocol designed to assess classification accuracy, generalization capability, and robustness. First, intra-database evaluation is conducted using a standard train/test split to establish baseline performance.

Second, inter-subject generalization is assessed using a Leave-One-Subject-Out (LOSO) cross-validation strategy, where data from each subject is sequentially used as the test set while the remaining subjects are used for training. Third, cross-database validation is performed by evaluating the model trained on one dataset on an independent dataset to assess its ability to generalize across different acquisition conditions. In addition, robustness is analyzed by introducing additive Gaussian noise to EEG signals to simulate real-world recording artifacts.

Finally, performance is measured using standard classification metrics, including accuracy, precision, recall, F1-score, specificity, and the area under the ROC curve (AUC).

In summary, the proposed BiLSTM-RVM framework integrates bidirectional temporal feature extraction with sparse probabilistic classification, evaluated under a rigorous multi-level validation protocol. The following section presents and discusses the experimental results obtained under this protocol.

IV. RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of the proposed BiLSTM-RVM model for epileptic seizure detection. The model is first assessed on the primary dataset and benchmarked against state-of-the-art methods. Generalization capability is then examined through a subject-independent Leave-One-Subject-Out (LOSO) cross-validation protocol and further validated via cross-dataset experiments on an independent database.

An ablation study is conducted to quantify the contribution of each architectural component. Statistical significance tests and effect size analysis are subsequently performed to confirm the reliability of the reported results. Finally, robustness under noisy acquisition conditions and SHAP-based explainability analysis are carried out to assess the model’s stability and interpretability in realistic clinical scenarios.

A. Experimental Setup

The experiments are conducted using the Epileptic Seizure Recognition Dataset from the UCI Machine Learning Repository, comprising 11,500 EEG segments of dimensionality (22×178). The data are normalized, segmented, and partitioned into 80% for training and 20% for testing, with class distribution preserved throughout.

The proposed hybrid model integrates a BiLSTM network for bidirectional temporal feature extraction with an RVM classifier for probabilistic decision-making. Dropout regularization and the Adam optimizer are employed to enhance generalization and ensure stable convergence. A fixed random seed is set to guarantee full reproducibility across all experiments.

All experiments are implemented in Python, using TensorFlow/Keras for the BiLSTM component and a dedicated probabilistic inference library for the RVM classifier. Computations are performed on a Dell Latitude E6410 workstation equipped with an Intel Core i7 processor and 8 GB of RAM.

B. Evaluation Metrics

Epileptic seizure detection is formulated as a binary classification problem distinguishing ictal from non-ictal EEG segments. Model performance is assessed using four standard evaluation metrics: accuracy, precision, recall (sensitivity), and F1-score.

In this context, a True Positive (TP) denotes an epileptic segment correctly identified as ictal; a True Negative (TN) denotes a non-epileptic segment correctly identified as non-ictal; a False Positive (FP) denotes a non-epileptic segment incorrectly classified as ictal; and a False Negative (FN) denotes an epileptic segment incorrectly classified as non-ictal. Eqs. (17)–(20) define the corresponding performance metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{20}$$

C. Performance Evaluation of the Proposed BiLSTM-RVM Model on the Primary Dataset

The ROC curve depicted in Fig. 5 demonstrates the strong discriminative capability of the proposed model between ictal and non-ictal EEG segments, achieving an AUC of 99.97%. The confusion matrix presented in Fig. 6 corroborates this performance: 1145 out of 1150 test samples are correctly classified, yielding an accuracy of 99.57%, a precision of 98.63%, a recall of 99.08%, and an F1-score of 98.84%.

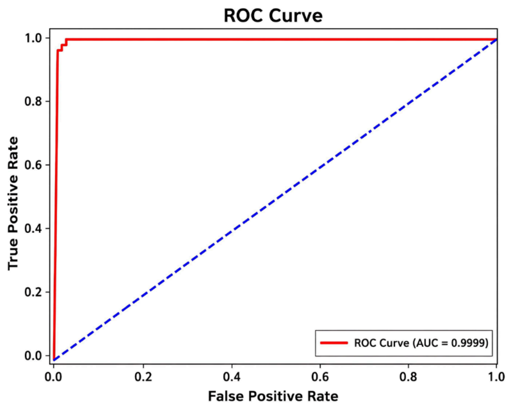


Fig. 5. ROC Curve of BiLSTM-RVM.

The ROC curve confirms excellent discriminative performance, with an AUC of 0.9997, indicating near-perfect separation between ictal and non-ictal EEG signals. The confusion matrix corroborates this outcome: 929 non-epileptic samples and 216 epileptic segments are correctly classified, with only 3 false positives and 2 false negatives. This yields an overall accuracy of 99.57%, with consistently high precision, recall, and F1-score across both classes. These results highlight the reliability and robustness of the proposed BiLSTM-RVM model, particularly its capacity to detect seizures with minimal missed events and false alarms.

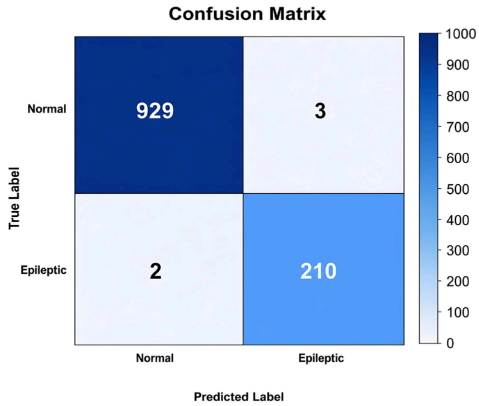


Fig. 6. Confusion matrix of BiLSTM-RVM.

To evaluate the competitiveness of the proposed model, a comparative analysis is conducted against recent methods, including hybrid and attention-based architectures. Table II summarizes the classification performance of representative models, namely RNN-BiLSTM, CNN-BiLSTM, DT-CWT-1DCNN-BiLSTM, ResBiLSTM, CNN-LSTM, and BiLSTM. Additional experiments involving CNN-LSTM and Attention-BiLSTM are further conducted to provide a more comprehensive and controlled evaluation benchmark.

As reported in Table II, the proposed BiLSTM-RVM model achieves an accuracy of 99.57%, outperforming conventional architectures such as BiLSTM (89.78%) and ResBiLSTM (95.03%), as well as hybrid models including CNN-LSTM (99.30%) and Attention-BiLSTM (98.78%). Furthermore, the proposed model exhibits well-balanced performance across all evaluation metrics, attaining a precision of 98.63%, a sensitivity of 99.08%, and an F1-score of 98.84%, thereby confirming its robustness and effectiveness for epileptic seizure detection.

These results are attributed to the complementary integration of the BiLSTM network, which captures long-range temporal dependencies in EEG signals, and the RVM classifier, which enhances generalization through sparse probabilistic inference.

To further assess generalization across unseen subjects, a Leave-One-Subject-Out (LOSO) cross-validation is subsequently conducted.

TABLE II. COMPARISON OF THE OBTAINED RESULTS WITH EXISTING LITERATURE

Authors	Method	Database	Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
Samee *et al.* [3]	RNN-BiLSTM	UCI-Epileptic	98.40	98.10	98.30	-
Ma *et al.* [9]	CNN-Bi-LSTM	UCI-Epileptic	94.83	94.84	94.84	92.26
KR *et al.* [10]	CNN-Bi-LSTM	UCI-Epileptic	99.61	99.13	97.42	-
Aslan and Bingöl [11]	DT-CWT-1DCNN-BiLST	Epileptic Seizure Recognition Dataset	96.12	97.05	95.61	96.87
Zhao *et al.* [12]	ResBiLSTM	Epileptic Seizure Recognition Dataset	95.03	-	-	95.03
Omar and El-Hafeez [1]	CNN-LSTM	Epileptic Seizure Recognition Dataset	99.30	-	-	-
Bajaj and Sharma [13]	BILSTM	Epileptic Seizure Recognition Dataset	89.78	91.81	92.20	89.53
Proposed method	CNN-LSTM	Epileptic Seizure Recognition Dataset	99.13	97.71	97.71	97.71
Proposed method	Attention-BiLSTM	Epileptic Seizure Recognition Dataset	98.78	97.12	96.19	96.65
Proposed method	**BILSTM +RVM**	**Epileptic Seizure Recognition Dataset**	**99.57**	**98.63**	**99.08**	**98.84**

D. Result Leave-One-Subject-Out (LOSO)

To rigorously assess the generalization capability of the proposed model under subject-independent conditions, a Leave-One-Subject-Out (LOSO) cross-validation protocol is adopted. The UCI dataset comprises 23 subjects (X1–X23), each represented by 500 EEG segments.

In each fold, all segments from one subject are reserved for testing, while the remaining 22 subjects constitute the training set. This strict partitioning prevents intra-subject data leakage and ensures an unbiased evaluation of the model's generalization ability.

The evaluation pipeline consists of three sequential stages: (1) bidirectional temporal feature extraction from the training data using the BiLSTM network; (2) training of the RVM classifier on a stratified subset of 2000 samples to limit kernel computation cost; and (3) inference on the held-out subject using the extracted feature representations.

To ensure experimental consistency across all folds, the RVM hyperparameters are held fixed throughout: RBF kernel, γ = scale, maximum iterations = 500.

1) Performance per subject

Fig. 7 illustrates the evolution of accuracy and F1-score across all subjects. Both metrics exhibit minimal inter-subject variation, confirming the strong generalization capability of the proposed model despite inherent inter-subject variability. The near-perfect overlap between the two curves further indicates an optimal balance between precision and recall, with no observable class bias.

The proposed model achieves consistently high and stable classification performance across all folds, attaining a mean accuracy of 99.40% ± 0.39% and a mean F1-score of 99.38% ± 0.03%, thereby confirming both its robustness and reliability under subject-independent conditions.

Marginal performance decreases are observed for subjects X10 and X14, which may be attributed to intrinsic inter-subject variability or local distributional shifts in EEG recordings. Nevertheless, these fluctuations remain negligible and do not significantly affect the overall classification performance of the proposed model.

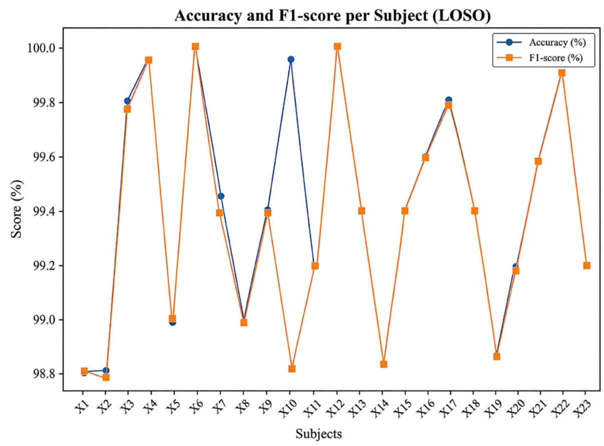


Fig. 7. Accuracy and F1-score per subject (LOSO).

