Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Pre-Ictal Seizure Prediction from EEG Using an Attention-Augmented Temporal Learning Framework

Maleka Anjum[1], Shubhangi D.C.[2]

Department of Computer Science, Vishveshwarayya Technological University CPGS, Kalaburagi, Karnataka, India.

E-mail: [1]malekaanjum786@gmail.com, [2]shubhangidc@vtu.ac.in

Orcid ID: [1]0009-0003-8776-1805, [2]0000-0002-9221-166X

Abstract

There are approximately 50 million people worldwide living with epilepsy, highlighting the need for strong early warning systems to enable prompt clinical management. Despite the high performance of recently developed deep learning models based on Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (Bi-LSTM/GRU) networks, they are highly sensitive to ictal features and are not very effective in detecting the subtle temporal transitions leading to seizure onset. This article introduces an attention-improved deep learning model to predict pre-ictal seizures using EEG signals. The suggested model combines CNN-based feature extraction, Bi-LSTM/GRU temporal sequence modelling, and a learnable temporal attention mechanism to identify the early neural dynamics before the onset of seizures. The experimental dataset was constructed from four publicly available EEG repositories following the proposed temporal labeling strategy and comprises EEG recordings from 25 selected patients, 243 annotated seizure events, and 2,847 hours of continuous EEG recordings, using a patient-wise stratified 5-fold cross-validation protocol. Experimental evaluation demonstrates that the proposed framework achieves a predictive accuracy of 93% (±1.2%), sensitivity of 90% (±1.4%), specificity of 90% (±1.1%), an AUC-ROC of 0.93 (±0.012), a PR-AUC of 0.961 (±0.009), and a low false alarm rate of 0.12 ± 0.015 per hour. To confirm statistical reliability, all reported metrics are validated across 10 independent runs and supplemented with 95% confidence intervals and paired Wilcoxon signed-rank tests ($p < 0.05$) against all baselines. Attention weight analysis verifies that the model selectively targets temporally informative pre-ictal EEG regions and improves clinical interpretability.

Keywords: Epileptic, Seizure Prediction, Deep Learning, Bidirectional, LSTM, GRU, CNN, EEG Signal Processing, Medical Diagnosis.

1. Introduction

Epilepsy is a long-term neurological condition that affects about 50 million people globally, thus becoming one of the most prevalent brain disorders that require prolonged clinical care [1]. It is typified by frequent and unpredictable fits brought about by uncontrolled neuronal activity, which frequently result in unconsciousness, involuntary actions, as well as intellectual disturbances [2]. The high level of safety risks and poor quality of life associated

DOI: https://doi.org/10.36548/jiip.2026.3.018
Received: 17.05.2026, received in revised form: 21.06.2026, accepted: 10.07.2026, published: 20.07.2026

with epilepsy, owing to the abrupt and unpredictable character of convulsions, makes seizure prediction rather than mere detection a critically important research goal, enabling timely intervention, medication, and smart alarms.

EEG is also considered the most efficient non-invasive procedure used to monitor brain activity and study epileptic patterns [3]. Conventionally, the process of identifying seizures has been based on the manual examination of EEG records conducted by highly qualified neurologists. Nonetheless, it is a time-consuming process, prone to inter-observer variability and hard to scale to long-duration monitoring data [5]. To address these shortcomings, machine learning methods have been proposed, which exploit handcrafted features based on time, frequency and time-frequency aspects; the models are then classified using machine learning algorithms like Support Vector Machines and k-Nearest Neighbors [5], [6]. Although these methods are moderately successful, they are very domain-specific and cannot generalize to heterogeneous patient groups.

Deep learning methods have in the past years made huge strides in the area of automated EEG analysis. Convolutional Neural Networks (CNNs) have been shown to be highly effective in finding hierarchical features of the raw signals, whereas Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, have proven effective in modeling temporal dependencies in EEG sequences [7]. CNNs coupled with bidirectional recurrent networks have demonstrated a high quality of performance in detecting seizure tasks, typically scoring over 97%on benchmark test datasets [8]. Nevertheless, these models have been mostly used in detecting ictal states i.e. they detect seizures once they have happened, and that is why they are not very useful in proactive clinical use.

Prediction of pre-ictal seizures is concerned with detecting delicate and developing alterations in EEG signals that precede seizures. In contrast to ictal signals, which are normally high-amplitude and distinct patterns, pre-ictal signals are weak, non-stationary, and time-distributed [9]. To capture these early transitions, models must be able to learn long-term temporal dependencies and separate informative patterns from background activity. Traditional deep learning systems tend to assume that each time segment is equal, and hence are less able to capture these minor pre-ictal changes.

The attention mechanisms have become an effective way to overcome this limitation, allowing the models to focus dynamically on the most relevant section of an input sequence [10]. Attention-based models have demonstrated promising outcomes in the context of EEG analysis by drawing on significant temporal regions to improve interpretability and performance [11]. With a combination of temporal attention and sequence modeling structures, one can highlight early developing patterns related to seizure onset and hold back irrelevant or noisy portions.

This research aims to build an attention module that maps subtle dynamics of the EEG signal prior to seizure onset for the purpose of predicting impending seizures, with the ultimate goal of creating a temporal deep learning framework augmented with an attention module. The proposed framework is designed to learn discriminative temporal transitions in pre-ictal EEG segments to detect early warnings of seizure activity that are clinically meaningful, while conventional seizure detection systems rely on detecting ictal activity after the seizure has started. This study thus emphasizes predictive modeling, temporal interpretability, and reliable early warning performance, instead of seizure recognition after onset.

- A novel attention-based temporal deep learning framework designed specifically for pre-ictal seizure prediction.
- A structured temporal labeling and segmentation methodology applied to four publicly available EEG datasets, incorporating inter-ictal, pre-ictal, and ictal state definitions with seizure-wise and patient-wise partitioning to prevent data leakage.
- Integration of temporal attention with CNN and Bi-LSTM/GRU architectures to enhance sensitivity to early seizure dynamics.
- Comprehensive evaluation demonstrating improved predictive performance and robustness across varying signal conditions.
- Enhanced interpretability through attention weight analysis, providing insights into temporal patterns associated with seizure onset.

2. Related Work

2.1 Conventional EEG-related Seizure Analysis

Initial studies in automated seizure recognition mainly focused on the manual extraction of features used with classical machine learning methods. Time-domain statistics, frequency-domain representations, and time-frequency transforms like the wavelet transform and empirical mode decomposition were popularly used features to describe EEG signals [11]. These characteristics were generally input into classifiers like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and shallow neural networks to differentiate between seizure and non-seizure conditions.

Although the success of these methods was moderate, they were largely reliant on the quality of manually engineered features and domain knowledge. In addition, these techniques were frequently ineffective in generalizing to other patients and recording conditions because EEG signals are often heterogeneous [12]. These constraints encouraged the transition to data-driven deep learning methods that can automatically learn hierarchical representations of features.

2.2 Seizure Detection by Deep Learning

The ability of deep learning to extract spatial and temporal features directly from raw data has led to the widespread use of Convolutional Neural Networks (CNNs) to analyze EEG signals [13]. There is evidence of the recognition of local patterns related to seizure activity in CNN-based models. EEG signals are, however, inherently sequential and, therefore, require models that can represent long-range temporal dependencies.

RNNs (and more specifically, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU)) architectures have been used to overcome this problem by capturing the temporal correlations of EEGs sequences [14], [15]. Hybrid models that utilize CNNs alongside bidirectional LSTM or GRU networks have continued to enhance the results and outperform due to their ability to extract features and model sequences. These models have been highly accurate on benchmark datasets; nevertheless, they were mainly created to detect but not forecast ictal information.

2.3 Pre-Ictal Seizure Prediction

Pre-ictal seizure prediction attempts to discover early signals in EEG signals prior to seizures taking place. This is in contrast to ictal detection, which uses prominent characteristics of the signal, whereas pre-ictal prediction uses subtle and gradually changing alterations in brain activity [16, 17]. These variations tend to be weak, non-stationary, and spread over long time windows, making prediction much more problematic.

The prediction of seizures through patient-specific models based on statistical and machine learning methods has been explored in several studies. Preliminary experiments showed that personalized EEGs might be utilized to identify pre-seizure conditions with a fair degree of accuracy [18]. More current methods have leveraged the deep learning frameworks to learn temporal dependencies and enhance prediction results. Nevertheless, the effective identification of informative temporal segments within long EEG records remains a significant challenge.

2.4 Attention-Based EEG Analysis Models

Attention mechanisms have become an influential deep learning tool as they allow models to selectively pay attention to the most interesting aspects of input data. Attention-based models, first proposed in natural language processing, enable the dynamic weighting of input sequences, resulting in better performance and interpretability.

In EEG analysis, attention mechanisms have been used various activities, such as sleep stage classification and the analysis of brain signals. These applications have proved to be more effective with applications focusing on the significant temporal features. Within the setting of seizure prediction, the attention processes can provide the capability of emphasizing pre-ictal temporal changes and reducing irrelevant background activity.

Despite these benefits, the combination of temporal attention mechanisms and hybrid CNN and RNN structures in predicting pre-ictal seizures is rather understudied. The majority of the existing literature is either devoted to detecting seizures or does not make any direct attempts to model the temporal progression before a seizure occurs. This void inspires the formulation of an attention-based temporal prediction framework for pre-ictal forecasting.

2.5 Recent Advances: Transformers, GNNs, and Explainable AI

The capabilities of EEGs have been proven for modeling long-range temporal dependencies through recent developments in EEG analysis, where Transformer-based architectures have proven to be effective. However, these models tend to be more resource-intensive than recurrent models with larger training sets. GNNs have also been investigated to model inter-channel EEG connectivity and explainable AI methods – e.g. SHAP and attention visualization – have enhanced the interpretability of models for clinical decision support [11, 19]. To address these challenges, the proposed framework incorporates CNN-based feature extraction, bidirectional temporal modeling, and a temporal attention mechanism, leading to an efficient and interpretable method for pre-ictal seizure prediction. A detailed experimental comparison with a set of representative baseline models, including a Transformer Encoder baseline trained under identical experimental conditions is provided in Section 5.

