SeizureFormer: A Multi-Scale Transformer for Seizure Risk Forecasting from RNS-Derived Biomarkers

Tianning Feng[3†], Juntong Ni[1†], Wei Jin[1], Ezequiel Gleichgerrcht[2]
[1]*Department of Computer Science, Emory University, Atlanta, GA, USA*
[2]*Department of Neurology, Emory University, Atlanta, GA, USA*
[3]*School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA*
[†]*E-mail: tfeng24@seas.upenn.edu, juntong.ni@emory.edu*
E-mail: wei.jin@emory.edu, eze.gleich@emory.edu

We present **SeizureFormer**, a Transformer-based model for long-horizon seizure risk forecasting (1–14 days) using structured biomarkers—interictal epileptiform activity (IEA) and long episodes (LE)—extracted from responsive neurostimulation (RNS) systems. Unlike prior models based on raw scalp EEG, SeizureFormer leverages stable RNS biomarkers and integrates multi-scale CNN patch embedding, cross-variable temporal convolution, and squeeze-and-excitation attention to capture both short-term fluctuations and long-term seizure cycles. Tested across five patients and multiple prediction windows (1–14 days), SeizureFormer achieved state-of-the-art performance with **mean ROC AUC** of *79.44%* and **mean PR AUC** of *76.29%* across five patients and four prediction windows. Compared to statistical, classical ML, and deep learning baselines, it demonstrates superior generalizability under class imbalance. Clinically, it enables actionable multi-day forecasting, supporting personalized and proactive intervention in epilepsy care by forecasting seizure-related events 1 to 14 days ahead.

Keywords: Seizure Forecasting; Epilepsy; Transformer; Deep Learning; RNS System

1. Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent and unpredictable seizures, affecting over 50 million people worldwide.[1,2] A central clinical need is the ability to detect or forecast seizure risk in advance, enabling timely intervention and improved patient management.[2,3]

Seizure risk detection generally falls into two categories: seizure classification and seizure forecasting. Early research predominantly focused on seizure classification, which aims to detect ongoing or imminent seizures by identifying ictal or pre-ictal segments from recent EEG recordings. In contrast, seizure forecasting aims to estimate the probability of future seizure occurrence over longer horizons-often days in advance-without access to EEG signals from the forecasted time window, relying solely on historical neural activity.[4,5] While classification supports reactive interventions, forecasting offers a critical opportunity for preventive care and long-term planning.

Early efforts in seizure forecasting primarily relied on scalp EEG recordings combined

with statistical or deep learning models.[4,5] However, scalp EEG signals are inherently non-stationary and highly susceptible to physiological and environmental factors such as brain state transitions, medication effects, fatigue, and electrode shifts.[6,7] These fluctuations make it difficult to model consistent long-term patterns, often requiring frequent retraining and resulting in poor generalization across patients—especially for multi-day prediction settings.[8]

As an alternative, responsive neurostimulation (RNS) systems extract seizure-related signals such as interictal epileptiform activity (IEA) and long episodes (LE), which are validated seizure risk indicators that reflect sustained EEG abnormalities associated with cortical excitability.[9,10] These structured features offer a more stable and interpretable basis for seizure risk assessment compared to scalp EEG. Recent studies leveraging RNS biomarkers have explored statistical,[8] machine learning, and deep learning methods[11] for seizure risk forecasting, achieving promising results yet several critical limitations remain:

- **Limited multi-scale modeling ability:** RNS data exhibits multi-scale temporal patterns, including daily, weekly, and potentially longer-term rhythms.[12] However, existing work often fails to capture these complex temporal dependencies, focusing either on narrow time windows or specific periodicities.
- **Limited capacity to model both short-term and long-term dependencies:** Seizure risk is influenced by both immediate neural fluctuations and long-range historical patterns. While statistical models effectively capture long-term trends, they lack the resolution for short-term dynamics. Conversely, common deep learning models excel in local pattern recognition but struggle with integrating seizure periodicities across time scales.[8,11] A unified approach that can jointly model both short-term variability and long-term seizure cycles is needed.
- **Overreliance on raw RNS data, leading to high personalization requirements:** Many existing models rely heavily on raw RNS recordings, which are noisy, complex, and patient-specific. This often requires extensive per-patient training and limits generalizability.[9,10] Shifting towards more structured and stable biomarkers could improve model scalability and clinical applicability.

To address these challenges, we propose SeizureFormer—a deep learning model that jointly captures multi-scale seizure dynamics, bridges short- and long-term dependencies, and reduces reliance on raw patient-specific data. In essence, we are faced with three key challenges. **First, how to extract seizure-relevant features from time-series signals with varying periodicity?** We address this by employing a CNN-based patch embedding module with multiple kernel sizes, allowing the model to capture both short-term fluctuations and long-term cycles from the input signals. Additionally, we introduce a Cross-Variable Temporal 2D Conv (CVT-Conv) module to explicitly model local interactions and aggregate contextual information across different A+B pattern channels. The integration of these two modules is critical for effectively modeling epileptic activity across multiple temporal scales and compensating for the Transformer's limited inductive bias toward locality and temporal hierarchy. **Second, how to bridge the gap between short-term and long-term forecasting, which are often handled separately in prior work?** We tackle this by introducing a hierar-

chical Transformer architecture that jointly models local and global temporal dependencies, enabling unified forecasting across different time horizons. **Third, how to reduce model dependence on raw, highly individualized EEG signals?** Instead of using raw RNS data, we leveraged the RNS system's continuous monitoring of EEG-based hyperexcitable patterns (A/B spike detections and long episodes) to reduce the reliance on raw signal processing. Together, these components form SeizureFormer, a robust solution for generalized, multi-timescale seizure risk forecasting.