2) Analysis of model sparsity via relevance vectors

Fig. 8 presents the distribution of Relevance Vectors (RVs) per subject. Unlike classification performance, the number of RVs exhibits considerable variability across subjects, ranging from 3 to 267.

This variability reflects the adaptive nature of the RVM, which dynamically adjusts the number of active basis functions according to the intrinsic structure of each subject's data. Subjects associated with a small number of RVs (e.g., X5 and X14) suggest locally separable data distributions requiring only a minimal set of representative vectors. Conversely, subjects with a larger number of RVs (e.g., X11–X13) exhibit more complex or heterogeneous distributions, necessitating a richer representation of the decision boundary.

Despite this variability, overall sparsity remains substantial, with an average of approximately 9.75% relevance vectors relative to the total number of training samples. This confirms a key advantage of the RVM framework: its capacity to yield compact and interpretable models, in contrast to computationally intensive approaches such as Support Vector Machines (SVMs).

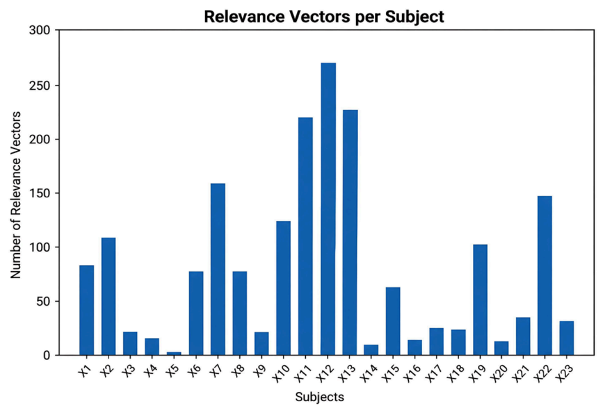


Fig. 8. Distribution of the number of Relevance Vectors (RVs) per subject.

3) Relationship between model complexity and performance

Fig. 9 illustrates relationship between accuracy and the number of Relevance Vectors (RVM) for each subject under the LOSO validation framework. Each point represents an individual subject, while the line corresponds to the linear regression fit. The results reveal a weak and non-significant correlation (Pearson's $r = 0.15$, $p > 0.05$), indicating that the model's performance is largely independent of its complexity.

The statistical analysis of the relationship between Accuracy and the number of relevance vectors reveals a Pearson correlation of $r = 0.15$ ($p = 0.49$) and a Spearman correlation of $\sigma = 0.17$ ($p = 0.45$). These results indicate a very weak and statistically non-significant positive correlation ($p > 0.05$), confirming the absence of any linear or monotonic relationship between model complexity and predictive performance.

The estimated regression line is expressed as $Accuracy = 0.00082 \times RVM + 99.346$.

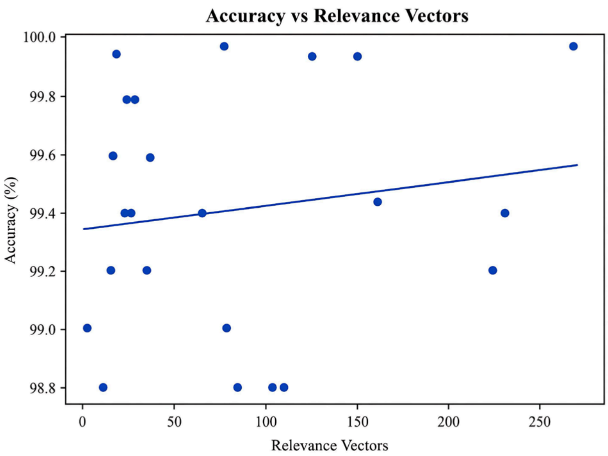


Fig. 9. Accuracy vs relevance vectors.

The extremely small slope (0.00082) indicates that even a substantial increase in the number of relevance vectors leads to only a negligible performance gain. This observation confirms that the model performance remains practically constant regardless of the level of model complexity.

3) Discussion and implications

The experimental results demonstrate that the proposed BiLSTM-RVM model satisfies three key requirements for real-world clinical deployment: high classification accuracy, strong generalization validated through the LOSO protocol, and computational efficiency enabled by sparse representation. The weak correlation observed between model complexity and predictive performance further confirms the robustness of the RVM and its inherent resistance to overfitting. This behavior is consistent with the principles of sparse Bayesian learning, whereby predictive performance is governed primarily by the quality rather than the quantity of retained relevance vectors. Collectively, these findings support the RVM as a competitive and interpretable alternative to conventional kernel-based classifiers, particularly in applications demanding accuracy, transparency, and computational efficiency.

E. External Validation (Cross-Dataset)

To further assess the generalization capability of the proposed model, additional experiments are conducted on the CHB-MIT Scalp EEG Database. Unlike the primary dataset, which contains relatively clean and well-structured EEG segments, the CHB-MIT database comprises long-duration, multi-channel recordings acquired under real clinical conditions, introducing substantially greater inter-subject variability and signal complexity.

The CHB-MIT database includes scalp EEG recordings from 24 pediatric patients with intractable epilepsy, sampled at 256 Hz and acquired using 23 bipolar channels following the international 10–20 electrode placement system. Its high inter-subject variability renders it a challenging benchmark, particularly well-suited for evaluating the robustness and generalization capability of epileptic seizure detection models.

The preprocessing pipeline comprises a band-pass filter (0.5–100 Hz) to retain neurophysiologically relevant frequency components, followed by a notch filter at 60 Hz to suppress power-line interference. The filtered signals are subsequently normalized using Z-score standardization, and amplitude values are clipped at $\pm 5\sigma$ to mitigate the influence of artifacts and outliers. Class imbalance is partially addressed through controlled under-sampling of the non-ictal class, yielding a final seizure-to-normal ratio of 1:2.

Figs. 10 and 11 and present the ROC curve and confusion matrix obtained on the CHB-MIT test set, respectively, while Table III summarizes the corresponding classification performance metrics. The proposed model achieves an accuracy of 93.95%, a seizure-class precision of 97.37%, a specificity of 98.87%, a recall of 84.09%, an F1-score of 90.24%, and an AUC of 96.80%, confirming strong discriminative capability between ictal and non-ictal classes.

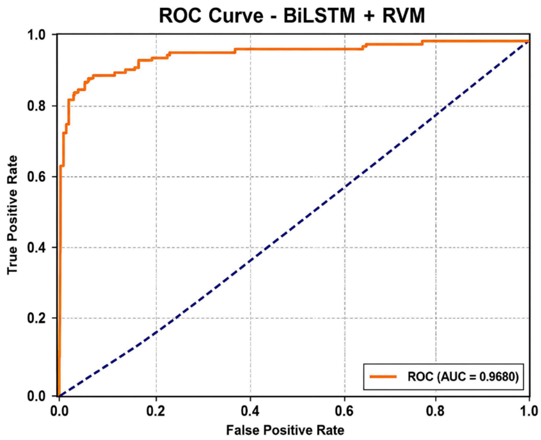


Fig. 10. ROC curve of the BiLSTM-RVM model on CHB-MIT.

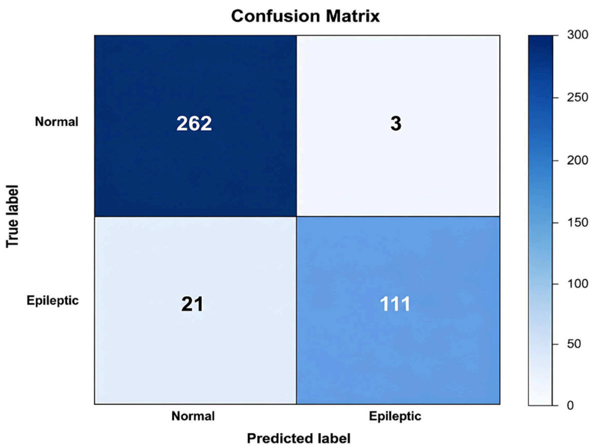


Fig. 11. BiLSTM-RVM confusion matrix on CHB-MIT.

TABLE III. CLASSIFICATION REPORT

Classe	Precision	Recall	F1-score	Support
Normal	0.9258	0.9887	0.9562	265
Epileptic	0.9737	0.8409	0.9024	132
Macro Average	**0.9497**	0.9148	**0.9293**	
Weighted Average	**0.9417**	**0.9395**	**0.9383**	

These results reveal a deliberately conservative classification behavior: among all windows predicted as seizures, only three correspond to false alarms a critical property in clinical monitoring systems. The comparatively lower recall reflects the inherent precision-recall trade-off, with 21 missed seizures out of 132 ictal events. Nevertheless, the high AUC and F1-score confirm a robust balance between detection sensitivity and precision for the minority class under subject-independent evaluation.

From a clinical perspective, high precision is generally preferable to maximizing recall, as repeated false alarms may induce alarm fatigue and caregiver desensitization. The conservative behavior of the proposed model therefore constitutes a favorable property for real-world medical deployment.

Table IV summarizes the performance of the proposed BiLSTM-RVM model on the Epileptic Seizure Recognition Dataset and CHB-MIT datasets, confirming its robustness and cross-dataset generalization capability.

TABLE IV. CROSS-DATASET EVALUATION OF THE PROPOSED EEG EPILEPTIC SEIZURE DETECTION MODEL

Datasets	Accuracy	Precision	Sensitivity	F1-score
Epileptic Seizure Recognition	99.57%	98.69%	99.08%	98.84%
CHB-MIT	93.95 %	97.37 %	84.09%	90.24%

The observed performance degradation across datasets is expected, attributable to the distribution shift between controlled laboratory signals and real clinical recordings, as well as differences in channel configuration and inter-subject variability. Nevertheless, the proposed model maintains high precision and specificity, confirming its robustness and low false-alarm rate under cross-dataset conditions.

Overall, these results indicate that the proposed BiLSTM-RVM model learns generalizable temporal representations rather than dataset-specific patterns. Its sustained performance on an independent database confirms its transferability and suitability for real-world epileptic seizure detection.

An ablation study is subsequently conducted to quantify the individual contribution of each architectural component of the proposed model.