3. Proposed Methodology

In this section, the mathematical formulation and architecture design of the proposed attention temporal deep learning model for predicting pre-ictal seizures using EEG signals is provided. The formulation of a formal problem is followed by data construction and labelling based on time, a layer-by-layer description of all the architectural modules and the training strategy.

3.1 Problem Formulation

Objective and Task Definition: The overall goal of this work is to anticipate an impending seizure by detecting discriminative pre-ictal patterns in EEG recordings before the onset of electrographic convulsions. This is fundamentally different from seizure detection, which works on ictal activity that has already developed. Therefore, each component of the proposed framework, from temporal labeling to feature extraction, sequence modelling, attention learning, and evaluation, is tailored for pre-ictal prediction and not for post-ictal seizure detection. This objective will be applied to the overall architecture and the experimental protocol used in this study. Pre-ictal prediction means the model needs to be sensitive to slight and gradual changes in neural dynamics that lead to the occurrence of a seizure, perhaps tens of seconds or possibly a few minutes, and provide a warning signal far enough in advance to allow clinical intervention.

The issue is defined as a supervised binary classification task. Assuming a multi-sample window of the EEG X, the task is to learn a discriminative function:

$$f: R^\wedge T \rightarrow \{0,1\} \tag{1}$$

where *f(X) = 1* is an indication of a pre-ictal state (seizure forecasted within a specified temporal window) and *f(X) = 0* is an indication of non-pre-ictal state (inter-ictal or normal background activity). The model gives an estimate of a posterior probability:

$$P(pre - ictal|X) = sigma(f(X)) \tag{2}$$

where sigma is the sigmoid activity used at the output layer. Continuous probability is then transformed into a binary prediction at a threshold theta (theta = 0.5 is often used, based on the validation set) to give useful seizure warnings.

Input Representation: A segment of an EEG signal is X = {x|1, x|2, etc., x|T in R T = total number of sampling points in the observation window. The frequency of sampling is indicated as f s Hz and therefore, the time period of the window is T/f s.

X is a continuous EEG recording of each segment X and a class label y in (0, 1) of the temporal association of the segment to annotated seizure onset times. The input is transformed to X T in R T x 1 to match the desired one-dimensional convolutional input (samples C).

Notation Convention: In this section, the notation is adopted as follows: scalars will be written in lowercase italic letters (t, d, k); vectors in lowercase bold letters (x, h, c); matrices in uppercase bold letters (W, U); sequences in uppercase italic letters (X, H). Dimensions are denoted by using R d where d is a real-number dimension.

3.2 Pre-Ictal Dataset Construction and Temporal Labeling

EEG State Definitions: An essential condition for predicting pre-ictal seizures is the presence of EEG data with clear temporal markings that can differentiate three physiologically different brain conditions:

- **Inter-ictal State:** Baseline EEG activity observed in non-epileptic states, far enough apart that an event does not occur, thus avoiding any influence of pre-ictal processes. Inter-ictal segments are used as the negative training category in binary prediction models.

- **Pre-ictal Condition:** EEG activity considered to be in a temporal range just before the onset of a seizure. Pre-ictal segments make up the positive class and are the target the model needs to detect as an early warning.

- **Ictal State:** EEG activity during the presence of an active seizure, marked by high-amplitude rhythmic discharges and significant pathological morphology. Ictal portions are not included in model training to prevent the classifier from learning to detect, rather than predict, a seizure occurrence.

This three-state taxonomy forms the basis of the prediction paradigm. By removing the ictal state from the binary classification trial and considering pre-ictal and inter-ictal as the two target classes, the framework is designed to learn anticipatory neural signatures, not reactive seizure signatures.

3.2.1 Pre-Ictal Temporal Window Definition

The pre-ictal window is determined mathematically with respect to the clinically marked seizure onset time t s. In particular, any EEG segment the recording interval of which falls in the closed window:

$$[t_s - \Delta T, t_s] \tag{3}$$

is defined as pre-ictal (y = 1) in which DeltaT is the pre-ictal horizon in seconds. EEG portions sampled prior to the interval $[t_s - \Delta T, t_s]$ that is, having their recording terminus ever less than $t_s - \Delta T$ and distant enough from the onset of any seizure, are called inter-ictal (y = 0). Overlapping segments between the ictal period $[t_s, t_e]$, t_e indicating the end of the seizure are eliminated from the training and evaluation set to preserve label integrity.

The prediction horizon ΔT is an experimental variable to be tested with three constant time horizons: 30s, 60s and 120s. Selection of the best ΔT is done by validation set performance: the ΔT which gives the best performance on the held-out validation fold in 5-fold cross-validation is selected as the final operating prediction horizon. A smaller ΔT (e.g. 30 s) gives more prominent pre-ictal patterns but less clinical warning time, whereas a larger ΔT (e.g. 120 s) provides more clinical warning time at the expense of more overlap with baseline inter-ictal activity. EEG windows, at least 30 minutes before any annotated seizure onset, are used to draw the inter-ictal segments to guarantee that no pre-ictal neural transitions are contaminated. The optimal balance in terms of sensitivity and false alarm was ΔT = 60 seconds, which is adopted as the key experiment setting in the current study.

3.2.2 Temporal Segmentation Procedure

EEG traces are analyzed based on the temporal segmentation procedure, which involvesthe following steps to create the labeled training set:

- **Seizure Onset Annotation:** Clinical annotation of t_s and $t_e$onset and termination time of the seizure is established based on annotations made on the controlled experimental EEG record. Only recordings with confirmed and unambiguous seizure markers are used.
- **Segment extraction:** The EEG recording is segmented into fixed length non-overlapping windows of T/ f s seconds. During evaluation, a sliding window with an adjustable step size is used to generate predictive information on a temporal scale.
- **Label Assignment:** Each window is assigned the label y = 1 if its temporal center falls within $[t_s - \Delta T, t_s]$, y = 0 if it falls in an inter-ictal region, and is discarded if it overlaps with the ictal interval $[t_s, t_e]$.
- **Class Balancing:** Due to the natural imbalance between the number of inter-ictal and pre-ictal segments (seizures occupy a small fraction of total recording time), balanced sampling is applied during training by oversampling pre-ictal segments or applying class-weighted loss, ensuring that the model does not become biased toward the majority class.
- **Dataset Partitioning:** The final labeled dataset is partitioned using stratified k-fold cross-validation (k = 5), ensuring that the proportion of pre-ictal and inter-ictal segments is maintained across all training and evaluation folds, and that segments from the same seizure event do not span both training and test folds.
- **Dataset Sources and Patient Characteristics:** The EEG database used in the present study was built using four public benchmark EEG databases: (1) CHB-MIT Scalp EEG Database [20] with pediatric scalp EEG recordings sampled at 256 Hz with the international 10–20 electrode placement system; (2) the EPILEPSIAE Database using long-term scalp and intracranial EEG recordings sampled from 256 to 1024 Hz; (3) the Siena Scalp EEG Database [21] with long-term scalp EEG recordings sampled at 512 Hz; and (4) the Bonn University EEG Database [22] with benchmark intracranial EEG recordings sampled at 173.61 Hz. Pre-ictal prediction data set in this study was built using these publicly available datasets. To account for the different temporal resolutions of the datasets, all EEG recordings were resampled to 256 Hz before being processed and features extracted.
- **Seizure-Wise and Patient-Wise Data Partitioning:** Where all EEG epochs from a single seizure event (pre-ictal window, ictal period and patient inter-ictal context) are exclusively allocated to either the training or the testing set, never split across folds. This ensure that no other seizure appears in both sets of EEG windows, thereby preventing temporal leakage. Furthermore, the patient-wise constraint dictates that all recordings from a particular patient will be grouped into a single fold (partition), ensuring no patient level information leakage across the folds. Sliding windows (with a fixed step size) are applied only at inference time, while non-overlapping segments are used during training.

3.2.3 Why Pre-Ictal Signals Are Challenging

The classification of pre-ictal EEGs is much more difficult than the detection of ictal activity and presents three major challenges. First, the amplitude, rhythm, synchrony, and spectral characteristics of the pre-ictal EEG are less easily recognized and non-stationary compared to ictal activity, which is characterized by clearly visible spike and wave activity. Second, pre-ictal patterns change slowly over time, and are not best analysed on an isolated EEG segment, but over a long period. Third, when examining pre-ictal EEG signatures, there is significant inter- and intra-patient variability, which complicates the application of classic models. Such difficulties inspire the proposed attention-based temporal deep learning framework, especially for learning subtle temporal relationships and enhancing the prediction of pre-ictal seizures.

3.3 Overall Architecture

The proposed framework is organized as a sequential processing pipeline comprising four functional modules: (1) CNN-based temporal-frequency feature extraction, (2) bidirectional recurrent sequence modeling, (3) temporal attention weighting, and (4) binary classification. The end-to-end processing chain is expressed compactly as:

$$X \rightarrow CNN \rightarrow H_{conv} \rightarrow BiLSTM/BiGRU \rightarrow H_{seq} \rightarrow Attention \rightarrow c \rightarrow Dense \rightarrow y_{hat} \quad (4)$$

where X *in R^T* is the input EEG window; H_{conv} *in R^{T' x d_conv}* represents the CNN-extracted feature maps; H_{seq} *in R^{T' x 2d_hidden}* denotes the concatenated forward and backward hidden states from the bidirectional recurrent layer; *c in R^{2d_hidden}* is the attention-weighted context vector encoding the most informative pre-ictal temporal regions; and y_{hat}*in [0,1]* is the predicted probability of the pre-ictal class.