Our model demonstrates leading performance across multiple evaluation settings. Extensive experiments show that SeizureFormer achieves the highest **ROC AUC (79.44%)** and **PR AUC (76.29%)**, significantly outperforming baseline statistical and recurrent models. Furthermore, our approach extends seizure forecasting beyond conventional short-term windows, enabling **multi-day (1-14 days) risk estimation**, which is critical for clinical decision-making and patient management. Our contributions can be summarized as follows:

(1) We develop the first Transformer-based model for seizure risk forecasting, SeizureFormer, capable of capturing both short-term fluctuations and long-horizon seizure cycles. This enables accurate seizure forecasting across different time scales up to 14 days ahead, addressing the limitations of prior approaches.
(2) Our model relies on A+B spike patterns and Long Episodes (LE) as seizure risk proxies, reducing dependence on large volumes of raw EEG data while maintaining predictive robustness and generalizability.
(3) Extensive experiments demonstrate that SeizureFormer outperforms statistical and deep learning baselines in both short-term and long-term seizure forecasting, achieving superior predictive accuracy and clinical applicability.

2. Related Works

EEG-based Seizure Risk Forecasting: Traditional seizure forecasting models rely on scalp EEG data, using statistical and deep learning methods to estimate seizure risk. Statistical approaches such as Poisson regression, GLMs, and logistic regression model seizure probability through probabilistic frameworks.[13] Proix's study[8] showed that multi-day seizure cycles could be captured using Poisson regression on long-term iEEG, achieving above-chance forecasting in approximately 66% of patients. While effective for modeling cyclic trends, these methods struggle with the fine-grained temporal resolution required for real-time forecasting. To address this, deep learning models like RNNs, GRUs, and LSTMs have been applied to capture sequential dependencies in EEG signals.[14,15] Though they improve short-term prediction, they face limitations in modeling long-range patterns and generalizing across patients,[16] partly due to the variability of raw scalp EEG.[6,7] Recent studies have attempted to incorporate multi-day periodicity and additional physiological biomarkers such as heart rate (HR) and skin conductance (SC).[16] However, few approaches integrate both short-term fluctuations and long-term seizure cycles, limiting their clinical applicability.

RNS-Based Seizure Risk Forecasting: The RNS system, compared to raw scalp EEG, provides more stable and temporally correlated biomarkers such as interictal epileptiform ac-

tivity (IEA) and long episodes (LE),[9,10] making it a reliable source for seizure-related forecasting. Several studies have explored this potential: a past study[17] demonstrated the feasibility of using statistical and classical machine learning methods, including SVM and random forests, to forecast seizure frequency from long-term RNS recordings, capturing meaningful patient-specific temporal patterns with AUC ranging from 70% to 89%. Deep learning approaches have also been applied; for example, Constantino[18] and Peterson[11] utilized CNN-based models for seizure detection and onset prediction, achieving high accuracy on short-term tasks. While these studies show strong potential for leveraging structured RNS data, they still face limitations: most models heavily rely on patient-specific features and require extensive per-patient training due to the raw, individualized nature of RNS signals. Moreover, few models have explored unified architectures capable of simultaneously capturing short-term variability and long-term seizure periodicity from RNS-derived features.

Long-Term Seizure Risk Forecasting: Recent work has extended seizure prediction from short- to long-term horizons using RNS data. For example, Yang's study[19] showed that Poisson regression with SVMs achieves AUCs above 70% for up to 6-day forecasts, highlighting the potential of cyclic RNS features. However, most models target short or long-term patterns in isolation, and few explore learning both simultaneously. To address this, we propose a Transformer-based model that captures short-term fluctuations and long-range dependencies via multi-head self-attention. Compared to GRUs and LSTMs, Transformers mitigate vanishing gradients and better capture long-range periodicity through global attention mechanisms. We also benchmark statistical, classical ML, and deep learning baselines, showing that our approach achieves superior performance in modeling seizure dynamics across time scales.

3. Methods

In this section, we first present a general problem formulation for seizure risk forecasting. We then describe how RNS-derived features are utilized to support the forecasting process.

3.1. *Problem Formulation and Feature Selection*

Problem Formulation: We formulate seizure risk forecasting as a binary risk estimation task, where the objective is to predict whether a given day falls into a high-risk seizure period based on historical biomarkers of brain excitability. Given an input time series $\mathbf{X} = [\mathbf{x}_{t-n}, \mathbf{x}_{t-n+1}, ..., \mathbf{x}_t] \in \mathbb{R}^{n \times d}$, where $\mathbf{x}_t$ represents the extracted feature set at time t, d is the feature dimension (i.e., the number of input variables), and n is the sequence length, defining the number of past time steps. The goal is to estimate the seizure risk for the subsequent h days. In this study, we evaluate four forecasting horizons: $h \in 1, 3, 7, 14$.

$$y_h = f(\mathbf{x}_{t-n}, \mathbf{x}_{t-n+1}, ..., \mathbf{x}_t), \tag{1}$$

where $y_h \in \{0, 1\}$ is a binary label determined by the presence of long episodes.