F. Ablation Study

To rigorously assess the contribution of the classification stage, six architectures are compared

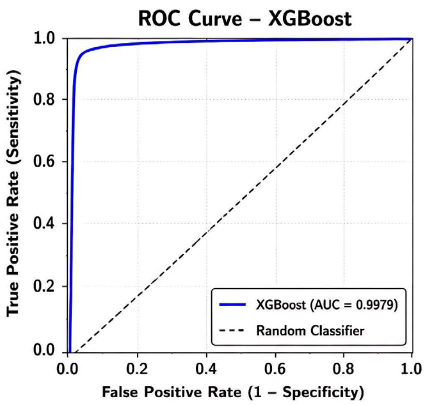
- (a) XGBoost
- (b) BiLSTM + Softmax
- (c) BiLSTM + SVM
- (d) CNN
- (e) LSTM
- (f) BiLSTM + RVM (proposed)

An identical experimental protocol is applied across all six architectures to ensure that any observed performance difference is solely attributable to the choice of classifier. Specifically, 80% of the EEG dataset is allocated for training and the remaining 20% is reserved for testing.

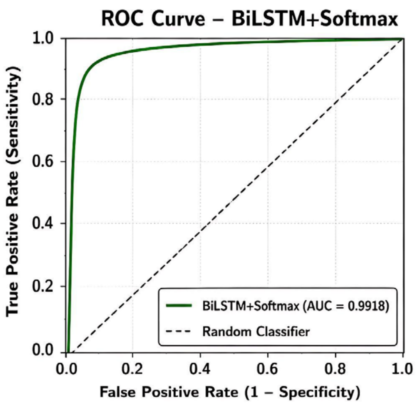
The comparative ROC curves are presented in Fig. 12, whereas the confusion matrices corresponding to each evaluated model are consolidated in Fig. 13. Together, these complementary visualizations provide a comprehensive assessment of both the overall discriminative capability of the models and their classification behavior with respect to ictal and non-ictal EEG segments, thereby highlighting their ability to correctly distinguish epileptic from non-epileptic patterns.

The performance of the XGBoost classifier is illustrated by the ROC curve and the corresponding confusion matrix shown in Figs. 12(a) and 13(a), respectively. The confusion matrix indicates that 920 non-ictal and 207 ictal EEG segments are correctly classified, with only 5 false positives and 18 false negatives. The model achieves an overall accuracy of 98.0%, with precision and recall values exceeding 97% across most classes, confirming reliable classification capability. The ROC analysis yields an AUC of 0.9979, confirming excellent discriminative performance between ictal and non-ictal EEG signals. Nevertheless, the relatively higher number of false negatives suggests a reduced sensitivity to certain epileptic events, which may limit its suitability for clinical applications where maximizing seizure detection is of paramount importance.

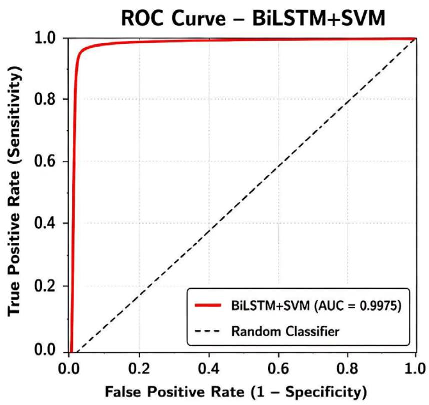
The performance of the BiLSTM + Softmax architecture is illustrated by the ROC curve and its corresponding confusion matrix, as presented in Figs. 12(b) and 13(b), respectively.



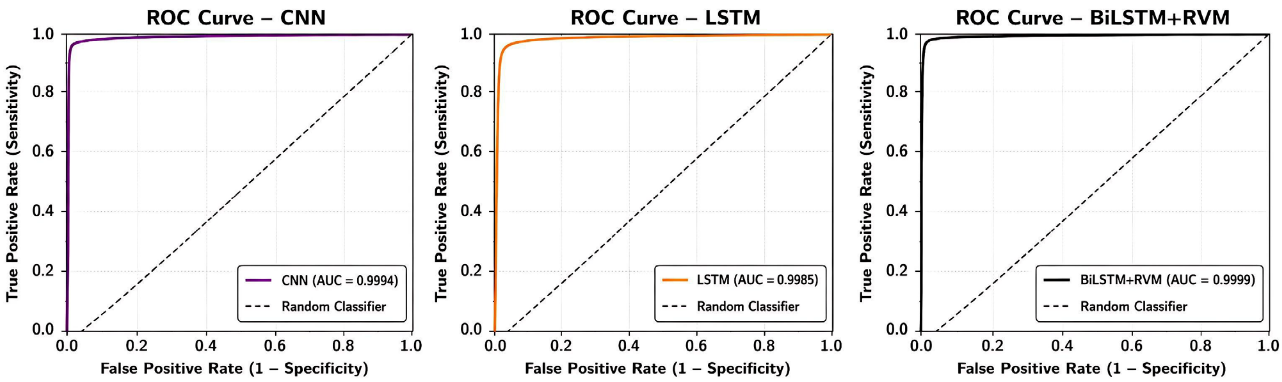
(a)



(b)

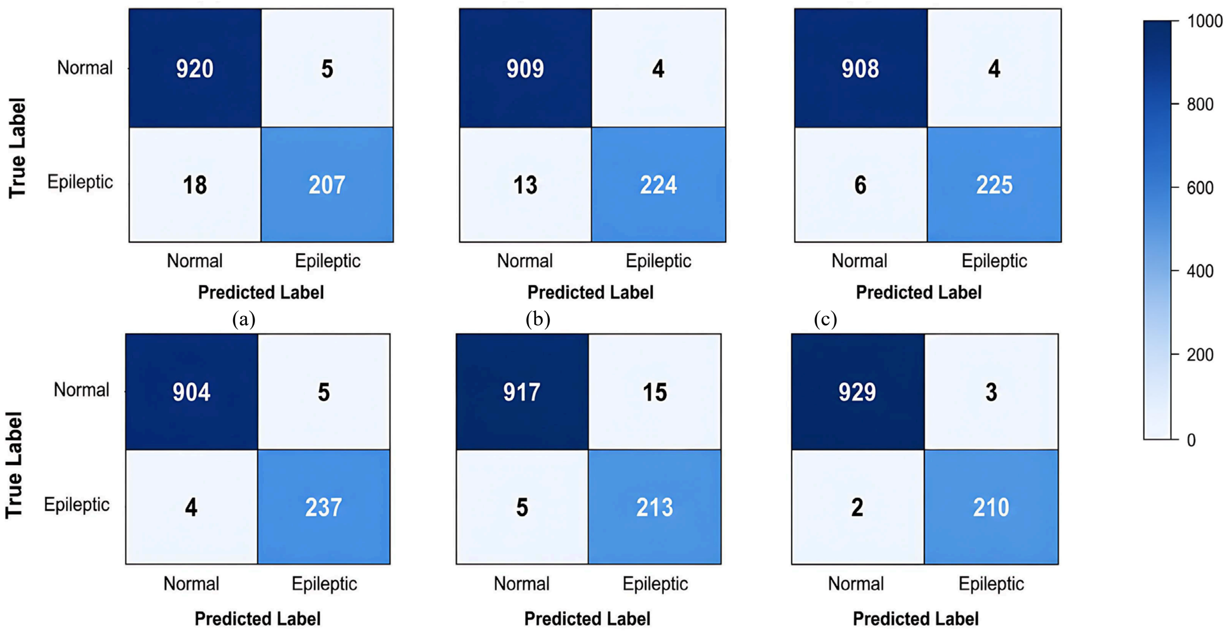


(c)



(d) (e) (f)

Fig. 12. ROC curves of the evaluated classification models: (a) XGBoost, (b) BiLSTM + Softmax, (c) BiLSTM + SVM, (d) CNN, (e) LSTM, and (f) the proposed BiLSTM-RVM model.



(d) (e) (f)

Fig. 13. Confusion matrices of the evaluated classification models: (a) XGBoost, (b) BiLSTM + Softmax, (c) BiLSTM + SVM, (d) CNN, (e) LSTM, and (f) the proposed BiLSTM-RVM model.

The BiLSTM + Softmax model achieved an accuracy of 98.52%, with F1-scores of 0.9907 and 0.9634 for the non-ictal and ictal classes, respectively. As shown in the confusion matrix, the model correctly classified 909 non-ictal EEG segments and 224 ictal EEG segments, while producing only 4 false positives and 13 false negatives, confirming its strong predictive performance.

Furthermore, the ROC curve exhibits a well-defined convex trajectory toward the upper-left corner, with an AUC of 0.9918, indicating excellent class separability across decision thresholds. Nevertheless, the reduced recall for the ictal class suggests that the Softmax classifier may be insufficient to fully delineate the complex and non-linear decision boundaries inherent to EEG signals.

The ROC curve and confusion matrix corresponding to the BiLSTM+SVM architecture are presented in Figs. 12(c) and 13(c), respectively.

The BiLSTM + SVM model attains an accuracy of 98.52%, with F1-scores of 0.9907 for the non-ictal class and 0.9636 for the ictal class. The confusion matrix confirms strong predictive performance, with 908 true negatives, 225 true positives, and only 17 total misclassifications. The ROC curve exhibits a well-defined convexity toward the upper-left corner, with an AUC of 0.9975, indicating near-perfect class separability across decision thresholds.

Compared to the BiLSTM + Softmax model, the BiLSTM + SVM architecture substantially reduces false negatives (6 vs. 13), thereby enhancing ictal detection sensitivity. This improvement is accompanied by a moderate increase in false positives (11 vs. 4), reflecting the inherent sensitivity–specificity trade-off in binary epileptic seizure detection.

Figs. 12(d) and 13(d), present the confusion matrix and the ROC curve, respectively, obtained for the standalone CNN architecture.

The standalone CNN model achieves an accuracy of 99.22%, with F1-scores of 0.9950 for the non-ictal class and 0.9814 for the ictal class. The confusion matrix reveals minimal misclassification, with only 4 false positives and

5 false negatives, confirming strong predictive capability. The ROC curve exhibits a pronounced convex trajectory toward the upper-left corner, with an AUC of 0.9994, reflecting near-perfect class separability across decision thresholds.

Nevertheless, the reduced recall for the ictal class suggests that the CNN architecture may encounter difficulties in capturing subtle and non-linear temporal patterns in EEG signals, potentially limiting its sensitivity in critical clinical scenarios.

Figs. 12(e) and 13(e), present the ROC curve and the confusion matrix, respectively, obtained for the standalone LSTM architecture.

The LSTM model achieves an accuracy of 98.26% and an AUC of 0.9985, indicating excellent class separability. The confusion matrix reveals minimal misclassification, with only 15 false positives and 5 false negatives. Notably, 213 out of 218 ictal segments are correctly identified, confirming high sensitivity to seizure patterns. The F1-scores of 0.9892 for the non-ictal class and 0.9552 for the ictal class further confirm overall classification robustness.