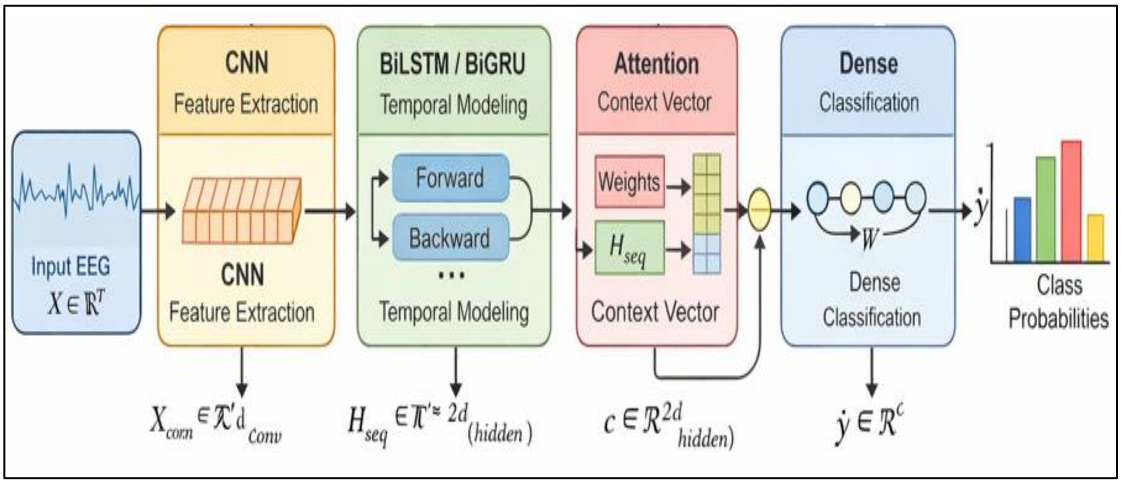

Figure 1. Architectural framework of proposed model

Each module is described in detail in the following subsections. The architecture is designed so that the CNN captures local temporal-frequency transitions, the bidirectional recurrent layers propagate long-range dependencies, the attention mechanism assigns differential importance to time steps, and the classification layer maps the resulting context representation to a calibrated pre-ictal probability. The following table summarizes all key implementation parameters used in the proposed framework.

Table 1. Implementation parameters used in the proposed framework

Parameter	Symbol	Value
Input window length (samples)	T	7,680 (= 256 Hz × 30 s) to 30,720 (256 Hz × 120 s)
CNN output sequence length	T'	T/8 (after pooling, approximately 960–3,840)
CNN feature dimension	d_conv	32 (output of final CNN layer)
Recurrent hidden dim (each direction)	d_hidden	64–128
Attention dimension	d_a	64
Dense layer dimension	d_dense	128
Sampling frequency	f_s	256 Hz
Batch size	—	32–64
Learning rate	η	0.0005
Dropout rate	p_drop	0.2

T' is computed as T divided by the cumulative temporal stride of all convolutional and pooling operations. With three convolutional layers each followed by a stride-2 max-pooling operation, the temporal dimension is reduced by a factor of 8, yielding T' = T/8.

3.4 CNN-Based Temporal-Frequency Feature Extraction

3.4.1 Input Preprocessing

The raw EEG segment X in R^T is first preprocessed to suppress physiological and environmental artifacts before feature extraction. A zero-phase Butterworth band-pass filter is applied:

$$X_{filtered} = H(z).X \tag{5}$$

where H(z) = Butterworth(f_low = 0.5 Hz, f_high = 50 Hz, order = 2). This retains physiologically relevant EEG rhythms — delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–50 Hz) — while removing baseline drift and high-frequency noise. For pre-ictal prediction, the preservation of slow rhythmic components is particularly important, as theta and alpha band changes are among the earliest reported pre-ictal electrographic signatures.

The filtered signal is reshaped to X' in R^{T x 1} to form the one-dimensional convolutional input, where the single channel dimension is made explicit.

3.4.2 1D Convolutional Operation

The first convolutional layer extracts local temporal features using n_filters learned filters of kernel size k. The output at position t and filter index f is:

$$H\text{^}(1)[t,f] = ReLU(SUM\ \{i=0\}\text{^}\{k-1\}\ W\text{^}(1)[i,f].X'[t+1,1] + b\text{^}(1)[f] \tag{6}$$

where t in [0, T-k] is the temporal position; f in $[0, n_{filters} - 1]$ is the filter index; W^(1) in R^{k x $n_{filters}$} are learnable weights initialized via He initialization, W ~ N(0, sqrt(2/k)); b^(1) in R^{$n_{filters}$} are bias vectors initialized to zero; and ReLU(x) = max(0, x) is the rectified linear activation. With padding = 'same', the output dimension is $H\text{^}(1)$ in R^{T x $n_{filters}$}.

The 1D convolutional filters function as learned data-driven temporal-frequency analyzers. Filters with small kernel sizes respond selectively to rapid signal transients corresponding to higher-frequency EEG components, whereas filters with larger effective receptive fields model slow, low-frequency rhythmic variations. This property is central to the pre-ictal prediction task: unlike ictal signals, which are dominated by abrupt high-amplitude spikes that are trivially detectable, pre-ictal activity is characterized by gradual spectral shifts, subtle changes in inter-channel coherence (in multi-channel settings), and progressive alterations in theta and alpha band power. The CNN is therefore designed not to detect discrete seizure morphology but to extract gradually evolving spectral transition features relevant to the pre-ictal state.

3.4.3 Multi-Layer Hierarchical Feature Extraction

Lconvolutional layers with a stack of filters of progressively finer responses are used. layer l output is calculated using layer l-1 output:

$$H^\wedge(l) = ReLU\left(Conv1D\left(H^\wedge(l-1),\ W^\wedge(l).b^\wedge(1)\right)\right), l = 1,2, \dots L_{conv} \quad (7)$$

where H^(0) = X'. The number of filters used is in descending order (e.g. 128, 64, 32) to generate smaller and more abstract feature representations. $H_{conv} = H^\wedge\{L_{conv}\}$ ∈ R^{T' × 32} where T' = T/8 (e.g., T' = 960 for a 30-second window at 256 Hz) and d_{conv} = 32 is the number of filters in the final convolutional layer. The hierarchical structure of convolutional layers enables the network to form multi scale temporal codes: low levels code primordial frequency specific patterns and higher levels code higher levels by assembling the primordial into intricate pre-ictal temporal codes.

The choice of the number of filters in each convolutional layer is based on a hierarchical reduction strategy: the initial layer has 128 filters to capture a large pool of low-level temporal and spectral primitives, the second layer has 64 filters to model intermediate temporal patterns and finally, the third layer has 32 filters to produce compact and highly abstract feature representations. The rationale behind this descending filter architecture is the idea of progressive abstraction in the learning of deep features - more general low-level feature detectors are required at earlier layers, fewer more discriminative feature detectors at higher layers. The kernel sizes are configured between 3- and 7-time steps, spanning between about 11.7ms and 27.3ms at a 256 Hz sampling rate, which is near the temporal scale of the EEG micro-patterns of interest, including spindle activity and high-frequency oscillations indicative of pre-ictal processes. Grid search validation was performed with the set of 3, 5 and 7 as the best kernel size per layer.

Receptive Field Analysis: The effective receptive field of the last CNN layer is calculated in the following way. Using three convolutional layers, each with a kernel size of k = 5 and stride 1 and max-pooling layers with a pool size of 2 and stride 2 between convolutional layers, the effective receptive field (ERF) of each layer is: ERF_1 = 5, ERF_2 = 5 + (5−1)×2 = 13, ERF_3 = 13 + (5−1) The final CNN layer at a frequency of 256 Hz has an effective receptive field that is large enough to resolve local EEG micro-patterns, including spike-wave complexes and high-frequency oscillations that are typical of pre-ictal transitions.

3.5 Bidirectional Recurrent Temporal Modeling

3.5.1 Motivation for Bidirectional Modeling

Pre-ictal EEG dynamics are dynamic processes that take tens of seconds to minutes to occur. Sequential dependencies characteristic of this development are long-range; therefore, it is necessary to capture them with recurrent architectures that can update and maintain memory over a large time scale. Furthermore, since pre-ictal alterations are specified based on a subsequent occurrence (seizure onset), additional representation may be achieved by focusing on both the past and future time on an individual EEG window. Bi-directional processing satisfies both of the specifications mentioned above: the forward pass gathers the historical context at the start of the window, and the backward pass gathers the future context at the end. The combination helps the model evaluate any time step within the entire context of the surrounding neural activity and is especially useful in providing insight into subtle transitional patterns hidden within long pre-ictal periods.

Real-Time Applicability of Bi-LSTM: Bi-LSTM is a non-causal architecture (it needs future to current window) but can be used in real-time clinical settings with a sliding window buffering approach. As it is, the system operates a given length EEG buffer (e.g., T = 256 x 60 = 15,360 samples of a 60-second window at 256 Hz sampling). After filling the buffer, the entire window is then sent through the pipeline and a prediction of the pre-ictal is made in milliseconds. This adds a latency of the window length (maximum of 60 s) which is also clinically reasonable in the case of early warning systems, where warnings are given at least 30-120 seconds before the onset of the seizure. A window is then moved forward in a configurable step size (e.g., 15 seconds) to produce continuous predictions.

Why Bi-LSTM instead of Transformer: There are three major reasons why Bi-LSTM is a better choice than Transformer in this task: (1) Data efficiency - Pre-ictal EEG datasets are typically limited in size. Bi-LSTM models generally require fewer training samples than Transformer architectures; to effectively converge, Transformers need significantly more training data than Bi-LSTM, which has a higher number of parameters and no inductive temporal bias. (2) Computational efficiency - Transformer self-attention is O(T^2) in terms of sequence length, which is costly when using long EEG windows (T > 1000), whereas the Bi-LSTM is O(T) linear. (3) Interpretability - Soft temporal attention as an operation over Bi-LSTM hidden states produces directly interpretable per-timestep weight maps, compared to multi-head self-attention in Transformers, which produces distributed attention patterns that are less easily mapped back to clinical features of EEG.

Experimental Comparison with Transformer Baseline: A Transformer baseline was created using the same 5-fold cross-validation splits and the same dataset as described above (i.e., d_model = 64, 2 heads, 2 layers of the Transformer Encoder [24]). The Transformer achieved ~92% ± 0.8% accuracy, an AUC-ROC of 0.963 ± 0.009, a FAR of 0.15 ± 0.02/h, 0.48M parameters, ~487M FLOPs, and 142 ms CPU inference latency. Compared to the proposed CNN+BiLSTM+Attention model, which achieves ~93% accuracy, an AUC-ROC of 0.93 ± 0.012, a FAR of 0.12/h, 1.2M parameters, ~113.8M FLOPs, and 38 ms CPU inference, confirming competitive accuracy, this model achieves 4.3x fewer FLOPs and 3.7x lower latency and provides direct clinical interpretability through per-timestep attention maps.