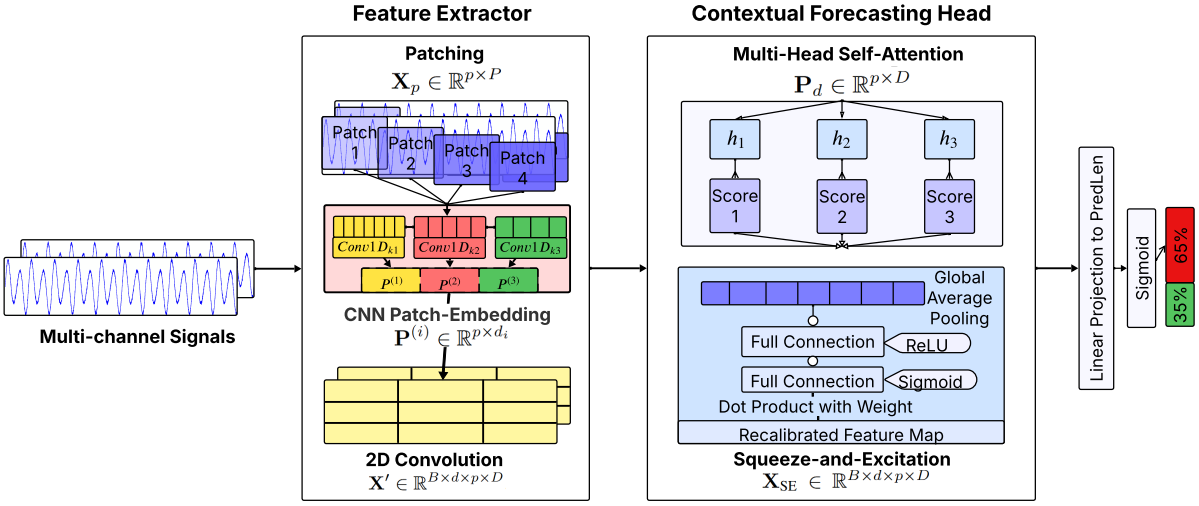
RNS-Based Feature and Label Selection: Due to the non-stationary nature of raw scalp EEG, we chose to use data collected by the RNS system. While raw RNS data can still be

varying among different patients and across time, some clinically relevant biomarkers can be extracted from raw RNS data to represent seizure proxy in a more stable way. To capture these clinically relevant dynamics, we extract two structured biomarkers from the RNS system:

- **A+B Patterns (IEA Surrogate Biomarkers):** The RNS system continuously monitors electrocorticographic activity and detects patient-specific epileptiform discharges. These patterns provide a structured time-series representation of cortical excitability, making them a more stable biomarker for seizure forecasting compared to raw EEG. For each device with two channels, we extract Pattern A and B from both channels to get two combined features: *Pattern A + B on Channel 1* and *Pattern A + B on Channel 2*. These serve as the input features $\mathbf{x}_t \in \mathbb{R}^2$, and the full input sequence $\mathbf{X} = [\mathbf{x}_{t-n}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{n\times 2}$ is used to predict the seizure risk label y_h.
- **Long Episodes (LEs) as Seizure Risk Indicators:** The RNS system also records Long Episodes (LEs), defined as sustained abnormal electrocorticographic activity exceeding a predefined duration threshold. Prior studies have shown that LE occurrences correlate strongly with seizure likelihood, making them an effective proxy for defining high-risk periods.[9] In our formulation, the binary label $y_h \in \{0, 1\}$ is determined based on LE activity: we assign $y_h = 1$ if the cumulative LE count or duration in the forecasting window exceeds a patient-specific threshold, indicating a high-risk period; otherwise, $y_h = 0$.

3.2. *Our Proposed Model: SeizureFormer*

Fig. 1. Overall framework of SeizureFormer.



As shown in Fig. 1, *SeizureFormer* is composed of two functional components that first extract multi-scale temporal features and then model contextual risk dependencies for long-term seizure forecasting. Both of the modules will be explained below:

Multi-Scale Temporal Feature Extractor: The first stage of our model is the **Multi-Scale Temporal Feature Extractor**, which decomposes multichannel A+B signals into temporally and spatially enriched representations suitable for downstream modeling.

We begin by applying **patching** to each input channel independently. To effectively capture local seizure-relevant patterns, each univariate time series $\mathbf{X} \in \mathbb{R}^{N\times 1}$ is segmented into overlapping temporal windows, or patches, of length P and stride S:

$$\mathbf{X}_p = \{\mathbf{X}_{t:t+P} \mid t = 1, S, 2S, ..., N - P\}, \tag{2}$$

$$\mathbf{X}_p \in \mathbb{R}^{p\times P}, \tag{3}$$

where P is the *patch length*, S is the *stride*, and $p = \lfloor (N - P)/S \rfloor + 1$ is the number of resulting patches. This process reduces redundancy and enables the model to encode fine-grained temporal structure.

After patching, we perform **CNN patch embedding** to extract temporal features from each patch at multiple resolutions. This design mimics clinical intuition: seizures can manifest as both brief transients and prolonged patterns. Multi-kernel CNNs enable the model to concurrently detect different temporal signatures. The patch sequence $\mathbf{X}_p$ is processed by K parallel 1D convolutional layers with kernel sizes $\{k_1, \ldots, k_K\}$, each producing a feature map:

$$\mathbf{P}^{(i)} = \mathrm{Conv1D}_{k_i}(\mathbf{X}_p), \tag{4}$$

$$\mathbf{P}^{(i)} \in \mathbb{R}^{p\times d_i}, \tag{5}$$

$$d_1 = \cdots = d_K, \quad i = 1, \ldots, K, \tag{6}$$

where $\mathrm{Conv1D}_{k_i}(\cdot)$ denotes a convolution with kernel size k_i. The outputs are concatenated to form the final patch embedding:

$$\mathbf{P} = \mathrm{Concat}\left(\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(K)}\right), \quad \mathbf{P} \in \mathbb{R}^{p\times d'}, \tag{7}$$

where $d' = \sum_{i=1}^{K} d_i$. To preserve the ordering of patches, we add a learnable position encoding:

$$\mathbf{P}_d = \mathbf{W}_P\mathbf{P} + \mathbf{W}_{\mathrm{pos}}, \quad \mathbf{P}_d \in \mathbb{R}^{p\times D}, \tag{8}$$

where $\mathbf{W}_P$ is a trainable projection matrix and $\mathbf{W}_{\mathrm{pos}}$ is a learnable position encoding.