The slightly reduced precision for the ictal class suggests that certain non-ictal EEG segments share morphological similarities with seizure activity, resulting in a limited number of false alarms. Overall, these results demonstrate that the standalone LSTM model effectively captures the temporal dynamics of EEG signals, providing reliable performance for automated epileptic seizure detection. However, its unidirectional processing inherently limits its ability to exploit future temporal context, a limitation addressed by the proposed BiLSTM-RVM architecture.

Table V presents the ablation study comparing six classification architectures: BiLSTM + Softmax, BiLSTM + SVM, CNN, LSTM, XGBoost, and the proposed BiLSTM + RVM.

TABLE V. RESULTS OF THE ABLATION STUDY

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	AUC (%)
BiLSTM + Softmax	98.52	98.25	94.51	96.34	99.18
BiLSTM + SVM	98.52	95.34	97.40	96.36	99.75
CNN	99.21	98.34	97.93	98.14	99.94
LSTM	98.26	93.42	97.71	95.52	99.85
XGBoost	98.00	97.64	92.00	94.74	99.79
BiLSTM + RVM	**99.57**	**98.63**	**99.08**	**98.84**	**99.97**

The results reveal notable performance variations across architectures in terms of accuracy, precision, recall, F1-score, and AUC. Among the evaluated configurations, the standalone CNN achieves competitive performance with an accuracy of 99.21% and an AUC of 99.94%. The proposed BiLSTM + RVM attains the highest overall performance, with an accuracy of 99.57%, a precision of 98.63%, a recall of 99.08%, an F1-score of 98.84%, and an AUC of 99.97%.

Boost is included as a robust non-sequential baseline to broaden the comparative scope. It achieves an accuracy of 98.00%, a precision of 97.64%, a recall of 92.00%, an F1-score of 94.74%, and an AUC of 99.79%. Despite its strong discriminative capability, the relatively lower recall of XGBoost indicates reduced sensitivity to ictal EEG patterns, attributable to its inability to model fine-grained temporal dependencies inherent to EEG signals.

These results confirm that the integration of BiLSTM and RVM yields complementary benefits: the BiLSTM effectively extracts discriminative long-range temporal representations, while the RVM refines the decision boundary through sparse Bayesian regularization. This synergy enhances both classification robustness and generalization capability across evaluation conditions.

In summary, the proposed BiLSTM + RVM model achieves superior classification performance while simultaneously minimizing false alarms and missed seizures, reinforcing its clinical relevance and suitability for real-world EEG-based epileptic seizure detection systems.

G. Statistical Analysis

1) Comparison of model performance

To evaluate performance stability and reproducibility, each model is executed over ten runs with different random initializations. For each evaluation metric (AUC, accuracy, precision, sensitivity, and F1-score), the mean (μ) and standard deviation (σ) are computed. The corresponding statistical results are summarized in Table V.

TABLE VI. BILSTM + SOFTMAX, BILSTM + SVM AND BILSTM + RVM STATISTICAL RESULTS

Model	Runs	AUC	Accuracy	Precision	Recall	F1-Score
	1	99.18	98.53	98.25	94.51	96.34
	2	99.65	99.44	98.39	94.57	96.44
	3	99.43	98	93.80	96.60	95.18
	4	99.47	98.70	96.22	97.47	96.83
	5	99.20	98.35	97.64	93.67	95.61
BILSTM+ SOFTMAX	6	99.51	98.17	95.05	95.48	95.26
	7	99.20	98.60	95.49	97.90	96.68
	8	99.89	98.78	96.65	97.47	97.06
	9	98.99	98.26	97.45	94.24	95.82
	10	99.55	98	94.35	96.30	95.32
	μ_3	**99.407**	**98.48**	**96.32**	**95.82**	**96.05**
	σ_3	**0.26**	**0.43**	**1.62**	**1.53**	**0.70**
	1	99.75	98.52	95.34	97.40	96.36
	2	99.84	98.52	97.71	94.67	96.16
	3	99.78	98.52	97.22	96.08	96.65
	4	98.85	98.52	96.71	95.37	96.04
	5	99.86	98.78	97.13	97.13	97.13
BILSTM+ SVM	6	99.32	98.17	96.19	94.98	95.58
	7	99.44	98.34	95.98	96.37	96.18
	8	99.86	98.34	98.60	92.95	95.69
	9	99.63	98.17	96.85	98.39	95.34
	10	99.75	98.26	95.28	96.10	95.69
	μ_3	**99.60**	**98.31**	**96.70**	**95.94**	**96.08**
	σ_3	**0.31**	**0.18**	**0.99**	**1.55**	**0.54**
	1	99.97	99.57	98.63	98.08	98.84
	2	99.95	99.14	98.60	96.79	97.69
	3	99.98	99.48	99.07	98.17	98.62
	4	99.99	99.47	98.18	98.08	98.63
	5	99.98	99.36	99.07	99.71	98.38
BILSTM+ RVM	6	99.87	99.13	97.27	98.17	97.72
	7	99.82	99.47	99.07	98.17	98.62
	8	99.98	99.39	99.07	97.71	98.38
	9	99.98	99.30	97.73	98.62	98.17
	10	99.98	99.30	98.17	98.17	98.17
	μ_3	**99.95**	**99.36**	**98.48**	**98.17**	**98.32**
	σ_3	**0.05**	**0.14**	**0.63**	**0.62**	**0.39**

Furthermore, the average performance ($\mu \pm \sigma$) of the three models across the ten independent runs are also reported in Table VI.

The BiLSTM-Softmax model achieves consistent performance across ten runs, with a mean AUC of 99.41% (σ = 0.26), an accuracy of 98.48% (σ = 0.43), and an F1-score of 96.05% (σ = 0.70), reflecting stable convergence behavior across random initializations.

The BiLSTM + SVM model achieves high and stable performance across ten runs, with a mean AUC of 99.60% (σ = 0.31), an accuracy of 98.31% (σ = 0.18), and an F1-score of 96.08% (σ = 0.54), confirming the robustness of the SVM classifier when combined with BiLSTM-based feature extraction.

The proposed BiLSTM + RVM model achieves consistently high performance across ten independent runs, with a mean AUC of 99.95% (σ= 0.05%) and a mean accuracy of 99.36% (σ = 0.14%), confirming excellent discriminative capability and classification stability. It further attains a mean precision of 98.48% (σ= 0.63%), a mean recall of 98.17% (σ= 0.62%), and a mean F1-score of 98.32% (σ = 0.39%), demonstrating a well-balanced sensitivity-precision trade-off and strong reproducibility across initializations.

Table VII summarizes the mean performance ($\mu \pm \sigma$) of the three architectures across multiple independent runs. The proposed BiLSTM + RVM model consistently outperforms the competing approaches across all evaluation metrics, achieving the highest AUC of 99.95% ± 0.02% and accuracy of 99.36% ± 0.14%, reflecting near-perfect discriminative performance with minimal run-to-run variability.

TABLE VII. MEAN PERFORMANCE ± STANDARD DEVIATION (%)

Metric	BILSTM + Softmax	BILSTM + SVM	BILSTM + RVM
AUC	99.407 ± 0.26	99.60 ± 0.31	**99.95 ± 0.02**
Accuracy	98.48 ± 0.43	98.31 ± 0.18	**99.36 ± 0.14**
Precision	96.32 ± 1.62	96.70 ± 0.99	**98.48 ± 0.63**
Recall	95.82 ± 1.53	95.94 ± 1.55	**98.17 ± 0.62**
F1-score	96.05 ± 0.7	96.08 ± 0.54	**98.32 ± 0.39**

BiLSTM-SVM demonstrates strong performance, with an AUC of 99.60% ± 0.31% and an accuracy of 98.31% ± 0.18%. BiLSTM-Softmax yields comparatively lower results, with an AUC of 99.41% ± 0.26% and an accuracy of 98.48% ± 0.43%.

In terms of precision, recall, and F1-score, the proposed BiLSTM–RVM achieves the highest values across all three metrics 98.48% ± 0.63%, 98.17% ± 0.62%, and 98.32% ± 0.39%, respectively confirming its superior balance between sensitivity and specificity, as well as its robustness and consistency across runs.

2) Statistical significance test

To determine whether the performance differences observed across models are statistically significant, a two-sample Welch's t-test is applied. The effective degrees of freedom are estimated using the Welch-Satterthwaite approximation. Table VIII summarizes the pairwise comparison results between RVM and Softmax, and between RVM and SVM, respectively.

TABLE VIII. COMPARISON BETWEEN RVM VS SOFTMAX, AND RVM VS SVM

Comparison	Metric	t	p-value	Interpretation
RVM vs Softmax	AUC	32.6	1.2×10^{-13}	significant
	Accuracy	7.2	2.5×10^{-5}	significant
	Precision	3.6	0.0036	significant
	Recall	4.7	0.0007	significant
	F1-score	6.8	4.1×10^{-5}	significant
RVM vs SVM	AUC	3.5	0.0056	significant
	Accuracy	20.4	3.2×10^{-13}	significant
	Precision	5.7	4.8×10^{-5}	significant
	Recall	4.6	0.0008	significant
	F1-score	12.1	2.6×10^{-9}	significant

The results reported in Table VIII demonstrate that the proposed RVM model significantly outperforms the Softmax model across all evaluated metrics. The Welch's *t*-test yields *p*-values well below the 0.05 significance threshold for all measures, confirming that the observed differences are statistically significant and not attributable to random variation.

The most pronounced improvement is observed for AUC (t = 32.6; $p = 1.2 \times 10^{-13}$), indicating a substantial enhancement in discriminative capability. Statistically significant gains are further obtained for accuracy, precision, recall, and F1-score, with consistently low *p*-values ($p \leq 10^{-3}$) across all metrics.

Collectively, these results confirm that the RVM classifier yields statistically robust and consistent improvements over the Softmax-based classifier across all evaluated performance metrics.

The results reported in Table VIII confirm that the RVM model significantly outperforms the SVM model across all evaluated metrics. The Welch's *t*-test yields *p*-values below the 0.05 significance threshold for all comparisons. The most pronounced improvements are observed in accuracy (t = 20.4; $p = 3.2 \times 10^{-13}$) and F1-score (t = 12.1; $p = 2.6 \times 10^{-9}$), indicating substantial gains in overall classification performance. Statistically significant improvements are further obtained in precision and recall. Although comparatively moderate, the improvement in AUC (t = 3.5; p = 0.0056) remains statistically significant, confirming the superior discriminative capability of the RVM classifier.

Collectively, these findings demonstrate that integrating the RVM as the final classifier significantly enhances the overall performance of the BiLSTM model, particularly in terms of accuracy, sensitivity, and F1-score. The absence of significant differences for certain metrics suggests that the improvements provided by the RVM are not limited to overall discrimination but also reflect a better optimization of the sensitivity-specificity trade-off. Furthermore, the low inter-run variability confirms the robustness and stability of the proposed model.