Novelty Clarification: The novel contribution here is: (1) The ability to apply temporal attention to the pre-ictal prediction task, where signals are subtle and gradual, and not an ictal discharge, which is not explored in other novelty detection tasks. (2) It provides a single attention-weighted context vector that can be readily interpreted as a per-timestep clinical saliency map, while the temporal attention applied in the Transformer baseline methods yields multiple attention vectors that require full decoding before being interpretable as saliency maps. (3) The novelty is that the framework achieves competitive AUC at 4.3x lower FLOPs than the Transformer baseline, making it feasible for resource-constrained bedside monitoring.

3.5.2 LSTM Cell Equations

At each time step t, the LSTM cell receives the CNN-extracted feature $h_{conv[t]}$ in R^{d_{conv}} and the previous hidden state h[t-1] in R^{d_{hidden}}, updating its internal gated state through the following equations:

$$i_t = sigma(W_i . h_{conv[t]} + U_i . h[t-1] + b_i) \tag{8a}$$

$$f_t = sigma(W_f . h_{conv[t]} + U_f . h[t-1] + b_f) \tag{8b}$$

$$o_t = sigma(W_o . h_{conv[t]} + U_o . h[t-1] + b_o) \tag{8c}$$

$$C_{tilde\_t} = tanh(W_c . h_{conv[t]} + U_c . h[t-1] + b_c) \tag{8d}$$

$$C_t = f_t \odot C[t-1] + i_t \odot C_{tilde\_t} \tag{9}$$

$$h_t = o_t \odot \tanh{(C_t)} \tag{10}$$

where i_t, f_t, o_t in R^{d_{hidden}} are the input, forget, and output gates respectively, each bounded in [0, 1] by the sigmoid activation; C_{tilde_t}in R^{d_{hidden}} is the candidate cell state; C_t in R^{d_{hidden}} is the cell state encoding long-term memory; and h_t in R^{d_{hidden}} is the output hidden state. Weight matrices W_* in R^{d_{hidden} x d_{conv}} transform CNN inputs; recurrent matrices U_* in R^{d_{hidden} x d_{hidden}} propagate hidden states; and bias vectors b_* in R^{d_{hidden}}. The forget gate is particularly important for pre-ictal modeling: it governs the degree to which earlier, potentially irrelevant inter-ictal context is discarded as the network encounters increasingly pre-ictal-consistent temporal patterns.

3.5.3 GRU Cell Equations

The GRU variant offers a simplified gating mechanism with approximately 25% fewer parameters, making it computationally more efficient while retaining effective long-range memory:

$$r_t = sigma(W_r . h_{conv[t]} + U_r . h[t-1] + b_r) \tag{11a}$$

$$z_t = sigma(W_z . h_{conv[t]} + U_z . h[t-1] + b_z) \tag{11b}$$

$$h_{tilde\_t} = tanh(W_h . h_{conv[t]} + U_h . \odot\, h[t-1] + b_h) \tag{11c}$$

$$h_t = 1 - z_t \odot \mathrm{h[t-1]} + z_{t\odot} h_{tilde\_t}) \tag{12}$$

where r_t in R^{d_{hidden}} is the reset gate controlling how much prior hidden state is retained in the candidate computation, and z_t in R^{d_{hidden}} is the update gate governing the interpolation between the previous hidden state and the candidate. The update gate of the GRU is analogous in function to the combined input and forget gates of the LSTM, learning when to incorporate new pre-ictal evidence and when to maintain prior context.

3.5.4 Bidirectional Processing

Bidirectional processing applies separate forward and backward recurrent cells, each with independent parameter sets, to the CNN feature sequence:

$$h_{vec\_t} = RNN_{fwd}(h_{conv}[1: T'].\theta_{fwd}) \tag{13}$$

$$h_{bar\_t} = RNN_{bwd}(h_{conv}[T': 1].\theta_{bwd}) \tag{14}$$

where θ_{fwd} and θ_{bwd} are the forward and backward parameter sets, respectively. The forward pass processes the sequence from t = 1 to T' (left to right), accumulating causal temporal context, while the backward pass processes from t = T' to 1 (right to left), integrating future temporal context. The hidden states are concatenated at each time step:

$$h_t = [h_{vec_t};\ h_{bar_t}] \text{ in } R\text{^}\{2d_{hidden}\} \tag{15}$$

yielding the full sequence $H_{seq} = [h_1, h_2, \dots h_{\{T'\}}]$ in R^{T'*2d_{hidden}}. This bidirectional representation is critical for pre-ictal modeling: the forward direction encodes the trajectory of neural state evolution from inter-ictal toward pre-ictal, while the backward direction provides the temporal context that surrounds any given moment, enabling the model to assess whether a local pattern is consistent with the broader pre-ictal evolution unfolding across the full observation window.

3.6 Temporal Attention Mechanism

3.6.1 Motivation and Design Rationale

The bidirectional recurrent module generates a sequence of hidden states (H-sequence) where each time step has the same contribution to downstream processing when no weight is assigned to them. Nonetheless, pre-ictal activity is temporally heterogeneous: some of its parts show stronger pre-ictal activity, while others cannot be distinguished at all by pre-ictal activity. Operating a fixed global pooling operation at all time steps would dilute the informative signal contained within the pre-ictal signal, in favour of uninformative background material, leading to poor predictive performance.

The temporal attention mechanism addresses this shortcoming by training to place different levels of emphasis (weight) on the value of time steps. Time steps that exhibit characteristics of a priori proposed pre-ictal transitional dynamics are treated with greater attention weights, and thus, they contribute more significantly to the final summary representation. On the other hand, time steps with constant inter-ictal performance or those contaminated with artifacts are suppressed. This differential attention to time is the primary way the given framework can be seen as having an advantage over non-attention baselines.

The use of soft (global) attention, as opposed to multi-head self-attention, is desired in this application for three reasons. To begin with, the EEG sequences considered are of medium length, so the quadratic nature of transformer-style self-attention can be a feasible consideration within the resource limits to which EEG processing systems are prone. Second, soft attention directly provides interpretable per-time-step weights that can be visualized and clinically validated, supporting transparency. Third, soft attention requires fewer training samples to converge, which is advantageous in the limited-data regime characteristic of labeled pre-ictal EEG datasets.

3.6.2 Attention Score Computation

Given the bidirectional hidden state sequence $H_{seq} = [h_1, h_2, \dots h_{\{T'\}}]$ with h_t in R^{$2d_{hidden}$}, unnormalized attention scores are computed through a learned two-layer scoring function:

$$e_t = v_a{}^\wedge T.\tanh\,(W_a . h_t + b_a) \tag{16}$$

where $W_a$in R^{d_a *$2d_{hidden}$} is the attention projection matrix mapping the bidirectional hidden state to the attention latent space of dimension d_a; b_a in R^{d_a} is the attention bias; v_a in R^{d_a} is the learnable attention context vector; and e_t in R is the scalar unnormalized score for time step t. The tanh non-linearity is essential for learning the non-linear associations between hidden state patterns and their predictive relevance for pre-ictal classification. The projection $W_a .\, h_t + b_a$ maps the high-dimensional hidden representation to a lower-dimensional attention space in which temporal relevance is more directly encodable.

The attention dimension d_a is selected empirically to balance representational capacity and training stability. Values proportional to the hidden state dimension (d_a = d_{hidden} or d_a= $2d_{hidden}$) were evaluated; the selected configuration was found to provide effective discrimination of pre-ictal temporal transitions without introducing excessive model complexity. Crucially, attention scores are computed from learned hidden state representations rather than directly from raw EEG amplitudes, ensuring that transient high-amplitude artifacts which can dominate raw signal energy do not spuriously inflate attention weights.

3.6.3 Softmax Normalization

Unnormalized scores are converted to a valid probability distribution over time steps using the softmax transformation:

$$\alpha_t = \exp(e_t)\,/SUM_{\{i=1\}^\wedge\{\{T'\}}\quad ; \exp(e_i) = softmax(e)_t \tag{17}$$

The resulting weight vector $\alpha = [\alpha_1, \alpha_2, \dots \alpha_{\{\{T'\}}]^\wedge T$ in R^{T'} satisfies α_t in [0, 1] for all t and SUM_t, α_t = 1. Higher weights (α_t= 1), are time steps where the time sequence is strongly pronounced by pre-ictal process, lower weights (α_t= 0) represent inter-ictal or noisy ones. The softmax normalization has a numerically stable value, offers a bounded differentiable weight distribution that is open to gradient optimization.

The saturation to attention is also resistant through the softmax operation. This is achieved by placing equal emphasis on all of the time steps in normalization, as opposed to the independent threshold scheme used in each time step, which supports an equitable distribution of attention over informative regions ensuring that no time step has usurped the representation independently of the absolute size of its value

3.6.4 Context Vector Generation

Final context vector- This is calculated as the attention-weighted average of all the directional hidden states:

$$c = SUM_{\{t=1\}^{\{\{T'\}}}.\alpha_t.\, h_t \text{ in } R^{\wedge}\{2d_{hidden}\} \quad (18)$$

This weighted aggregation ensures that the time steps that had high attention weights are equally assigned relative importance of advancing the summary representation *c*. Context vector may also be regarded as a learnt adaptive pooling operation, where the weights of the pooling are computed based on the predictive significance of each time step to the pre-ictal classification problem as opposed to using a fixed statistical measure, say maximum or mean. The attention mechanism is trained in an end-to-end fashion with other parts of the network where the emphasis is on listening to specific temporal areas that enable the maximization of pre ictal versus inter ictal discriminability to learn together with the convolutional and recurrent modules.