Once patch-wise features have been extracted and embedded, we enhance the representation further using the **Cross-Variable Temporal 2D Conv** module. While the previous steps capture intra-channel patterns, this component explicitly models interactions across different A+B channelscapturing dependencies across spatially separated brain regions. This allows the model to learn coordinated dynamics that may underlie seizure generation. We first reshape the patch embeddings into a 4D tensor $\mathbf{X} \in \mathbb{R}^{B\times d\times p\times D}$, where B is the batch size, d is the number of channels, p is the number of patches, and D is the embedding dimension. Then, a 2D convolution is applied over the (d, p) grid:

$$\mathbf{X}' = \mathrm{Conv2D}_{\mathcal{K}}(\mathbf{X}), \quad \mathbf{X}' \in \mathbb{R}^{B\times d\times p\times D}. \tag{9}$$

These enhancements introduce inductive biases that help the model learn local and inter-channel patterns.

Contextual Risk Modeling Module: The second component of SeizureFormer is the **Contextual Risk Modeling Module**, which integrates local features extracted by the previous stage and models global temporal dependencies to predict seizure risk.

Starting from the output of the previous convolutional step, we have a 4D tensor $\mathbf{X}' \in \mathbb{R}^{B\times d\times p\times D}$, where B is the batch size, d the number of channels, p the number of patches, and D the embedding dimension. To capture long-range dependencies across time for each channel, we apply **Multi-Head Self-Attention (MHSA)** to the patch dimension. For each attention head $j \in \{1, \ldots, h\}$, the output is computed as:

$$\begin{aligned} \mathrm{SA}_j(\mathbf{P}_d) &= \mathrm{softmax}\left(\frac{(\mathbf{P}_d\mathbf{W}_j^Q)(\mathbf{P}_d\mathbf{W}_j^K)^\top}{\sqrt{d_k}}\right)(\mathbf{P}_d\mathbf{W}_j^V), \\ MultiHead(\mathbf{P}_d) &= \left(\|_{j=1}^{h}\mathrm{SA}_j(\mathbf{P}_d)\right)\cdot\mathbf{W}^O, \end{aligned} \tag{10}$$

where $\mathbf{W}_j^Q, \mathbf{W}_j^K, \mathbf{W}_j^V$ are projection matrices for the queries, keys, and values respectively, $\mathbf{W}^O$ is the output projection, and $d_k = D/h$ is the per-head dimension. This mechanism allows the model to learn contextualized representations of each patch in the sequence.

Following attention, we further enhance the representation using a **Squeeze-and-Excitation Feature Recalibration** block. This lightweight module adaptively emphasizes the most seizure-relevant channels by learning to reweight their contributions based on global context. It first summarizing the information from each channel via global average pooling:

$$g_{b,d} = \frac{1}{p\cdot D}\sum_{a=1}^{p}\sum_{c=1}^{D}\mathbf{X}_{b,d,a,c}, \tag{11}$$

and then passing this summary through a two-layer feedforward network:

$$\mathbf{s} = \sigma\left(\mathbf{W}_2\cdot\mathrm{ReLU}(\mathbf{W}_1\mathbf{g})\right), \quad \mathbf{X}_{\mathrm{SE}} = \mathbf{X}'\cdot\mathbf{s}.\mathrm{unsqueeze(-1)}.\mathrm{unsqueeze(-1)}, \tag{12}$$

where $\mathbf{W}_1, \mathbf{W}_2$ are learnable parameters and σ is the sigmoid activation. This recalibration allows the model to focus on the most informative spatial patterns while suppressing less relevant activity, improving robustness to noise and channel variability.

The final step is **Seizure Risk Prediction**. We flatten the recalibrated tensor $\mathbf{X}_{\mathrm{SE}}$ to obtain a global representation $\mathbf{H}$, which is passed through a fully connected layer with dropout and sigmoid activation:

$$\hat{Y} = \sigma(\mathbf{W}\mathbf{H} + \mathbf{b}). \tag{13}$$

This produces a scalar probability representing the likelihood of a high-risk seizure period for the given input.

The model is trained using **binary cross-entropy loss**, which is well-suited for imbalanced binary classification tasks:

$$\mathcal{L} = -\frac{1}{M}\sum_{i=1}^{M}\left(y_i\log\hat{Y}_i + (1-y_i)\log(1-\hat{Y}_i)\right), \tag{14}$$

where y_i is the ground truth label, $\hat{Y}_i$ is the predicted probability, and M is the batch size.

4. Experimental Settings

4.1. *Datasets*

Our study uses electrocorticographic data from patients with implanted RNS devices. Rather than raw spectrograms, we extract IEA surrogate biomarkers as features and LEs as seizure risk indicators. The dataset comprises recordings from five patients, each identified by a unique RNS Patient ID, and collected at the Emory Epilepsy Center. The daily recordings span from 3,030 to 6,953 days per patient. These patients were selected based on the availability of ultra-long-term monitoring data and clinically validated seizure annotations. The dataset is partitioned into 70% training, 10% validation, and 20% testing sets.

4.2. *Data Preprocessing and Labeling*

To account for inter-patient variability and device reprogramming effects, we normalize the A+B pattern counts using patient-specific Z-scores:

$$Z_t = \frac{X_t - \mu}{\sigma}, \tag{15}$$

where X_t is the A+B count at time t, and μ, σ are the mean and standard deviation across all visits. High-risk days ($Y_t = 1$) are defined as days where the LE count exceeds 70% of the mean over the past 60 days, a threshold that approximates when patients typically transition into higher seizure vulnerability states. This adaptive strategy, informed by clinician input, reflects evolving seizure patterns and maintains robustness across clinical stages.