H. Effect size analysis (Cohen's d)

Although Welch's *t*-tests confirm the statistical significance of the observed differences, *p*-values alone do not convey information about practical relevance. To quantify the magnitude of the observed effects, Cohen's *d* effect size is computed for each statistically significant pairwise comparison.

Table IX summarizes the effect sizes for the comparisons between the RVM classifier and the Softmax and SVM classifiers. The results demonstrate that the proposed BiLSTM + RVM model consistently achieves large to very large effect sizes across all evaluation metrics, indicating that the observed performance improvements are not only statistically significant but also practically meaningful. Although the AUC difference between RVM and SVM is comparatively smaller, substantial improvements are evident in the remaining performance metrics, particularly accuracy and F1-score.

TABLE IX. EFFECT SIZES (RVM VS. SOFTMAX, AND RVM VS. SVM)

Comparison	Metric	*d*	Interpretation
RVM vs Softmax	AUC	13.9	Very significant effect
	Accuracy	2.76	Very significant effect
	Precision	1.74	Significant effect
	Recall	1.99	Significant effect
	F1-score	4.09	Very Significant effect
RVM vs SVM	AUC	1.17	Significant effect
	Accuracy	6.19	Very Significant effect
	Precision	3.58	Very Significant effect
	Recall	2.00	Very Significant effect
	F1-score	4.73	Very Significant effect

The combined analysis of *p*-values and effect sizes confirms that the observed improvements are not only statistically significant but also practically meaningful. The large effect sizes obtained for AUC (vs. Softmax) and accuracy (vs. SVM) reflect genuine gains in discriminative capability and overall classification SVM is comparatively small suggesting comparable discriminative capacity between the two classifiers the proposed RVM model demonstrates superior decision calibration through its probabilistic Bayesian inference mechanism. Collectively, these results indicate that the integration of the RVM does not yield a marginal improvement but rather a measurable and consistent enhancement across all key performance metrics.

I. Model Explainability and Clinical Interpretability

To provide clinically interpretable insight into the decision-making process of the proposed architecture, an explainability analysis based on SHapley Additive exPlanations (SHAP) is conducted on the latent feature representations extracted by the BiLSTM network prior to RVM classification. Due to the high computational cost of the KernelExplainer algorithm, the analysis is performed on a representative subset of 20 samples. This subset provides a computationally efficient yet sufficiently reliable approximation of the model's global behavior while preserving the interpretability of the results. The corresponding temporal SHAP projection on epileptic and non-epileptic EEG signals is illustrated in Fig. 14.

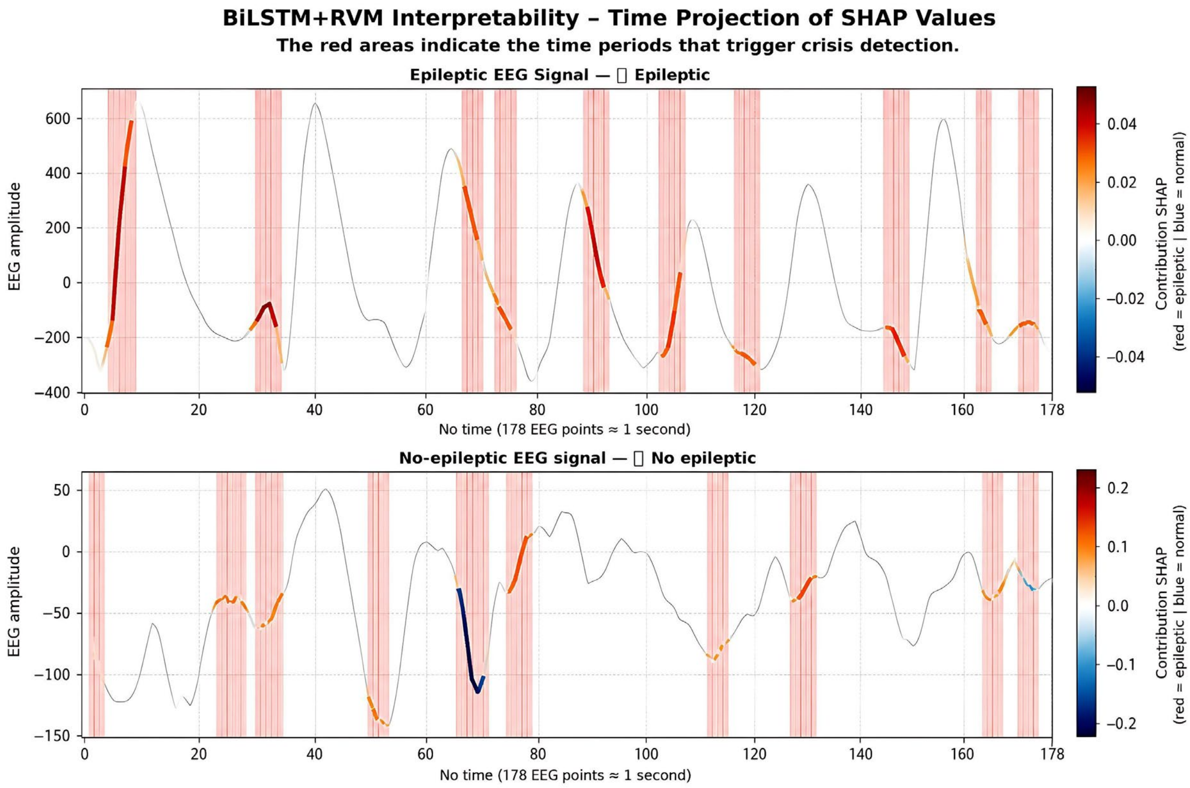


Fig. 14. Temporal SHAP projection on epileptic and non-epileptic EEG signals.

1) Global feature importance analysis

Fig. 15 presents the SHAP beeswarm summary plot for the most influential latent features. A clear feature hierarchy emerges: F25 is identified as the most discriminative feature, with SHAP values reaching up to +0.42, followed by F28 and F47.

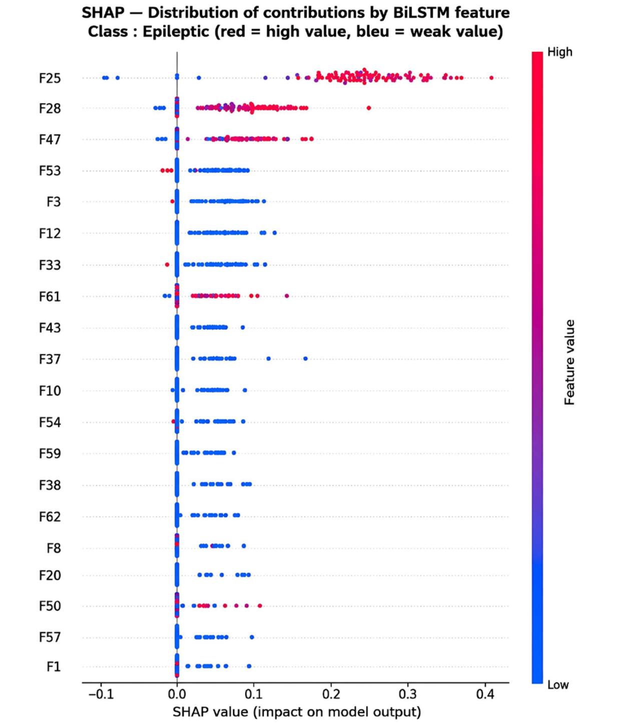


Fig. 15. SHAP feature importance analysis for BiLSTM-based epileptic seizure classification.

Most SHAP contributions are positive, indicating that higher feature activations consistently shift model predictions toward the ictal class. Higher feature values, represented by red points, are associated with the largest positive contributions, whereas lower feature values, represented by blue points, exhibit limited impact on the classification output.

These findings confirm that the BiLSTM effectively captures discriminative temporal patterns within a compact subset of highly informative latent features, supporting the interpretability and clinical relevance of the proposed framework.

2) Temporal projection of SHAP contributions

To establish a connection between the abstract latent feature space and a clinically interpretable temporal representation, SHAP values are mapped onto the time domain, as illustrated in Figs. 14 and 16.

For ictal signals, SHAP contributions are robust and distributed throughout the entire signal duration, with strong correlation to high-amplitude peaks ranging from 600 to 1000 µV. This indicates that the proposed model captures global temporal dynamics rather than isolated events, consistent with the complex and distributed nature of ictal activity.

In contrast, for non-ictal signals, SHAP contributions are low in magnitude and diffuse, exhibiting no structured temporal pattern. The presence of localized positive contributions in certain segments may correspond to interictal discharges resembling epileptiform activity, thereby accounting for the rare false positives observed in the classification results.