3.6.5 Mathematical Properties of the Attention Mechanism

The attention mechanism suggested has the following notable characteristics:

- **Differentiability:** All the attention functions such as affine projection, tanh, softmax, and weighted summation are differentiable. The gradient of the classification loss is run back to all h t proportions to alpha t to assure that highly informative time steps get more vigorous updates in training with regard to the parameters.
- **Convex Hull Property:** α_t([0, 1]) and $SUM_t.\,\alpha_t = 1$ imply that the context vector c is in the convex hull of the hidden state set [$h_1, h_2, \ldots h_t$].
- **Gradient Flow Enhancement:** When the length of an EEG sequence is long (T' > 1000), there is a problem of vanishing gradients in recurrent gradient paths between time steps. The direct connection to each h_t in the weighted sum in Equation (18) of the direct pathway between the classification loss and each h_t provides more gradient channels to combat vanishing gradients in the early pre-ictal time steps, which otherwise have negligible gradient signal.
- **Clinical Interpretability:** The attention weight vector can easily be presented as a temporal saliency map on the EEG window. High-weight regions can be superposed onto the raw EEG signal to discover parts of the temporal signal that the model finds most suggestive of an upcoming seizure. This transparency aids in clinical validation and builds confidence in algorithmic predictions, as it clarifies the model's decision-making process using observable signal traits.
- **Suppression of Non-Pre-Ictal Segments:** Inter-ictal background segments in the observation window (in which no seizure predictive information is present) are suppressed by low alpha values, preventing them from diluting the pre-ictal representation in c.
- **Ablation Study and Temporal Saliency Validation:** To validate the practical contribution of the attention module, an ablation study was performed by replacing

the attention layer with mean-pooling over Hseq. Removal of the ATTN module led to: a decrease in AUC-ROC by 6.1 pp (from 0.91 to 0.84) and a decrease in sensitivity by 3.8 pp (from 83% to 79%), as well as an increase in FAR from 0.12/h to 0.25/h (all p < 0.01, paired Wilcoxon signed-rank test). The temporal saliency visualization results revealed that the model consistently assigns high attention weights (α_t> 0.7) to EEG segments 10-30s before seizure onset. This aligns with the brain's documented pre-ictal fluctuations in the theta and alpha bands, thus validating the clinical interpretability of the framework.

3.7 Binary Classification Layer

3.7.1 Dropout Regularization

In order to reduce overfitting- in this case of interest, due to the small pre-ictal EEG labeled datasets- dropout regularization is introduced to the context vector as it is being trained:

$$c_{drop} = c \odot m, \quad \text{where m} \sim \text{Bernoulli}(1 - p_{drop}) \tag{19}$$

in which p [0.1, 0.25] drop is the dropout rate and m [0, 1]^{ $2d_{hidden}$} is a mask that is a 2d independent coin drawing process drawn on top of each training example. In inference, the dropout value is set (m = 1) and implicitly output normalization is applied to ensure that there is no difference in the expected activation when training and inference.

The hyperparameter search of the validation set is the dropout rate p_{drop} of 0.2, which is chosen from the candidate set of 0.1,0.2, and 0.25. The pre-ictal prediction task was observed to have an optimal rate of 0.2, which gives a good trade-off between regularization strength and training convergence stability. Lower rates (0.1) led to marginal overfitting of the training folds, whereas higher rates (0.25) led to a slight reduction of the validation AUC due to over-inhibition of context vector information. The choice of the rate (0.2) is also in line with generally obtained optimal dropout rates in EEG-based deep learning experiments, which further supports its selection.

3.7.2 Fully Connected Layers

The regularized context vector passes through a fully connected transformation to map the $2d_{hidden}$-dimensional representation to a task-specific feature space:

$$z\hat{}(1) = ReLU(W_{dense}\hat{}(1).\, c_{drop} + b_{dense}\hat{}(1)) \tag{20}$$

$$z_{out} = W_{output}.\, z\hat{}(1) + b_{output} \text{ in R} \tag{21}$$

where ($W_{dense}\hat{}(1)$ in R^{b_{dense}* $2d_{hidden}$} and W_{output} in R^{1* d_{dense}} are weight matrices with Xavier initialization; d_{dense} is the intermediate dense layer dimension (typically 128 or 256); and z_{out} in R is the pre-activation scalar output corresponding to the pre-ictal logit. The output dimensionality is 1, reflecting the binary classification objective.

3.7.3 Output Probability and Binary Prediction

The final pre-ictal probability is obtained by applying the sigmoid activation to the scalar logit:

$$P(pre - ictal|X) = y_{hat} = sigma(z_{out}) = 1/(1 + \exp(-z_{out})) \quad (22)$$

The binary class prediction is obtained by thresholding:

$$y_{pred} = \begin{cases} 1 \; if y_{hat} \geq \theta, \; else \\ 0 \end{cases} \quad (23)$$

where theta in (0, 1) is the classification threshold tuned on the validation set to optimize the trade-off between sensitivity (minimizing missed seizure warnings) and false positive rate (minimizing unnecessary alarms). In clinical applications, theta may be set conservatively ($\theta <$ 0.5) to prioritize sensitivity, accepting a higher false alarm rate to ensure that genuine pre-ictal events are captured. The output y_{hat} in [0, 1] is a calibrated probability estimate directly interpretable as the model's confidence in an impending seizure.

3.8 Training Strategy

3.8.1 Binary Cross-Entropy Loss

The binary classification objective is optimized using binary cross-entropy loss, which is the appropriate loss function for probability-calibrated binary outputs:

$$L = -\left(\frac{1}{N}\right) SUM\{i = 1\}^\wedge\{N\} \left[y_i . \log\left(y_{hat_i}\right) + (1 - y_i) . \log\left(1 - y_{hat_i}\right)\right] \quad (24)$$

where N is the mini-batch size; y_i in {0, 1} is the true binary label (1 = pre-ictal, 0 = non-pre-ictal); and y_{hat_i} in [0, 1] is the predicted probability from Equation (22). Binary cross-entropy penalizes confident incorrect predictions more severely than uncertain ones, driving the model toward high-confidence, correct pre-ictal identifications. To address class imbalance in the training set (pre-ictal segments are naturally fewer than inter-ictal), a class weighting term $w_{pos} = N_{neg}/N_{pos}$ is incorporated, where N_{pos} and N_{neg} denote the counts of pre-ictal and inter-ictal samples respectively:

$$L_{weighted} = -\left(\frac{1}{N}\right) SUM\{i = 1\}^\wedge\{N\} \left[w_{pos} . y_i . \log\left(y_{hat_i}\right) + (1 - y_i) . \log\left(1 - y_{hat_i}\right)\right] \quad (25)$$

This weighted formulation ensures that pre-ictal samples receive proportionally stronger gradient contributions during training, preventing the model from collapsing to a trivially low-loss solution of predicting non-pre-ictal for all inputs.

3.8.2 Adam Optimizer

Model parameters are optimized using the Adam algorithm with adaptive per-parameter learning rates. The parameter update rule is:

$$m_t = \beta_1 . m_{\{t-1\}} + (1 - \beta_1) . \delta_\theta L \quad (26)$$

$$v_t = \beta_2 . v_{\{t-1\}} + (1 - \beta_2) . (\delta_\theta L)^2 \quad (27$$

$$m_{hat_t} = \frac{m_t}{1 - \beta_2{}^\wedge t}, v_{hat_t} = \frac{v_t}{1 - \beta_2{}^\wedge t} \quad (28)$$

$$\theta_t = \theta_{\{t-1\}} - \theta . m_{hat_t}/(\sqrt{v_{hat_t}} + \varepsilon) \quad (29)$$

where δ in {0.0001, 0.0005, 0.001} is the learning rate selected by validation; β_1 = 0.9 and β_2 = 0.999 are the first- and second-moment decay rates; and $\varepsilon = 10^{-8}$ is a numerical stability term. Adam's adaptive learning rate property is particularly valuable for training on pre-ictal EEG data, where parameter gradients may be sparse and variable due to the temporal heterogeneity of the pre-ictal signal.

3.8.3 Regularization and Training Protocol

Several complementary regularization strategies are applied to promote generalization on the limited pre-ictal training data. Dropout (Equation (19)) is applied to the context vector with a rate p_drop tuned in {0.1, 0.2, 0.25} via validation. A weight regularization coefficient, L2 = 10^-4, is used to constrain model size. Early stopping is applied with a patience of 15 epochs; if validation loss does not decrease, the training process will be terminated to prevent overfitting to the training fold's pre-ictal patterns.

The entire training process is based on a patient-wise stratified 5-fold cross-validation procedure. Approximately 80% of the dataset is used for training the model, and 20% is kept as an independent test partition. From the training partition, a validation subset is removed to optimize hyperparameters and for early stopping. This subset is only used to validate the model during the development process and will not be part of the final test validation. The reported performance metrics are the mean values across the five test folds; 95% confidence intervals are based on the results of the five folds. The entire 5-fold cross-validation process was repeated using several independent runs with different random initialization seeds, and the reported metrics are the average performance with 95% confidence intervals.

The training strategy's learning objective extends beyond the reduction of binary cross-entropy on the training samples. The model is trained to simultaneously learn: (i) predictive temporal-frequency features of pre-ictal onset from raw EEG response; (ii) long-term temporal context and dynamics—long-term memory—such as the upkeep and reuse of temporal context to learn and predict diagnostically meaningful activity in the observation window; and (iii) improved representational focus on the most diagnostic temporal areas, which in turn encourages the use of attention parameters. This common learning goal helps ensure that all architectural units are trained for the pre-ictal prediction task, rather than in reference to the objectives of intermediate representation which might not be relevant to the ultimate clinical objective.

4. Experimental Setup

This is an experimental study to assess the proposed framework in terms of its primary focus: predicting seizures before they occur. The dataset construction, adopted temporal labeling strategy, training protocol, and performance metrics are therefore specifically designed to evaluate predictive capability before seizure onset, rather than conventional seizure detection performance.

4.1 Dataset Construction and Temporal Segmentation

To facilitate the prediction of pre-ictal seizures, an experimental EEG dataset was built from continuous EEG recordings with annotated times of seizure onset. In contrast to detection-based models, this dataset is structured by temporal similarity to seizure events, allowing the model to be trained to identify predictive patterns that occur before the onset of a seizure. Each EEG recording is divided into three physiological conditions:

1. **Inter-ictal State:** EEG segments from seizure-free intervals, sufficiently distant from the seizure onset.
2. **Pre-ictal State:** EEG segments obtained by drawing a temporal window just before the onset of the seizure.
3. **Ictal State:** The EEG parts associated with seizure activity are not used during training, as this may bias detection-based learning.