4.3. *Data Visulization*

To illustrate both macro-level trends and finer variability, we visualized 1000-day segments of normalized A+B patterns with corresponding LE risk labels in Fig.2.

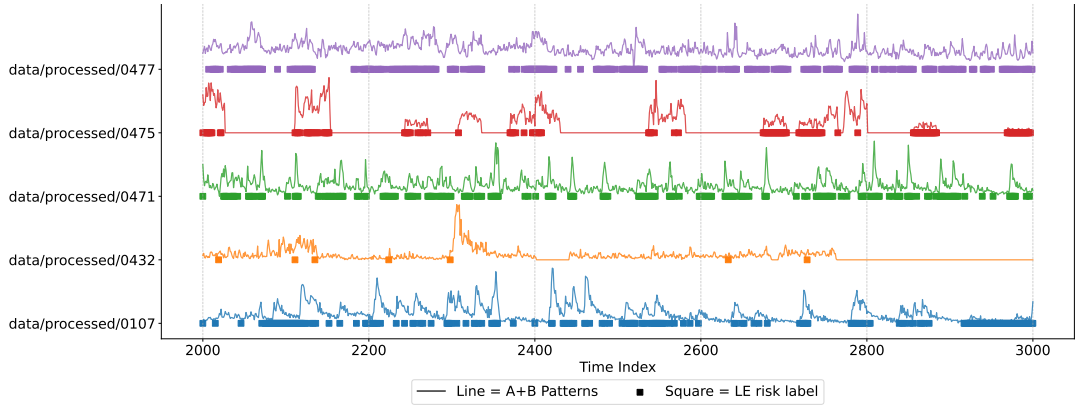


Fig. 2. A+B patterns and LE risk labels over a 1000-day period. The line represents normalized A+B patterns, while the square markers indicate high risk labels.

These samples were selected for their size and to represent diverse seizure risk profiles across patients. Fig. 2 shows that clear differences emerge across patients: *Patient 0475* and *Patient 0432* show periodic high-risk patterns—short cycles for 0475 and long, irregular ones

for 0432—affecting predictability. *Patient 0471* exhibits short, frequent cycles with high A+B variability, indicating unstable risk dynamics. In contrast, *Patients 0477* and *0107* show prolonged high-risk periods and noisy A+B patterns, suggesting persistent ictal-interictal states and increased forecasting difficulty.

4.4. *Evaluation Metrics*

According to the label distribution shown in Fig. 2, the label distribution for all patients is imbalanced. Thus we use Area Under the Receiver Operating Characteristic curve (ROC AUC) and Area Under the Precision-Recall curve (PR AUC) as the primary metrics.[20]

4.5. *Implementation Details*

Models undergo extensive hyperparameter tuning. We set the learning rate to 0.003, hidden dimension (D) to 128, batch size to 2048, and weight decay to 0.0001. The model uses 3 input channels ($enc\_in = 3$), 2 decoder input channels ($dec\_in = 2$), and 3 output channels ($c\_out = 3$). The Transformer encoder comprises 3 layers with 2 attention heads and feedforward dimension 1024. Dropout is set to 0.2 to prevent overfitting. To address class imbalance, we apply class weighting in the loss function using a *pos_weight* derived from the ratio of negative to positive samples per patient. Experiments are conducted on an NVIDIA TITAN RTX (24GB VRAM) with CUDA 12.4. Models are trained for up to 30 epochs with early stopping (patience = 5), based on validation AUC.

4.6. *Baseline Models*

We compare SeizureFormer with baselines across three categories: statistical, machine learning, and deep learning. Statistical models, including Generalized Linear Models (GLM) and Poisson Regression,[21,22] provide interpretable baselines suited for structured medical data. For classical machine learning, we test Support Vector Machines (SVM)[23] and Logistic Regression,[22] which are effective for binary classification using structured features. Deep learning models include GRU-LSTM[24] for capturing long-range dependencies, DLinear[25] for decomposing trend and seasonality in time series, Informer,[26] which introduces a ProbSparse self-attention mechanism for efficient long-sequence forecasting, and PatchTST,[27] a Transformer-based model that leverages patching and global attention for improved forecasting.

5. Experimental Results

In this section, we present our experimental results, structured around four key research questions (RQs) to systematically evaluate the performance of our proposed model, SeizureFormer, in seizure risk forecasting.

(1) **RQ1:** Does SeizureFormer outperform baseline models?
(2) **RQ2:** How does the prediction window ($PredLen = 1, 3, 7, 14$ days) affect performance?
(3) **RQ3:** How do different patients' ictal patterns influence model performance?
(4) **RQ4:** How do different components affect the performance of SeizureFormer?

Table 1. Seizure Risk Forecasting Results. The table presents each model's seizure risk forecasting performance under different settings. Mean ROC AUC and Mean PR AUC are also calculated to show the overall performance of each model. The best performance is bold and the second best is underlined.