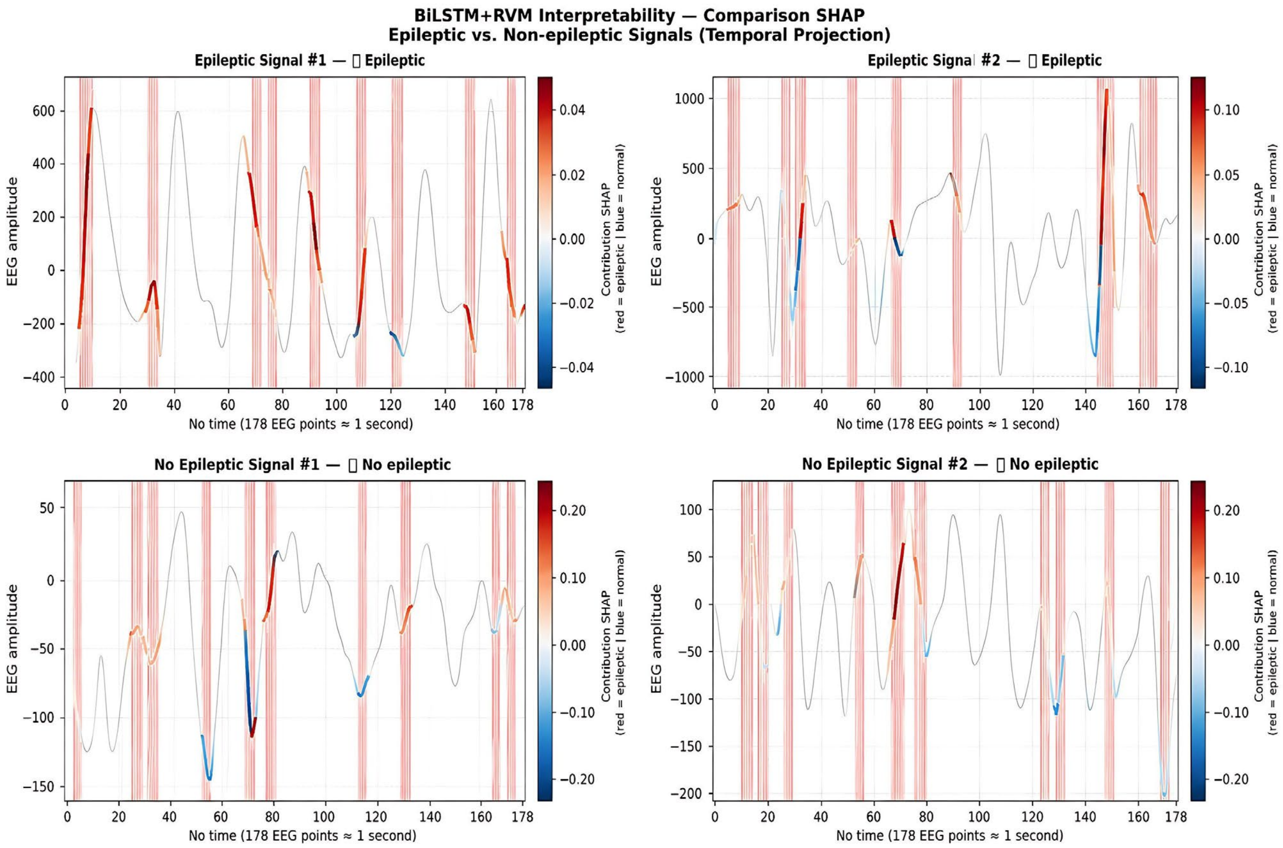


Fig. 16. Comparison of temporal SHAP patterns for epileptic and non-epileptic EEG signals.

3) Clinical Interpretation and Implications

From a clinical perspective, the SHAP analysis highlights three key properties of the proposed model.

- Feature interpretability: The identification of dominant latent dimensions F25, F28, and F47 indicates that discriminative information is concentrated within specific components of the BiLSTM feature vector, enabling targeted clinical interpretation.
- Temporal interpretability: Projecting SHAP values onto the time domain enables the localization of temporal intervals contributing most to epileptic seizure detection, thereby enhancing model transparency and supporting clinician trust.
- Decision consistency: The clear distinction between ictal profiles, characterized by strong and structured SHAP contributions, and non-ictal profiles, exhibiting weak and diffuse contributions, confirms that the proposed model relies on physiologically meaningful temporal patterns rather than spurious correlations.

4) Contribution of RVM to interpretability

The RVM classifier further enhances interpretability through its sparse Bayesian formulation, relying on a limited subset of relevance vectors. This property reduces model complexity and complements the SHAP analysis by emphasizing the most informative temporal patterns, thereby improving both generalization capability and decision transparency.

Overall, the SHAP analysis demonstrates that the proposed BiLSTM-RVM model does not operate as a black box, but grounds its decisions on physiologically consistent and clinically interpretable temporal patterns, thereby reinforcing its suitability for real-world clinical deployment.

J. Robustness Evaluation under Noisy Conditions

1) Robustness evaluation protocol

EEG signals acquired in real clinical environments are inherently corrupted by various noise sources and artifacts, including electromagnetic interference, Electromyographic (EMG) activity, Electrooculographic (EOG) artifacts, and motion-induced disturbances. Assessing the robustness of the proposed model to such perturbations is therefore essential for validating its applicability in real-world deployment scenarios.

To this end, a systematic robustness evaluation is conducted using two representative noise models: additive Gaussian noise and impulsive noise. For each noise type, seven perturbation levels are applied to the entire test set comprising 1150 segments. Noisy signals are generated by injecting noise directly into the original test data, converted to float32 format. Temporal feature extraction is subsequently performed using the BiLSTM network, followed by classification using the RVM trained exclusively on clean data, without any retraining or adaptation.

This evaluation protocol simulates a realistic clinical deployment scenario in which the model is exposed to unseen noisy signals without prior adaptation.

2) Results under gaussian noise

Fig. 17 illustrates the comparison between a clean EEG signal and the same signal corrupted by additive Gaussian noise. The presence of noise induces a progressive degradation of signal quality, characterized by increased stochastic fluctuations and a reduction in the Signal-to-Noise (SNR) ratio, defined as the ratio of signal power to noise power. This distortion directly affects the stability of temporal and spectral features, potentially reducing class separability at the classification stage.

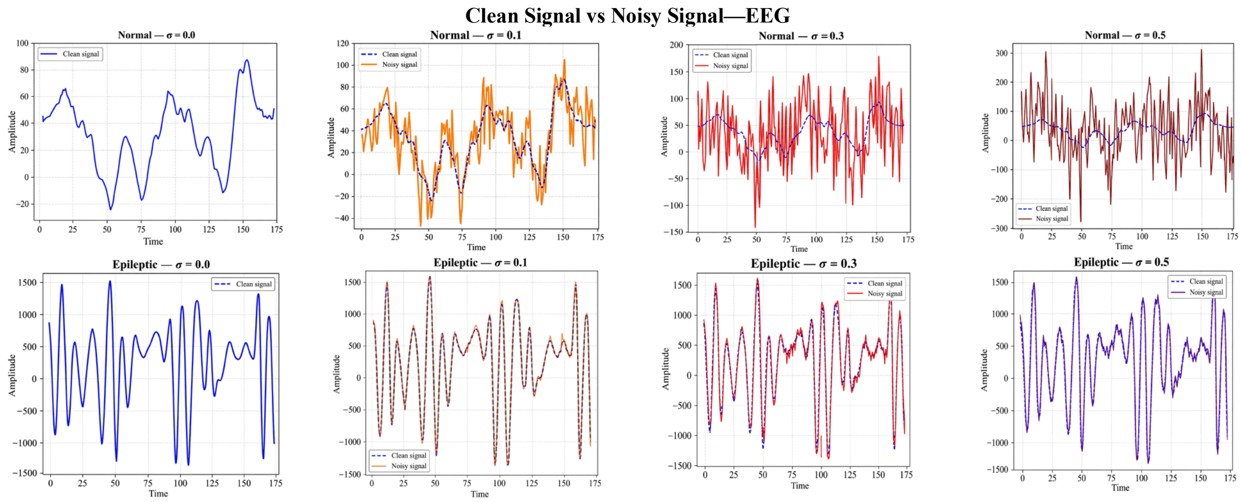


Fig. 17. Comparison between a clean EEG signal and its Gaussian noise corrupted version.

Table X reports the effect of increasing Gaussian noise standard deviation (σ) on model performance.

The results under additive Gaussian noise confirm the strong robustness of the proposed BiLSTM-RVM model. At low to moderate noise levels, the model maintains near-baseline performance: accuracy remains above 98% for $\sigma \le 0.05$ and above 96% for $\sigma \le 0.20$, corresponding to a maximum degradation of only −2.11% at $\sigma = 0.20$. This behavior reflects the model's capacity to effectively handle distributed low-amplitude perturbations, characteristic of real-world EEG acquisition conditions.

TABLE X. ROBUSTNESS EVALUATION UNDER ADDITIVE GAUSSIAN NOISE: CLASSIFICATION ACCURACY AND F1-SCORE VS NOISE LEVEL (σ)

Noise Level (σ)	Accuracy (%)	F1-score (%)	Degradation (%)
0.00 (Clean signal (baseline): obtained results)	**98.96**	**98.95**	—
0.01 (1% of the amplitude)	98.87	98.87	−0.09
0.05 (5% of the amplitude)	98.87	98.87	−0.09
0.10 (10% of the amplitude)	98.43	98.43	−0.53
0.20 (20% of the amplitude)	96.87	96.93	−2.11
0.30 (30% of the amplitude)	92.96	93.29	−6.06
0.50 (50% of the amplitude)	71.48	74.25	−27.77

A more pronounced performance degradation is observed at higher noise levels, with accuracy declining to 92.96% at $\sigma = 0.30$ and to 71.48% at $\sigma = 0.50$. However, such noise levels correspond to extremely low signal-to-noise ratios that are rarely encountered in clinical practice and may therefore be considered beyond realistic operating conditions.

Overall, the proposed model preserves more than 97% of its baseline performance under realistic Gaussian noise conditions, confirming its robustness to continuous stochastic perturbations.

3) Results under impulsive noise

Fig. 18 presents a comparison between a clean EEG signal and the same signal corrupted by impulsive noise. Unlike additive Gaussian noise, impulsive noise introduces sporadic high-amplitude outliers, resulting in strong localized distortions across the signal. These transient impulses may obscure discriminative temporal patterns and significantly impair feature extraction, thereby reducing classification robustness.

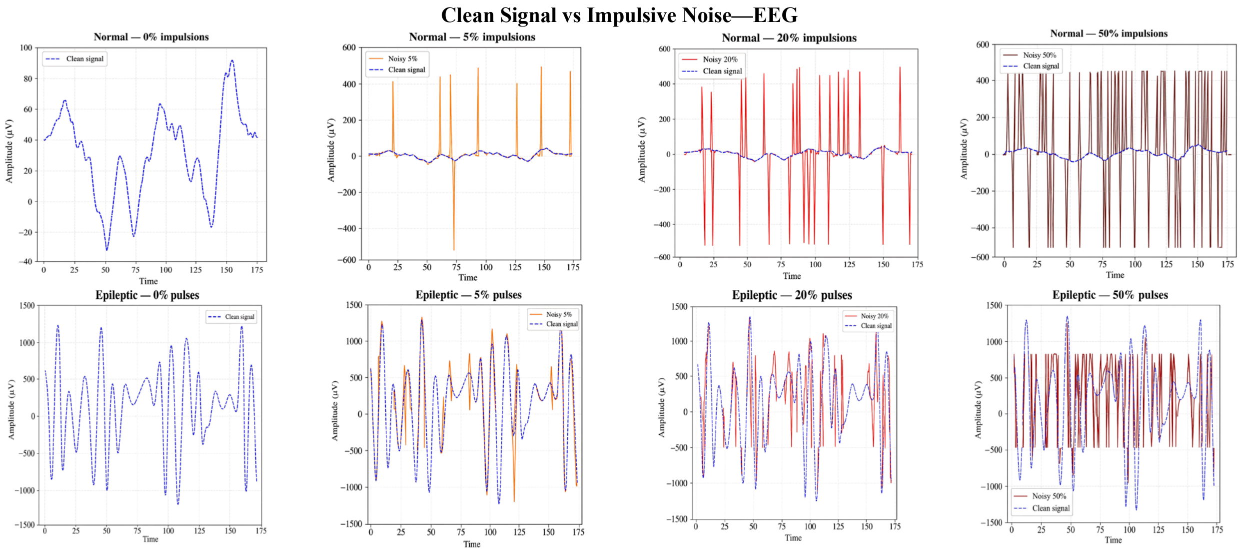


Fig. 18. Comparison between a clean EEG signal and its impulsive noise corrupted version.