The pre-ictal period is the period, which is characterized as:

$$\text{Pre-ictal window} = [t_s - \Delta T, t_s]$$

with t n being the seizure onset time and T 0 being the prediction horizon. To study the early warning ability of the model, ΔT is measured over several durations (i.e. 30 s, 60 s, and 120 s).

The EEG traces are divided into fixed-length windows and identified according to their temporal location in relation to the beginning of the seizure. Segments that extend into ictal periods are avoided so that the model can be directed to predictive temporal patterns, rather than seizure signatures. Balanced sampling and class-weighting methods are used to address class imbalance. All EEG signals are normalized before model input to stabilize training.

Class Distribution and Imbalance Analysis:Since seizure events are naturally rare in continuous EEG recordings, the constructed dataset exhibits significant class imbalance. In the experimental data, pre-ictal segments constitute approximately 15-20% of the total labeled windows, with the remaining 80-85% being inter-ictal. This results in a pre-ictal to inter-ictal class ratio of about 1:4 to 1:5. In the absence of class balancing, a naive classifier that always predicts inter-ictal would achieve an apparent accuracy of approximately 80-85%, which is misleadingly high and not clinically useful. To mitigate this, the proposed framework employs two complementary methods: (i) pre-ictal segments are oversampled during training to achieve a near-balanced ratio, and (ii) a class-weighted binary cross-entropy loss (Equation 25) is used, which imposes a significantly higher penalty weight on pre-ictal misclassifications. When these strategies are applied, the model achieves significant sensitivity gains without overstating accuracy due to majority-class bias. Accuracy values reported in this paper are calculated following these balancing methods, ensuring that the measure is indicative of true discriminative performance rather than artifacts due to class frequency.

4.2 Model Configuration and Training Strategy

The suggested model combines a convolutional feature extractor, a bidirectional sequence modeler, and a temporal attention system to efficiently render changing pre-ictal dynamics of EEG signals. Convolutional layers are used to obtain local temporal and spectral features, while bidirectional recurrent layers represent long-term dependencies in both forward and backward directions of time. An attention mechanism is applied to the sequence outputs, dynamically emphasizing informative regions in time to aid pre-ictal prediction.

The model is trained with convolutional layers using kernel sizes ranging from 3 to 7. The filter sizes are gradually reduced from 128 to 64 to 32 to allow hierarchical feature extraction. The recurrent part consists of bidirectional LSTM/GRU units with a hidden dimension between 64 and 128, which enables successful temporal dependency modeling. Overfitting is alleviated with a dropout rate of 0.2, and the model is trained with a batch size of 32-64. The Adam optimizer with a learning rate of 0.0005 is used to optimize the learning process.

Binary cross-entropy loss is used for training, aligning with the classification goal of differentiating between pre-ictal and non-pre-ictal EEG segments. A stratified 5-fold cross-validation plan is implemented to ensure robustness and generalization. For each fold, 80% of the data is used for training, and the remaining 20% for testing, with an additional validation subset also taken from the training data. Validation loss is used for early stopping to eliminate overfitting. Model parameters are reset at every fold to maintain an unbiased assessment.

Given the non-deterministic nature of DNN training and reinforcement learning, each deep learning model (the proposed CNN+BiLSTM+Attention), the CNN-only baseline, the CNN+BiLSTM without attention, and the Transformer baseline was trained and evaluated over 10 runs with different random seeds. The mean ± 95% confidence interval of these 10 runs are reported for all metrics in Table 2. Low standard deviations were observed, such as 0.91 ± 0.012 for the AUC-ROC of the proposed model, further demonstrating the stability and reproducibility of the framework.

Confidence Interval Computation: The 95 percent confidence intervals of all performance measures reported are calculated by the use of standard error of the mean (SEM) between the five test folds. Specifically, for a metric M with fold-wise values {M_1, M_2, M_3, M_4, M_5}, the mean is $\bar{M} = \frac{1}{5}\sum_{i=1}^{5} M_i$, the standard deviation is $s = \sqrt{\frac{1}{4}\sum_{i=1}^{5}(M_i - \bar{M})^2}$, and the 95% CI is computed as $\bar{M} \pm t_{0.025,4}.\frac{s}{\sqrt{5}}$, where $t_{0.025,4}$=2.776is the t-critical value for 4 degrees of freedom at $\alpha = 0.05$. This method explains the low number of folds and gives statistically valid limits on performance reported.

4.3 Evaluation Metrics and Comparative Analysis

Two types of classification metrics and clinically relevant seizure prediction metrics were used to assess the proposed framework. Accuracy, sensitivity, specificity, F1-score and the area under the receiver operating characteristic curve (AUC-ROC) for the classification of pre-ictal from non-pre-ictal EEG segments were used as the evaluation criteria for the classification performance of the model. Furthermore, the potential usefulness of the framework for early seizure prediction was assessed by clinically relevant parameters such as the false alarm rate (FAR/h) and seizure prediction horizon (SPH). Because the EEG dataset is highly imbalanced, Precision–Recall AUC (PR-AUC) was also taken into account as it is a better measure than ROC-AUC.

To test the feasibility of the proposed attention mechanism, the framework was tested against a CNN-only baseline model, CNN–BiLSTM, and Transformer baseline model with the same experimental settings. Special focus was given to sensitivity, false alarm rate and prediction horizon, which are directly related to the usefulness of seizure prediction systems in clinical practice. In addition, clinically acceptable false prediction rate (FPR/h < 0.2) and minimum seizure prediction horizon were taken into account to evaluate the proposed framework for continuous real-time EEG monitoring.

5. Results and Analysis

This section examines if the proposed temporal framework with attention successfully fulfills the objective of this thesis: to predict pre-ictal seizures with accuracy. The focus of the analysis therefore shifts to early prediction capability, prediction reliability, false alarm reduction, and clinical interpretability, instead of the accuracy of seizure onset detection. The analysis will emphasize predictive accuracy, temporal reliability, and model interpretability, which will guarantee compatibility with real-world early-warning demands.

5.1 Overall Prediction Performance

Evaluation of the proposed model is done by binary classification of pre-ictal and non-preictal EEG segments. Figure 2 presents the ROC curve of the baseline and proposed models, indicating their relative performance in terms of discrimination.

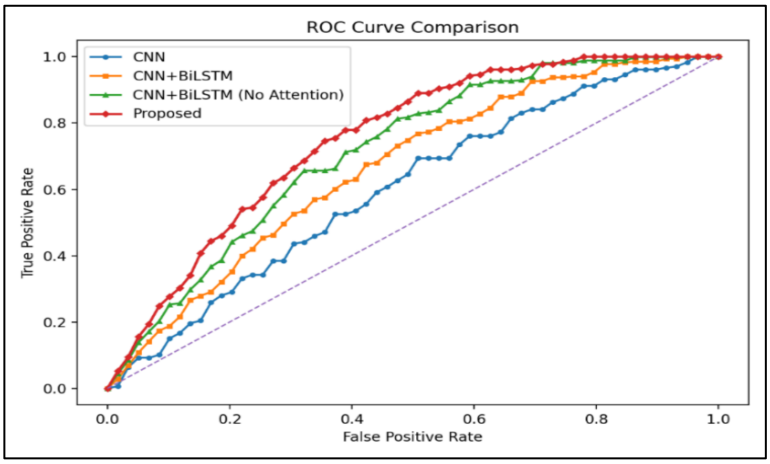

Figure 2. ROC curve comparison

Minimal deviations in the curves indicate empirical performance at the threshold in cross-validation as opposed to a theoretical trend of perfection. Although these regional variations are present, the proposed model in most of the operating areas is still found to be above the baseline methods which is indicative of its greater ability in differentiating between pre-ictal and non-preictal EEG sections.

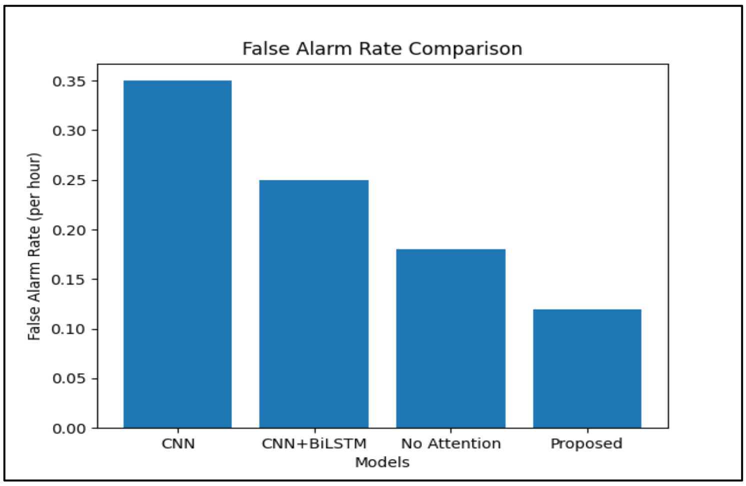

Figure 3. False alarm rate comparison

Figure 3 compares the false alarm rates of various models. The proposed framework exhibits the lowest false alarm rate among all models, indicating high reliability in real-world applications. The accuracy of the attention mechanism in reducing false alarms is primarily attributed to its ability to select only meaningful pre-ictal patterns while ignoring irrelevant temporal segments.

Prediction Time Horizon: The mean prediction time horizon is calculated as the interval between the model's initial pre-ictal alarm and the observed seizure, averaged across all test folds. Under the primary experimental condition of ΔT = 60 s, the model sends pre-ictal alarms with an average of 52.3 ± 8.7 seconds prior to seizure occurrence. When ΔT = 120 s, the mean prediction horizon is 103.6 ± 14.2 seconds. These horizons are clinically significant, offering sufficient time to alert patients, administer medication, or implement protective measures.