Patients		0107		0432		0471		0475		0477		Average	
Models	Pred	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR
GLM	1	59.5%	37.6%	77.8%	19.6%	68.9%	55.9%	95.9%	88.7%	81.2%	67.1%		
	3	55.2%	51.5%	69.8%	21.8%	67.6%	76.3%	92.5%	86.8%	76.8%	75.8%	**71.32%**	**64.34%**
	7	58.7%	70.2%	53.4%	25.9%	67.5%	87.3%	90.2%	87.0%	80.4%	88.4%		
	14	58.2%	81.7%	21.5%	69.4%	95.3%	86.4%	84.1%	NA	NA	NA		
Poisson Regression	1	62.7%	48.8%	78.0%	27.1%	67.7%	54.7%	96.2%	90.8%	80.8%	65.9%		
	3	54.6%	51.3%	78.8%	33.5%	63.0%	73.6%	93.6%	88.5%	71.5%	69.4%	71.20%	65.24%
	7	59.8%	70.6%	72.1%	31.2%	59.6%	85.7%	90.5%	87.6%	76.9%	87.2%		
	14	65.0%	82.8%	38.2%	17.2%	61.6%	93.7%	82.2%	79.9%	NA	NA		
Logistic Regression	1	62.4%	44.9%	82.3%	22.3%	68.4%	54.8%	94.8%	87.1%	79.3%	62.6%		
	3	53.9%	49.1%	77.9%	28.2%	65.3%	73.8%	93.2%	87.6%	76.6%	74.6%	73.13%	64.94%
	7	55.3%	66.7%	70.9%	31.7%	68.7%	88.1%	90.4%	87.4%	81.3%	89.2%		
	14	59.1%	82.2%	54.3%	25.3%	69.8%	95.3%	85.6%	82.9%	NA	NA		
SVM	1	60.4%	39.4%	64.4%	16.4%	68.8%	55.2%	95.8%	88.5%	79.3%	63.0%		
	3	55.5%	51.3%	72.9%	22.9%	68.1%	76.1%	93.3%	87.9%	75.5%	71.8%	71.35%	63.92%
	7	57.3%	68.4%	62.1%	27.0%	67.5%	88.0%	90.6%	87.4%	81.3%	89.2%		
	14	57.7%	81.4%	50.8%	22.9%	69.0%	95.2%	85.4%	82.4%	NA	NA		
DLinear	1	70.9%	57.7%	46.2%	8.96%	68.9%	55.9%	96.9%	92.0%	81.8%	67.6%		
	3	52.9%	52.9%	73.4%	22.7%	69.4%	78.0%	93.6%	88.9%	79.3%	78.6%	72.83%	65.97%
	7	47.1%	61.6%	84.2%	37.2%	69.1%	89.5%	88.2%	85.1%	79.0%	89.6%		
	14	52.5%	77.5%	80.6%	35.6%	69.5%	95.2%	80.3%	78.9%	NA	NA		
GRU_LSTM	1	58.2%	81.7%	82.5%	32.4%	68.5%	56.3%	**97.6%**	93.3%	83.9%	68.4%		
	3	67.9%	66.2%	87.8%	54.3%	70.8%	78.7%	94.7%	88.9%	76.7%	73.5%	73.16%	71.05%
	7	64.5%	75.1%	76.8%	30.8%	58.4%	88.1%	90.2%	86.2%	50.0%	85.2%		
	14	45.6%	77.6%	35.1%	51.9%	51.9%	85.7%	82.6%	NA	NA	NA		
Informer	1	54.2%	64.1%	53.0%	54.7%	54.0%	67.7%	90.2%	89.3%	77.8%	66.7%		
	3	61.3%	66.9%	83.7%	28.8%	61.4%	68.8%	88.7%	82.7%	66.8%	66.8%	65.65%	69.82%
	7	67.7%	80.1%	54.2%	29.8%	58.7%	82.7%	85.4%	86.1%	56.6%	79.9%		
	14	65.1%	**88.0%**	46.2%	38.5%	42.3%	90.1%	84.9%	84.8%	NA	NA		
PatchTST	1	**74.2%**	63.9%	93.2%	62.4%	66.4%	46.0%	97.1%	91.7%	78.2%	62.7%		
	3	52.0%	45.4%	92.6%	59.6%	69.6%	76.4%	94.5%	89.9%	77.4%	78.7%	76.41%	73.07%
	7	53.7%	70.8%	87.4%	49.5%	73.1%	91.0%	91.2%	88.0%	78.8%	89.2%		
	14	28.9%	64.3%	83.9%	**78.5%**	**74.6%**	**95.9%**	85.0%	84.5%	NA	NA		
SeizureFormer	1	**74.2%**	53.5%	**94.7%**	61.6%	70.3%	59.2%	97.1%	**93.4%**	**84.8%**	73.0%		
	3	61.4%	68.1%	92.2%	77.8%	71.0%	76.7%	94.8%	90.8%	80.6%	80.7%	**79.44%**	**76.29%**
	7	47.7%	64.6%	87.9%	52.3%	71.8%	90.6%	92.7%	88.3%	81.6%	**90.7%**		
	14	64.7%	83.1%	85.5%	64.6%	73.3%	95.7%	86.0%	84.9%	NA	NA		

5.1. *RQ1: Overall Model Performance*

To evaluate whether SeizureFormer outperforms baseline models, we first analyze the overall model performance, disregarding the effects of prediction length (*PredLen*) and patient-specific variations. Table 1 presents the ROC AUC and PR AUC scores across all patients and *PredLen* values. SeizureFormer achieves the highest average ROC AUC (*79.44%*) and PR AUC (*76.29%*), surpassing all baseline models. The second-best performing model, PatchTST, achieves *76.41%* ROC AUC and *73.07%* PR AUC, while other deep learning models, such as GRU-LSTM, also demonstrate strong performance but fall short of SeizureFormer. Traditional statistical and classical machine learning models, including Logistic Regression, Poisson Regression, GLM, and SVM, yield lower mean PR AUC scores, reflecting their limited capacity to model complex temporal dynamics and address severe class imbalance. These results establish SeizureFormer as the most effective model overall, though further analysis is needed to

assess its robustness across different prediction lengths and patient cases. Examining patient-specific performance, SeizureFormer achieves top performance for four out of five patients, demonstrating strong generalization across different seizure patterns. For Patient *0477*, where long-term forecasting becomes challenging due to sparse or inconsistent label availability at longer prediction windows *PredLen = 14*, SeizureFormer still outperforms other models in shorter forecasts. GRU-LSTM exhibits strong performance in some cases but struggles with consistency, particularly for longer *PredLen*. These results confirm SeizureFormer as the most effective model for seizure risk forecasting.