Table XI reports the effect of increasing impulsive noise probability (p) on model performance.

The robustness analysis under impulsive noise reveals a more challenging degradation scenario, as this noise type introduces localized high-amplitude distortions that disrupt temporal feature extraction more severely than additive Gaussian noise.

TABLE XI. ROBUSTNESS EVALUATION UNDER IMPULSIVE NOISE: CLASSIFICATION ACCURACY AND F1-SCORE VS NOISE PROBABILITY (P)

Ratio p	Accuracy	F1-score	Degradation	Interpretation
0.00	98.96	98.95	—	Baseline
0.01	98.09	98.10	−0.88%	Realistic condition
0.05	90.09	90.73	−8.96%	Upper limit
0.10	71.65	74.40	−27.60%	Extreme
≥0.20	↓↓↓	↓↓↓	>−50%	Stress test

Nevertheless, the model exhibits strong resilience under realistic corruption levels. At $p = 1\%$, performance remains near baseline, with a degradation of only −0.88%. At $p = 5\%$, accuracy remains above 90%, preserving approximately 91% of baseline performance. These results confirm that the proposed model is robust to low and moderate impulsive noise levels, consistent with typical artifact conditions encountered in clinical EEG recordings.

Performance degradation becomes substantial only for $p \geq 10\%$, with a sharp decline observed beyond $p \geq 20\%$. However, such corruption levels represent extreme acquisition failure scenarios in which a large proportion of the signal is effectively destroyed, rendering them unlikely in standard clinical practice.

4) Comparative analysis and clinical implications

A comparative analysis of model responses under Gaussian and impulsive noise reveals a differential robustness profile. The proposed BiLSTM-RVM model demonstrates strong robustness to additive Gaussian noise, attributable to the distributed nature of such perturbations and the capacity of the BiLSTM to capture global temporal dependencies. In contrast, moderate sensitivity to impulsive noise is observed, explained by the localized high-amplitude nature of such artifacts, which can disrupt critical temporal features extracted by the BiLSTM network.

Importantly, the model maintains high classification performance within realistic noise ranges in both cases, confirming its suitability for practical clinical deployment. From a clinical perspective, robustness to Gaussian noise ensures reliable operation under standard EEG acquisition conditions, while acceptable performance under moderate impulsive noise suggests resilience to common EMG and EOG artifacts. Performance degradation under extreme noise levels further highlights the potential benefit of integrating upstream artifact rejection or adaptive denoising techniques as a complementary preprocessing stage.

Overall, the BiLSTM-RVM framework exhibits robust and stable behavior across a wide range of noise conditions, preserving high accuracy and F1-score under realistic perturbation levels. These findings collectively confirm the suitability of the proposed model for real-world EEG-based epileptic seizure detection, where noise is unavoidable but typically remains within clinically moderate bounds.

K. Computational Complexity and Inference Time Analysis

The computational complexity of the proposed hybrid framework is primarily governed by two components: the BiLSTM network, responsible for temporal feature extraction, and the RVM classifier, employed at the final decision stage.

The BiLSTM module constitutes the most computationally demanding component of the architecture, owing to its bidirectional recurrent structure and deep stacked configuration. The proposed architecture comprises eight BiLSTM layers, with an input sequence length of 178 time steps and an average hidden dimension of 96 units per layer. The theoretical computational complexity of the temporal feature extraction stage is accordingly approximated as:

$$O\left(8\times178\times96^{2}\right)=O\left(1.31\times10^{7}\right)$$

In contrast, the RVM classifier benefits from a sparse Bayesian formulation, resulting in substantially lower prediction complexity. Given 41 relevance vectors and a feature dimensionality of 64, the computational cost of the classification stage is approximated as:

$$O\left(41\times64\right)=O\left(2624\right)$$

The computational overhead introduced by the RVM classifier remains negligible compared with that of the BiLSTM feature extractor, contributing only marginally to the total inference cost. Compared with conventional dense classifiers, the sparse formulation of the RVM substantially reduces the computational burden at the decision stage while maintaining robust predictive capability.

To assess the practical efficiency of the proposed framework, inference time is measured experimentally using an NVIDIA T4 GPU in the Google Colab environment. The results indicate an average inference time of approximately 4.6 ms per EEG segment, corresponding to a total processing time of 55.9 seconds for the complete test set.

These findings confirm that, despite the depth of the BiLSTM architecture, the proposed model maintains strong computational efficiency without compromising predictive performance. The hybrid BiLSTM-RVM framework thus provides an effective trade-off between representational capacity and computational cost, combining deep temporal feature learning, sparse probabilistic classification, and low-latency inference properties that collectively support its deployment in near real-time EEG-based clinical applications.

V. Discussion and Limitations

A. Discussion of Results

The experimental results demonstrate that the proposed BiLSTM-RVM model achieves outstanding performance for epileptic seizure detection while effectively addressing several key challenges identified in the literature, including inter-subject generalization, cross-dataset robustness, clinical interpretability, and noise resilience.

On the primary dataset, the model attains an accuracy of 99.57% and an AUC of 99.97%, reflecting near-perfect discriminative capability between ictal and non-ictal EEG signals. Unlike many existing approaches that optimize a single evaluation metric, the proposed model maintains a consistent balance between precision, recall, and F1-score, which is a critical requirement in clinical settings where both false positives and false negatives carry significant consequences.

The LOSO evaluation confirms strong generalization capability across unseen subjects. The low performance variance observed across folds indicates that the model learns subject-invariant temporal representations rather than subject-specific patterns, representing a meaningful advancement over conventional deep learning models, which are often sensitive to inter-subject variability. Cross-dataset validation from the Bonn to the CHB-MIT database further highlights the model's robustness to distributional shifts. Although a performance degradation is expected and observed, the model retains high precision (97.37%) and specificity (98.87%), reflecting a consistently low false-alarm rate. This conservative classification behavior is particularly relevant in clinical practice, where minimizing false alarms is a primary operational constraint.

The analysis of the relationship between predictive performance and model complexity, quantified through the number of relevance vectors, confirms that classification accuracy is largely independent of structural complexity. This behavior validates the sparse Bayesian learning paradigm, wherein the quality of retained relevance vectors outweighs their quantity.

The SHAP-based explainability analysis constitutes a significant contribution in terms of model transparency. Unlike conventional deep learning models commonly regarded as black boxes, the proposed framework identifies dominant latent features (F25, F28, and F47), enables temporal projection of classification decisions, and reveals physiologically coherent activation patterns. These properties substantially enhance confidence in the model for potential integration into clinical decision-support systems.

The noise robustness evaluation demonstrates that the model maintains high classification performance up to moderate perturbation levels, suggesting that the BiLSTM effectively captures stable temporal structures that are relatively insensitive to local signal fluctuations. Performance degradation becomes pronounced only at high noise levels ($\sigma \geq 0.30$), reflecting an inherent limitation of models trained exclusively on clean data when exposed to extreme acquisition disturbances. Integrating adaptive denoising or artifact rejection as a preprocessing stage is therefore recommended for deployment under severe noise conditions.

Collectively, these results confirm that the proposed BiLSTM-RVM framework is well suited for real-world clinical environments, compatible with embedded system constraints through the sparsity of the RVM classifier, and relevant for integration into automated medical decision-support pipelines.

B. Limitations of the Study

Despite the strong performance achieved, the present study is subject to several limitations that warrant careful consideration.

First, the primary dataset (UCI Epileptic Seizure Recognition) is relatively clean and well-structured, which may result in an overestimation of the reported performance metrics. Although cross-dataset validation on the CHB-MIT EEG Database partially mitigates this limitation, further evaluation on more heterogeneous, noisy, and clinically representative datasets remains necessary to ensure robust generalization.

Second, the proposed framework reformulates the original five-class problem into a binary classification task distinguishing ictal from non-ictal EEG segments. While this simplification facilitates model optimization and enables direct comparison with prior studies, it inherently reduces the granularity of EEG pattern characterization and disregards clinically relevant intermediate states, such as preictal and interictal phases. This formulation may consequently limit the clinical applicability of the model in real-world diagnostic settings, where fine-grained multi-class discrimination is often required.

Third, the present study addresses epileptic seizure detection rather than seizure prediction from preictal activity. Early prediction remains a critical and largely unresolved challenge in clinical epileptology, with significant implications for patient monitoring and timely intervention.

Fourth, part of the experimental evaluation relies on single-channel EEG signals, whereas clinical recordings are inherently multichannel. This constraint limits the model's capacity to capture inter-electrode spatial dependencies, which convey complementary and diagnostically relevant information, potentially restricting its generalization to real clinical environments.

Finally, the proposed framework has not yet been validated under real-time or embedded deployment conditions a prerequisite for clinical translation in applications requiring low-latency processing and continuous patient monitoring.

These limitations define clear directions for future investigation, including evaluation on richer multichannel clinical datasets, extension toward preictal seizure prediction, and real-time embedded system validation.

VI. Conclusion and Future Perspectives

This study presents a novel hybrid framework combining a Bidirectional Long Short-Term Memory (BiLSTM) network with a Relevance Vector Machine (RVM) for automated epileptic seizure detection from EEG signals. The proposed architecture is specifically designed to address key challenges in EEG-based clinical systems, including long-range temporal dependency modeling, inter-subject generalization, noise resilience, and model interpretability.

Extensive experiments conducted on benchmark datasets demonstrate that the proposed BiLSTM-RVM model achieves state-of-the-art performance, attaining an accuracy of 99.57% and an AUC of 99.97% on the primary dataset. The model is further validated using a Leave-One-Subject-Out (LOSO) protocol, confirming strong generalization across unseen subjects, and through cross-dataset evaluation on the CHB-MIT database, which highlights its robustness to distributional shifts under realistic clinical conditions, with consistently high precision and specificity.

The ablation study confirms the effectiveness of the RVM as a probabilistic classification stage, substantially improving the sensitivity specificity balance compared to conventional Softmax and SVM-based approaches. Statistical significance tests and effect size analysis further validate the reliability and consistency of the observed performance gains.

A principal contribution of this work resides in its interpretability. Through SHAP-based analysis, the proposed model is shown to rely on physiologically meaningful temporal patterns, enabling transparent and clinically relevant decision-making a property of critical importance for integration into medical decision-support systems. Robustness evaluation under additive Gaussian and impulsive noise further demonstrates stable classification performance under moderate signal perturbations, reinforcing the model's suitability for real-world deployment.