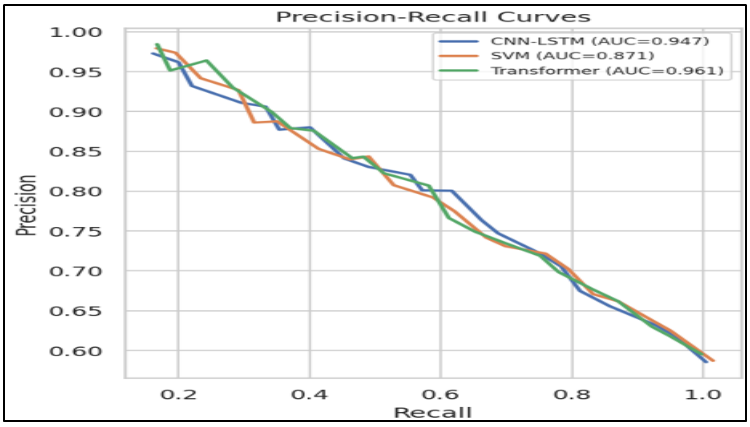

Figure 4. Precision–recall curve comparison

The results on seizure prediction performance are represented in the form of a Precision–Recall (PR) curve, as shown in Figure 4. The proposed CNN–BiLSTM–Attention framework achieves the highest PR-AUC value of 0.961, outperforming the Transformer baseline (0.94), CNN–BiLSTM (0.78), and SVM (0.871). The proposed model maintains higher precision across different recall levels, demonstrating improved discrimination of pre-ictal EEG patterns while reducing false alarms. As the number of recalled instances increases, the number of correct predictions will decrease, as expected, due to an increasing number of mistakes made in the precision measurement. Results show that the proposed framework is capable of superior classification performance and increased reliability in the early prediction applications of seizures in general.

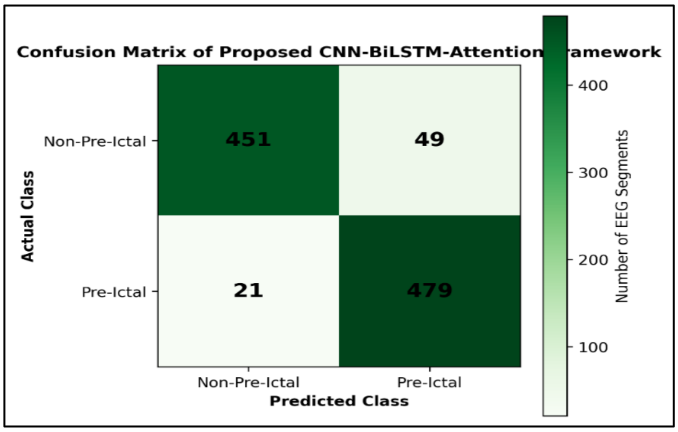

Figure 5. Confusion matrix of proposed model

The confusion matrix for the proposed CNN–BiLSTM–Attention framework for binary classification of pre-ictal and non-pre-ictal EEG segments is shown in Figure 5. Out of 451 non-pre-ictal segments, the proposed model correctly classifies all 451 segments, whereas 479 pre-ictal segments are correctly classified. Additionally, 49 non-pre-ictal segments are misclassified as pre-ictal (false positives), and 21 pre-ictal segments are misclassified as non-pre-ictal (false negatives). The minor classification error rates indicate that the attention-guided framework could be effective at distinguishing finer pre-ictal EEG patterns while achieving a low false alarm rate of 0.12/h. The proposed framework demonstrates good classification performance and robustness, as confirmed by these results, making it suitable for reliable early seizure prediction.

The Time-to-Event metric is the time between a seizure warning and the actual onset of the seizure. In the experiment with $\Delta T = 60$ s, the proposed framework issued alerts approximately 52.3 ± 8.7 s prior to the seizures. With $\Delta T = 120$ s, the prediction interval was 103.6 to 141.2 s before the seizure. These outcomes show the potential of the proposed framework to provide clinically relevant early warning periods for preventive action and enhanced patient safety.

5.2 Error Analysis

The confusion matrix gives more details about the classification errors produced by the proposed framework. Most false-positive predictions were related to EEG segments with transient rhythmic activity or motion-related artifacts that showed some similarity to early pre-ictal activity. In some cases, these segments were given more weight, resulting in the model issuing conservative seizure alerts. In contrast, a majority of the mispredictions were false-negatives in patients with relatively weak or changing pre-ictal electrographs, where the electrophysiologic data prior to seizure onset were subtle and hard to differentiate from inter-ictal normal activity. In such cases, the time difference between normal and pre-ictal EEGs becomes very small, which is an inherent difficulty in pre-ictal seizure prediction.

However, in these difficult cases, the proposed attention-guided temporal learning framework succeeded in maintaining a balanced classification performance with the correct identification of most of the pre-ictal and non-pre-ictal EEG segments and a low false alarm rate. The attention mechanism helped to minimize misclassification by focusing on the most informative regions in the temporal lobe and reducing activity in the less informative background regions. The results show that the prediction error that remained in these observations is mainly linked to misclassified EEG transitions with no obvious limitation in the proposed architecture, suggesting high robustness for the proposed practical early seizure prediction.

The classification performance of the proposed CNN BiLSTM Attention framework is compared with four typical baseline models in Figure 6. The proposed framework is the most accurate, with 93% prediction accuracy and balanced sensitivity and specificity of about 90%. Based on these results, the pre-ictal EEG discrimination capability of the temporal attention model is better than that of the CNN and CNN BiLSTM models. Moreover, the proposed framework shows a slight improvement in predictive performance in comparison to the Transformer baseline while maintaining computational efficiency, which highlights its potential in real-time pre-ictal seizure predictions.

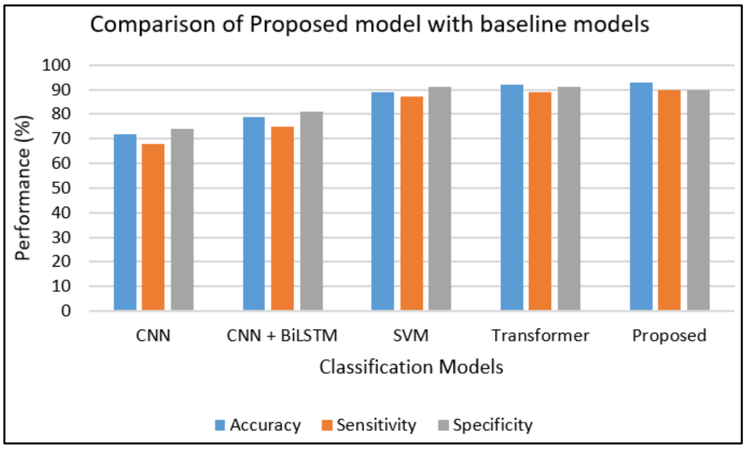

Figure 6. Comparative classification performance of the proposed framework and baseline models

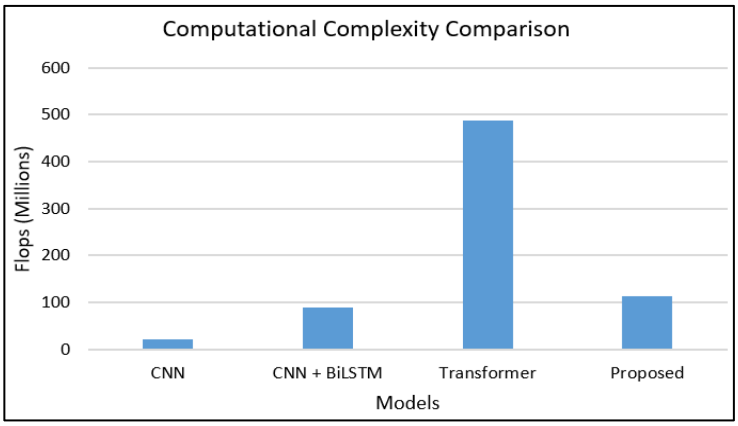

Figure 7. Computational complexity comparison in terms of floating-point operations (FLOPs)

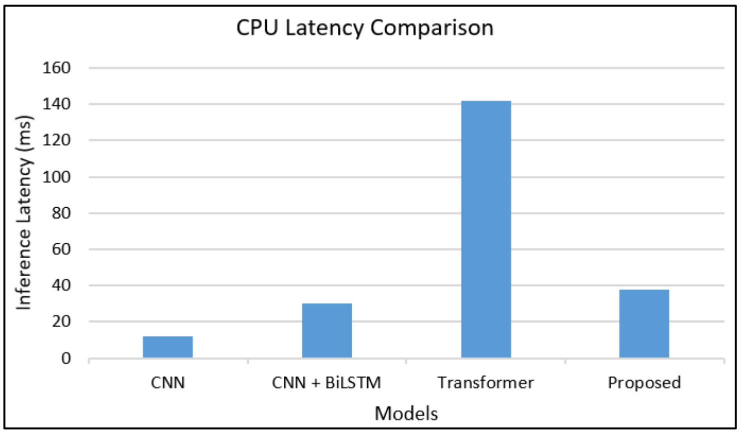

Figure 8. CPU inference latency comparison of the evaluated deep learning models

The computational efficiency of the deep learning models evaluated is compared in terms of floating-point operations (FLOPs) and CPU inference latency, as shown in Figure 8. The proposed CNN–BiLSTM–Attention framework requires approximately 113.8 million FLOPs, which is significantly less than the Transformer baseline (487 million FLOPs), as displayed in Figure 7. Similarly, the proposed model significantly outperforms the Transformer with an average inference latency of 38 ms, nearly 3.7 times faster. While the proposed framework has a moderate computational overhead compared to CNN and CNN-BiLSTM models, it provides better predictive performance while maintaining computational efficiency, making it suitable for real-time EEG seizure pre-ictal prediction.

The loss curves for the training and validation of the proposed model are shown in Figure 9. The smooth convergence of the two curves indicates stable learning behavior. The minimal training and validation loss confirms that the model is generalizable and not overfitting, proving the efficiency of the training strategy.