5.2. *RQ2: Impact of Prediction Window*

To analyze the impact of prediction window length on model performance, we evaluate trends in both ROC AUC and PR AUC scores across different *PredLen* values. Table 1 provides the full performance breakdown. A general pattern emerges where shorter *PredLen* values (*1, 3* days) yield higher ROC AUC scores, while longer *PredLen* values (*7, 14* days) lead to higher PR AUC scores. For *PredLen = 1, 3* days, models show higher ROC AUC, with deep learning models (SeizureFormer, PatchTST, GRU-LSTM) performing consistently well. Traditional models (e.g., Poisson Regression, SVM) show weaker and less stable performance. For *PredLen = 7, 14* days, PR AUC improves, reflecting better long-term risk trend capture. Deep learning models handle class imbalance more effectively, with SeizureFormer exceeding *80%* PR AUC and GRU-LSTM often above *70%*. Statistical models frequently fall below *50%*. The trade-off between ROC AUC and PR AUC highlights the balance between short-term precision and long-term generalization. SeizureFormer performs robustly across all *PredLen* values. However, while most models benefit from longer prediction horizons, we observe a consistent PR AUC decline for **Patient 0475** (e.g., SeizureFormer: *93.4%→84.9%*, PatchTST: *78.2%→80.9%*, Poisson Regression: *90.8%→79.9%*, SVM: *88.5%→82.4%*). As shown in Fig.2, this patient's highly periodic seizure pattern likely causes temporal misalignment in long-window predictions, leading to false positives during low-risk phases and reduced PR AUC.

5.3. *RQ3: Inter-Patient Variation*

To investigate the impact of different patients' ictal patterns on model performance, we analyze how seizure forecasting varies across individuals. Given the variability in seizure occurrence and EEG characteristics, it is essential to assess whether certain models maintain consistent performance across different patients. Table 1 shows that forecasting performance varies widely across patients. Some (e.g., *0475*) consistently achieve high scores, while others (e.g., *0432*) show large fluctuations, indicating varying signal clarity and model sensitivity. SeizureFormer performs most consistently, while GRU-LSTM shows more variation. Statistical models often fail on patients with complex patterns. Model performance is strongly patient-dependent, with SeizureFormer generalizing better than other baselines.

5.4. *RQ4: Ablation Study of Model Components*

To assess each component's contribution, we performed ablation studies on **Patient 0477**, a representative case with stable high-risk dynamics. As shown in Table 2, we individually

Table 2. Ablation experiment results on Patient 0477. The best performance is bold and the second best is underlined.

Model Variant	**PredLen = 1**		**PredLen = 3**		**PredLen = 7**	
Metrics	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
Full Model	**84.83%**	**73.03%**	**80.62%**	**80.70%**	<u>81.63%</u>	**90.70%**
w/o CNN Patch Embedding	<u>84.39%</u>	<u>71.89%</u>	73.31%	75.60%	64.61%	81.59%
w/o SE Block	82.80%	68.68%	<u>77.95%</u>	<u>78.21%</u>	**82.76%**	<u>90.20%</u>
w/o Cross-Variable Temporal convolution	77.71%	63.55%	77.66%	77.24%	75.69%	88.65%
w/o All Modules	76.35%	61.86%	70.87%	72.40%	72.97%	87.56%

removed three core modules—CNN Patch Embedding, Squeeze-and-Excitation (SE) Block, and Cross-Variable Temporal Convolution (CVT)—as well as all together. The table shows that removing any module reduces overall performance across multiple prediction lengths. Detailed comparisons reveal the following patterns:

- **CNN Patch Embedding** plays a critical role in long-horizon prediction. At *PredLen = 7*, removing it leads to the largest performance degradation: ROC AUC drops by 17.02% and PR AUC by 9.11%, confirming its effectiveness in capturing long-term features.
- **CVT (Cross-Variable Temporal Convolution)** is crucial for short-term forecasting. At *PredLen = 1*, its removal causes a 7.12% drop in ROC AUC and a 9.48% drop in PR AUC, indicating its importance in modeling high-resolution temporal dependencies.
- **SE Block** subtly improves short-horizon performance. At *PredLen = 1*, its removal results in a 2.03% drop in ROC AUC and a 4.35% drop in PR AUC. Interestingly, at *PredLen = 7*, its removal causes a slight increase in ROC AUC (+1.13%), likely due to reduced overfitting or over-suppression of weak channels. However, PR AUC still drops by 0.50%, which is more clinically relevant in imbalanced settings. This suggests that SE Block may help improve early seizure precision while trading off certain long-horizon sensitivity.

These findings highlight the complementary strengths of the three modules in addressing both short-term and long-term forecasting demands. Their combination enables *SeizureFormer* to achieve consistent performance across different horizons.

6. Conclusion

We present **SeizureFormer**, a Transformer-based model that advances seizure risk forecasting using RNS-derived IEA biomarkers. By integrating multi-kernel CNNs, cross-variable convolution, self-attention, and a squeeze-and-excitation module, it captures seizure-relevant temporal dynamics across multiple time scales. Experiments demonstrate state-of-the-art performance and strong generalizability across patients and forecasting horizons.