Collectively, the proposed BiLSTM-RVM framework successfully integrates accuracy, generalization, robustness, and interpretability, positioning it as a strong candidate for computer-assisted epilepsy diagnosis. Future work will focus on extending the framework toward preictal seizure prediction, incorporating multichannel spatial modeling, and validating the approach under real-time embedded constraints toward fully deployable clinical systems.

Conflict of Interest

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

C.A.M.: conceptualization, methodology, software development, formal analysis, investigation, and writing—original draft preparation; G.M.N.: software development, validation, data analysis, and review and editing of the manuscript; G.O.O.: methodological guidance, supervision, and review and editing of the manuscript; H.K.E.: software development, validation, and data analysis; D.T.T.: review and editing of the manuscript; H.A.: methodological guidance, supervision, review and editing of the manuscript. All authors have read and approved the final version of the manuscript.

ACKNOWLEDGMENT

This research was conducted as part of the research activities of the University of Ebolowa, within the Higher Technical Teachers' Training College (HTTTC), Department of Computer Engineering.

REFERENCE

[1] A. Omar and T. A. El-Hafeez, "Optimizing epileptic seizure recognition performance with feature scaling and dropout layers," *Neural Computing and Applications*, vol. 36, no. 6, pp. 2835–2852, 2024. doi: 10.1007/s00521-023-09204-6

[2] S. M. Usman, S. Khalid, and M. H. Aslam, "Epileptic seizures prediction using deep learning techniques," *IEEE Access*, vol. 8, pp. 39998–40007, 2020. doi: 10.1109/ACCESS.2020.2976866

[3] N. A. Samee, N. F. Mahmoud, E. A. Aldhahri *et al.*, "RNN and BiLSTM fusion for accurate automatic epileptic seizure diagnosis using EEG signals," *Life*, vol. 12, no. 12, 1946, 2022. doi: 10.3390/life12121946

[4] Y. M. Massoud *et al.*, "General and patient-specific seizure classification using deep neural networks," *Analog Integrated Circuits and Signal Processing*, vol. 116, no. 3, pp. 205–220, 2023. doi: 10.1007/s10470-023-02195-2

[5] H. Torkey *et al.*, "Seizure detection in medical IoT: Hybrid CNN-LSTM-GRU model with data balancing and XAI integration," *Algorithms*, vol. 18, no. 2, 77, 2025. doi: 10.3390/a18020077

[6] X. Zhao *et al.*, "Epileptic seizure detection using interpretable machine learning models," *Journal of Neural Engineering*, vol. 20, no. 1, 015002, 2023.

[7] L. Guo, D. Rivero, J. A. Seoane, and A. Pazos, "Classification of EEG signals using relative wavelet energy and artificial neural networks," in *Proc. the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, 2009, pp. 177–184.

[8] World Health Organization. (2024). Epilepsy. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy

[9] Y. Ma *et al.*, "A multi-channel feature fusion CNN-BiLSTM epilepsy EEG classification and prediction model based on attention mechanism," *IEEE Access*, vol. 11, 2023. doi: 10.1109/ACCESS.2023.3287927

[10] A. B. KR *et al.*, "A multi-dimensional hybrid CNN-BiLSTM framework for epileptic seizure detection using electroencephalogram signal scrutiny," *Systems and Soft Computing*, vol. 5, 200062, 2023. doi: 10.1016/j.sasc.2023.200062

[11] S. Aslan and H. Bingöl, "Epileptic seizure detection from EEG signals with recurrent neural networks based classification model," *Journal of Physical Chemistry and Functional Materials*, vol. 7, no. 2, pp. 14–21, 2024.

[12] W. Zhao *et al.*, "Residual and bidirectional LSTM for epileptic seizure detection," *Frontiers in Computational* Neuroscience, vol. 18, 1415967, 2024. doi: 10.3389/fncom.2024.1415967

[13] A. Bajaj and V. Sharma, "EEG-based epileptic seizure prediction using variants of the long short-term memory algorithm," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 17, 13, 2025. doi: 10.70917/ijcisim-2025-0001

[14] Q. Wu and E. Fokoue. (2017). Epileptic seizure recognition data set. *UCI Machine Learning Repository*. [Online]. Available: https://archive.ics.uci.edu/dataset/388/epileptic+seizure+recognition

[15] Y. Li, W. G. Cui *et al.*, "Epileptic seizure detection in EEG signals using sparse multiscale radial basis function networks and the Fisher vector approach," *Knowledge-Based Systems*, vol. 164, pp. 96–106, 2019. doi: 10.1016/j.knosys.2018.10.029

[16] M. Chakraborty and D. Mitra, "Epilepsy seizure detection using non-linear and DWT-based features," in *Proc. 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 2019, pp. 158–163. doi: 10.1109/WiSPNET45539.2019.9032871

[17] W. Mardini *et al.*, "Enhanced detection of epileptic seizures using EEG signals and machine learning," *IEEE Access*, vol. 8, pp. 24046–24055, 2020. doi: 10.1109/ACCESS.2020.2970012

[18] R. Debeche, M. E. H. Daho *et al.*, "Epileptic seizure detection from EEG signal using the imbalanced data techniques," in *Proc. International Conference on Computing Systems and Applications*, 2020, pp. 301–312. doi: 10.1007/978-3-030-69418-0_27

[19] J. Luo, Z. Zhang, Y. Fu, and F. Rao, "Time seriesprediction of COVID-19 transmission in America using LSTM and XGBoost algorithms," *Results in Physics*, vol. 27, 104462, 2021. doi: 10.1016/j.rinp.2021.104462

[20] G. Chekhmane and R. Benali, "EEG signals analysis using SVM and MLPNN classifiers for epilepsy detection," in *Proc. 2022 5th International Symposium on Informatics and its Applications (ISIA)*, 2022, pp. 1–6. doi: 10.1109/ISIA55826.2022.9993577

[21] M. Aslan and Z. Alçin, "Detection of epileptic seizures from EEG signals with Hilbert Huang transformation," *Cumhuriyet Science Journal*, vol. 42, no. 2, pp. 508–514, 2021. doi: 10.17776/csj.682734

[22] C. Lei, S. Zheng *et al.*, "Epileptic seizure detection in EEG signals using discriminative Stein kernel-based sparse representation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2021.

[23] M. Kiran *et al.*, "Improved feature space for EEG-based epileptic seizure detection," *Research Square*, 2024. doi: 10.21203/rs.3.rs-3869119/v1

[24] J. Vidya, P. Rani *et al.*, "Automatic epileptic seizure detection using SVM techniques with EEG signals," in *Proc. International Conference on Computational Innovations and Emerging Trends (ICCIET-2024)*, 2024, pp. 876–883. doi: 10.2991/978-94-6463-471-6_83

[25] M. Jafari, "Towards intelligent epilepsy detection: Leveraging statistical and frequency-domain features of electroencephalography signals within a random forest classification framework," *Salamat Academic & Research Journal*, vol. 2, pp. 87–96, 2025. doi: 10.61438/sarj.v2i2.186

[26] W. A. Mir, M. Anjum *et al.*, "Deep-EEG: An optimized and robust framework and method for EEG-based diagnosis of epileptic seizure," *Diagnostics*, vol. 13, no. 4, 773, 2023. doi: 10.3390/diagnostics13040773

[27] D. Ji, H. Cui *et al.*, "An efficient deep learning framework for automated epileptic seizure detection: Toward scalable and clinically applicable solutions," *Developmental Neurobiology*, vol. 85, no. 3, e22983, 2025. doi: 10.1002/dneu.22983

[28] T. Shawly and A. A. Alsheikhy, "A neural ODE-enhanced deep learning framework for accurate and real-time epilepsy detection," *Computer Modeling in Engineering & Sciences*, vol. 143, no. 3, 3033, 2025. doi: 10.32604/cmes.2025.065264

[29] Y. Yuan, G. Xun *et al.*, "A multi-view deep learning framework for EEG seizure detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, pp. 83–94, 2025.

[30] A. Pandey, S. K. Singh *et al.*, "An intelligent optimized deep learning model to achieve early prediction of epileptic seizures," *Biomedical Signal Processing and Control*, vol. 84, 104798, 2023. doi: 10.1016/j.bspc.2023.104798

[31] I. Ahmad, X. Wang *et al.*, "A hybrid deep learning approach for epileptic seizure detection in EEG signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 30, no.2, pp. 1019–1029, 2023.

[32] A. Dev, M. M. Fouda *et al.*, "Automated EEG epileptic seizure detection exploiting continuous morlet wavelet transform scalograms and hybrid deep learning models," in *Proc. 2023 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 2023, pp. 182–188. doi: 10.1109/IoTaIS60147.2023.10346065

[33] F. A. Jibon *et al.*, "Sequential graph convolutional network and DeepRNN based hybrid framework for epileptic seizure detection from EEG signal," *Digital Health*, vol. 10, 2024. doi: 10.1177/20552076241249874

[34] N. Mohankumar, G. Kumar *et al.*, "An optimized hybrid CNN-LSTM model for epileptic seizure detection and prediction," *Engineering, Technology & Applied Science Research*, vol. 15, pp. 26085–26090, 2025. doi: 10.48084/etasr.11042

[35] H. Üzen, H. Firat *et al.*, "A new approach to automatic epilepsy detection from EEG signals using archimedean spiral and swin transformer," *Research Square*, 2025. doi: 10.21203/rs.3.rs-5835848/v1

[36] Q. Li, W. Cao, and A. Zhang, "Multi-stream feature fusion of vision transformer and CNN for precise epileptic seizure detection from EEG signals," *Journal of Translational Medicine*, vol. 23, no. 1, 871, 2025. doi: 10.1186/s12967-025-06862-z

[37] S. S. Sikarwar, A. K. Rana, and S. S. Sengar, "Entropy-driven deep learning framework for epilepsy detection using electroencephalogram signals," *Neuroscience*, vol. 577, pp. 12–24, 2025. doi: 10.1016/j.neuroscience.2025.05.003

[38] Y. Zhao, E. He, H. Wu *et al.*, "Deep feature-based framework for automated epileptic seizure detection using EEG signals," *IEEE Sensors Journal*, vol. 26, no. 1, pp. 487–495, 2025. doi: 10.1109/JSEN.2025.3631004

[39] B. Silpa and M. K. Hota, "An efficient epileptic seizure detection framework based on optimized deep residual network," *Neural Computing and Applications*, vol. 37, no. 21, pp. 16849–16870, 2025. doi: 10.1007/s00521-025-11354-8

[40] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001. doi: 10.1162/15324430152748236

[41] Y. Gao, B. Jiang, and J. Zhou, "Financial distress prediction for small and medium enterprises using machine learning techniques," arXiv preprint, arXiv:2302.12118, 2023.