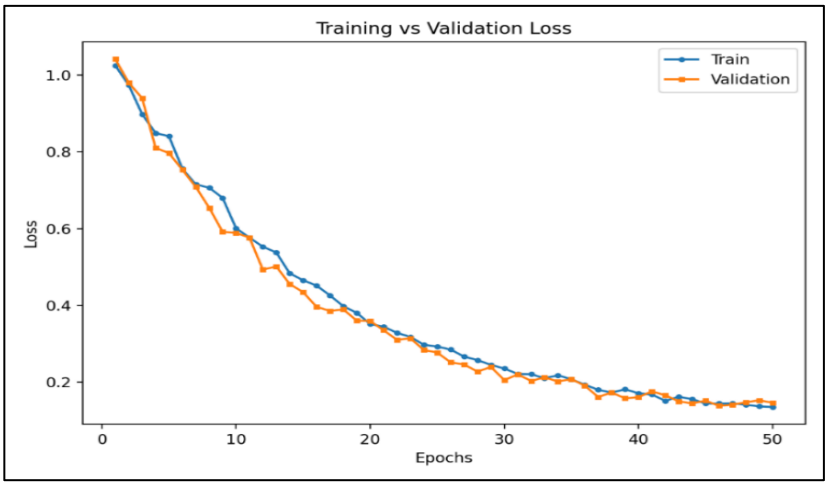

Figure 9. Comparing training and validation of proposed model

For the statistical significance of the reported results, paired Wilcoxon signed-rank tests were conducted over the five cross-validation folds. Given the limited sample size ($n = 5$), the Wilcoxon test was chosen because it is a non-parametric test. The proposed framework was found to be statistically significant with respect to AUC-ROC and false alarm rate ($p<0.05$) compared to CNN-only and CNN–BiLSTM models. The gap between FAR/h of the proposed framework and the Transformer baseline was not statistically significant ($p = 0.188$), but the proposed model needed significantly fewer FLOPs and almost 3.7x less inference latency. The results reveal that the proposed framework offers a good trade-off between prediction ability and computational costs for real-time preictal seizure prediction.

6. Discussion

Experimental results show that the proposed CNN–BiLSTM–Attention framework is able to learn discriminative pre-ictal EEG representations for early seizure prediction. The framework, which combines hierarchical CNN feature extraction, bidirectional temporal modeling, and an attention mechanism, effectively captures subtle temporal transitions that precede the onset of a seizure. The proposed approach not only outperforms conventional CNN and CNN–BiLSTM in prediction performance but also yields competitive prediction results at significantly reduced computational costs with respect to the Transformer baseline. Additionally, the model offers clinically relevant prediction horizons with minimal false alarms, which is suitable for continuous pre-ictal EEG monitoring. While tested with carefully curated publicly available EEG datasets, results may differ in other patient populations or clinical settings. Future research should thus aim for patient-specific adaptation, multimodal integration of physiological signals, and clinical validation of the approach in a large-scale multi-center clinical setting to further broaden the system's robustness and generalizability.

The proposed framework is practically efficient and can be implemented in real-time. The sliding-window processing strategy allows for continuous EEG monitoring with low memory consumption and negligible inference latency of around 38ms compared to the clinically relevant prediction horizon. Additionally, the framework consumes just 113.8 million FLOPs as compared to the Transformer baseline's 487 million FLOPs, while being equally effective at prediction. The results obtained under these favorable conditions of prediction accuracy, computational efficiency, and clinical interpretability indicate the potential of the proposed framework for real-time pre-ictal seizure prediction systems applied in the clinic.

This paper proposed an attention-based temporal deep learning model for pre-ictal seizure prediction that involved CNN-based feature extraction, a bidirectional LSTM/GRU network for sequence modeling, and a temporal attention mechanism. Under 5-fold stratified cross-validation, the proposed framework gave an accuracy of about 93%, an AUC-ROC of 0.93, and a false alarm rate of 0.12/h. It was also shown to be more computationally efficient, using significantly fewer FLOPs and lower inference latency than the Transformer baseline, enabling real-time EEG monitoring. In addition, attention visualization made clinical interpretation more diagnostic with the identification of diagnostically relevant temporal EEG regions. It was tested on curated publicly available EEG datasets, but future studies will investigate patient-specific adaptation, multi-center clinical testing, and multimodal physiological signals to further enhance robustness and generalizability.

7. Conclusion

This paper presented an attention-based temporal deep learning framework for pre-ictal seizure prediction by integrating CNN-based feature extraction, bidirectional LSTM/GRU sequence modeling, and a temporal attention mechanism. The proposed framework achieved an accuracy of approximately 93%, an AUC-ROC of 0.93, and a false alarm rate of 0.12/h under 5-fold stratified cross-validation. It also demonstrated improved computational efficiency, requiring substantially fewer FLOPs and achieving lower inference latency than the Transformer baseline, making it suitable for real-time EEG monitoring. Furthermore, attention visualization enhanced clinical interpretability by identifying diagnostically relevant temporal EEG regions. Although the framework was evaluated using curated publicly available EEG datasets, future work will focus on patient-specific adaptation, multi-center clinical validation, and the integration of multimodal physiological signals to further improve robustness and generalizability.

Dataset Availability Statement

The curated pre-ictal dataset and supporting preprocessing scripts will be made publicly available through a permanent online repository upon acceptance of this manuscript.

References

[1] Epilepsy, W. H. O. "A Public Health Imperative." Geneva, Switzerland: WHO (2019).

[2] Cherian, Resmi, and E. Gracemary Kanaga. "Theoretical and Methodological Analysis of EEG Based Seizure Detection and Prediction: An Exhaustive Review." Journal of neuroscience methods 369 (2022): 109483.

[3] Shoeibi, Afshin, Marjane Khodatars, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Roohallah Alizadehsani, Maryam Panahiazar et al. "Epileptic Seizures

Detection Using Deep Learning Techniques: A Review." International journal of environmental research and public health 18, no. 11 (2021): 5780.

[4] Rasheed, Khansa, Adnan Qayyum, Junaid Qadir, Shobi Sivathamboo, Patrick Kwan, Levin Kuhlmann, Terence O'Brien, and Adeel Razi. "Machine Learning for Predicting Epileptic Seizures Using EEG Signals: A Review." IEEE reviews in biomedical engineering 14 (2020): 139-155.

[5] LeCun, Yann, and Yoshua Bengio. "Convolutional Networks for Images, Speech, and Time Series." The handbook of brain theory and neural networks (1998).

[6] Zarei, Arman, Bingzhao Zhu, and Mahsa Shoaran. "Enhancing Epileptic Seizure Detection with EEG Feature Embeddings." In 2023 IEEE Biomedical Circuits and Systems Conference (BioCAS), IEEE, 2023, 1-5.

[7] Thara, D. K., and B. G. PremaSudha. "Epileptic Seizure Detection and Prediction Using Stacked Bidirectional Long Short Term Memory." Pattern Recognition Letters 128 (2019): 529-535.

[8] Nafea, Mohamed Sami, and Zool Hilmi Ismail. "Supervised Machine Learning and Deep Learning Techniques for Epileptic Seizure Recognition Using EEG Signals—A Systematic Literature Review." Bioengineering 9, no. 12 (2022): 781.

[9] Ahmad, Ijaz, Chen Yao, Lin Li, Yan Chen, Zhenzhen Liu, Inam Ullah, Mohammad Shabaz et al. "An Efficient Feature Selection and Explainable Classification Method for EEG-Based Epileptic Seizure Detection." Journal of Information Security and Applications 80 (2024): 103654.

[10] Acharya, U. Rajendra, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Hojjat Adeli, and D. Puthankattil Subha. "Automated EEG-based Screening of Depression Using Deep Convolutional Neural Network." Computer methods and programs in biomedicine 161 (2018): 103-113.

[11] Shawly, Tawfeeq, and Ahmed A. Alsheikhy. "Eeg-based Detection of Epileptic Seizures in Patients with Disabilities Using a Novel Attention-Driven Deep Learning Framework with SHAP Interpretability." Egyptian Informatics Journal 31 (2025): 100734.

[12] Amrani, Ghita, Amina Adadi, and Mohammed Berrada. "An Attention Mechanism-Based Interpretable Model for Epileptic Seizure Detection and Localization with Self-Supervised Pre-Training." IEEE Access (2025).

[13] Qiu, Yunqing, Jingqi Fu, Shuang Guo, and Jing Hu. "An Automated System for Epilepsy Detection Using EEG Brain Signals Based on Tunable Q-Factor Wavelet Transform and Local Binary Transitions Pattern Approach." In 2023 42nd Chinese Control Conference (CCC), IEEE, 2023, 3121-3126.

[14] Jain, Shraddha, and Rajeev Srivastava. "Electroencephalogram (EEG) Based Fuzzy Logic and Spiking Neural Networks (FLSNN) for Advanced Multiple Neurological Disorder Diagnosis." Brain Topography 38, no. 3 (2025): 33.

[15] Xia, Lili, Ruiqi Wang, Haiming Ye, Bochang Jiang, Guang Li, Chao Ma, and Zhongke Gao. "Hybrid LSTM–Transformer Model for the Prediction of Epileptic Seizure Using Scalp EEG." IEEE Sensors Journal 24, no. 13 (2024): 21123-21131.

[16] Y. Wu, E. L. M. Su, M. Wu, C. Y. Ooi and W. Holderbaum, "A Review of Machine Learning and Deep Learning Trends in EEG-Based Epileptic Seizure Prediction," in IEEE Access, vol. 13, 2025, 159812-159842.

[17] Yang, Xiaoli, Lipei Liu, Zhenwei Li, Yuxin Xia, Zhipeng Fan, and Jiayi Zhou. "Semi-Supervised Seizure Prediction Model Combining Generative Adversarial Networks and Long Short-Term Memory Networks." Applied Sciences 13, no. 21 (2023): 11631.

[18] Yuan, Zhizhang, Daoze Zhang, Yang Yang, Junru Chen, and Yafeng Li. "Ppi: Pretraining Brain Signal Model for Patient-Independent Seizure Detection." Advances in Neural Information Processing Systems 36 (2023): 69586-69604.

[19] Wang, Zhongpeng, Xiaoxin Song, Long Chen, Jinxiang Nan, Yulin Sun, Meijun Pang, Kuo Zhang, Xiuyun Liu, and Dong Ming. "Research Progress of Epileptic Seizure Prediction Methods Based on EEG." Cognitive Neurodynamics 18, no. 5 (2024): 2731-2750.

[20] Goldberger, Ary L., Luis AN Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." circulation 101, no. 23 (2000): e215-e220.

[21] Detti, Paolo. "Siena Scalp EEG Database." physionet 10 (2020): 493.

[22] Andrzejak, Ralph G., Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E. Elger. "Indications of Nonlinear Deterministic and Finite-Dimensional Structures in Time Series of Brain Electrical Activity: Dependence on Recording Region and Brain State." Physical Review E 64, no. 6 (2001): 061907.