In clinical settings, SeizureFormer could enable proactive seizure management. For instance, if a forecast indicates elevated seizure risk over the next three days, clinicians might preemptively adjust medication, modify behavioral protocols, or recommend precautions. Such foresight empowers both patients and providers to make timely, informed decisions.

Future work will explore transfer learning across patients, dynamic biomarker adaptation, and integration with closed-loop neurostimulation for end-to-end seizure risk mitigation.

Acknowledgments

This research was partially supported by the U.S. National Science Foundation under Award No. 2437345. We would like to thank our collaborators and colleagues for their valuable feedback and discussions that helped improve this work.

References

1. World Health Organization, Epilepsy: A public health imperative (2023).
2. W. T. Kerr, K. N. McFarlane and G. F. Pucci, The present and future of seizure detection, prediction, and forecasting with machine learning, including the future impact on clinical trials, *Frontiers in Neurology* **15**, p. 1425490 (2024).
3. L. Kuhlmann, K. Lehnertz, M. P. Richardson *et al.*, Seizure prediction—ready for a new era, *Nature Reviews Neurology* **17**, 618 (2021).
4. L. Kuhlmann, K. Lehnertz, M. P. Richardson, B. Schelter and H. P. Zaveri, Eeg datasets for seizure detection and prediction—a review, *Seizure* **65**, 3 (2019).
5. X. Zhang, X. Zhang, Q. Huang and F. Chen, A review of epilepsy detection and prediction methods based on eeg signal processing and deep learning, *Frontiers in Neuroscience* **18**, p. 1468967 (2024).
6. K. Rasheed, A. Qayyum, J. Qadir, S. Sivathamboo, P. Kwan, L. Kuhlmann, T. O'Brien and A. Razi, Machine learning for predicting epileptic seizures using eeg signals: A review, *IEEE Reviews in Biomedical Engineering* **14**, 139 (2021).
7. L. Yang, R. E. Stirling, M. Kerr, J. J. Howbert and S. N. Baldassano, Seizure forecasting using a long-term implanted neurostimulator: data non-stationarity and model retraining, *arXiv preprint* (2022).
8. T. Proix, M. O. W. Thiele, V. R. Rao, T. K. Tcheng, W. Stacey, M. Guye, D. M. Kühn, M. L. V. Quyen, M. J. Stigler, O. Blanke, V. K. Jirsa and M. O. Baud, Forecasting seizure risk in adults with focal epilepsy: a development and validation study, *The Lancet Neurology* **20**, 127 (2021).
9. A. N. Khambhati, E. F. Chang, M. O. Baud and V. R. Rao, Hippocampal network activity forecasts epileptic seizures, *Nature Medicine* **30**, 2787 (OCT 2024).
10. H. Yang, J. Müller, M. J. Cook, P. Jiruska and A. Schulze-Bonhage, Seizure forecasting with ultra long-term eeg signals, *Clinical Neurophysiology* **146**, 56 (2024).
11. V. Peterson, D. A. Bickel, R. E. Gross, D. L. Kreil, M. J. Cook, G. Krauth and G. Klambauer, Deep net detection and onset prediction of electrographic seizure patterns in responsive neurostimulation, *Epilepsia* **64**, e1 (2023).
12. C. Schroeder, A. Nowacki, S. Chiang, M. Zamora and G. Worrell, Tracking seizure cycles in chronic intracranial eeg using wearable and implanted devices, *arXiv preprint* (2022).
13. M. G. Leguia *et al.*, Learning to generalize seizure forecasts: Comparison of generalized linear models and deep learning approaches, *Epilepsia* (2022).
14. P. Thodoroff, J. Pineau and A. Lim, Learning robust features using deep learning for automatic seizure detection, *International Conference on Machine Learning (ICML)* (2016).
15. S. Raghu, N. Sriraam, W. M. Temesgen and S. Rao, Deep learning for epilepsy: A review of recent advances and future challenges, *IEEE Transactions on Biomedical Engineering* (2019).
16. R. G. Andrzejak *et al.*, Seizure forecasting: Where do we stand?, *Epilepsia* (2023).
17. Y. Yang, S. B. Dumanis, K. A. Bujarski, G. A. Worrell, E. Geller, L. J. Hirsch, U. Seneviratne, M. J. Morrell and C. W. Bazil, Machine learning algorithms for seizure frequency forecasting using electrocorticographic data from a neurostimulator, *Epilepsy Behavior* **124**, p. 108347 (2021).
18. A. C. Constantino, N. D. Sisterson, N. Zaher, A. Urban, R. M. Richardson and V. Kokkinos,

Expert-level intracranial electroencephalogram ictal pattern detection by a deep learning neural network, *Frontiers in Neurology* **12**, p. 603868 (MAY 2021).
19. H. Yang, J. Müller and M. Eberlein, Seizure forecasting with ultra long-term eeg signals, *Clinical Neurophysiology* (2024).
20. T. Saito and M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, *PLoS ONE* **10**, p. e0118432 (2015).
21. J. A. Nelder and R. W. M. Wedderburn, Generalized linear models, *Journal of the Royal Statistical Society: Series A (General)* **135**, 370 (1972).
22. J. M. Hilbe, *Logistic Regression Models* (Chapman and Hall/CRC, 2011).
23. C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning* **20**, 273 (1995).
24. K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, *arXiv preprint* (2014).
25. A. Zeng, Z. Zhang and Y. Xu, Dlinear: Modeling long-term temporal dependency with linear model for time series forecasting, *arXiv preprint* (2023).
26. H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong and W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (12)2021. ProbSparse self-attention, self-attention distilling, generative decoder.
27. Z. Nie, J. Yoon and R. Zhang, Time-series representation learning via patch-based transformer, *Advances in Neural Information Processing Systems (NeurIPS)* (2022).