A Self-Interpretable Deep Learning Model for Seizure Prediction Using a Multi-Scale Prototypical Part Network

Yikai Gao, Aiping Liu, *Member, IEEE*, Lanlan Wang, Ruobing Qian, and Xun Chen, *Senior Member, IEEE*

***Abstract*— The epileptic seizure prediction (ESP) method aims to timely forecast the occurrence of seizures, which is crucial to improving patients' quality of life. Many deep learning-based methods have been developed to tackle this issue and achieve significant progress in recent years. However, the "black-box" nature of deep learning models makes the clinician mistrust the prediction results, severely limiting its clinical application. For this purpose, in this study, we propose a self-interpretable deep learning model for patient-specific epileptic seizure prediction: Multi-Scale Prototypical Part Network (MSPPNet). This model attempts to measure the similarity between the inputs and prototypes (learned during training) as evidence to make final predictions, which could provide a transparent reasoning process and decision basis (e.g., significant prototypes for inputs and corresponding similarity score). Furthermore, we assign different sizes to the prototypes in latent space to capture the multi-scale features of EEG signals. To the best of our knowledge, this is the first study that develops a self-interpretable deep learning model for seizure prediction, other than the existing post hoc interpretation studies. Our proposed model is evaluated on two public epileptic EEG datasets (CHB-MIT: 16 patients with a total of 85 seizures, Kaggle: 5 dogs with a total of 42 seizures), with a sensitivity of 93.8% and a false prediction rate of 0.054/h in the CHB-MIT dataset and a sensitivity of 88.6% and a false prediction rate of 0.146/h in the Kaggle dataset, achieving the current state-of-the-art performance with self-interpretable evidence.**

***Index Terms*— Deep learning, interpretability, signal processing, seizure prediction, electroencephalography.**

Manuscript received 20 October 2022; revised 25 February 2023; accepted 17 March 2023. Date of publication 23 March 2023; date of current version 29 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 32271431, Grant 61922075, and Grant 82272070; in part by the Research Project of Health Commission of Anhui Province under Grant AHWJ2022b004; in part by the Fundamental Research Funds for the Central Universities under Grant KY2100000123; and in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2019-025. *(Corresponding author: Xun Chen.)*

Yikai Gao and Aiping Liu are with the School of Information Science and Technology, University of Science and Technology of China (USTC), Hefei 230027, China.

Lanlan Wang and Ruobing Qian are with the Epilepsy Center, Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230027, China.

Xun Chen is with the Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, and the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: xunchen@ustc.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2023.3260845, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2023.3260845

I. Introduction

EPILEPSY, one of the most common neurological diseases globally, affects 50 million people worldwide. The risk of premature death among people with epilepsy is up to three times higher than in the general population [1]. With the rapid growth of modern medicine in recent years, about 70 percent of epileptic patients could be seizure-free after proper diagnosis and treatment. However, unfortunately, 30 percent of patients still suffer from refractory epilepsy (i.e., medicines cannot control the seizures) [1], [2]. Hence, the study of epileptic seizure prediction (ESP) is crucial, which can significantly improve the quality of life of patients with epilepsy.

Electroencephalography (EEG) is a practical approach to recording electrical activity in the brain, which contributes to epilepsy diagnosis [3], [4], [5]. For the past few years, numerous studies have shown that EEG signals can be used for ESP [6], [7], [8], [9]. In general, researchers divide the longstanding EEG signals of epileptic patients into four states: preictal (the period before the seizure onset), interictal (the period between seizures), ictal (the period of seizure), and postictal (the period after seizures) [10]. Therefore, we can convert the ESP problem into a binary classification problem distinguishing preictal and interictal states. Many sophisticated approaches have been constructed following this pattern to solve this challenging problem.

Feature extraction and classification are two critical steps of conventional EEG-based ESP methods. Generally, researchers construct and select features manually based on experience or observations. Then a classifier is applied for decision making. For instance, Chisci et al. adopted the coefficient of the auto-regressive model as the feature of EEG signals. A support vector machine (SVM) is designed for subsequent classification [11]. Bedeeuzzaman et al. established a time domain feature set from EEG signals and classified these features by a linear classifier [12]. Zhang and Parhi applied spectral analysis to feature extraction of EEG signals. Then a novel feature selection approach is adopted, and an SVM classifier is used for classification [13]. Usman et al. pre-processed

EEG signals with empirical mode decomposition and then extracted time and frequency domain features to train the SVM model [14]. These studies laid a solid foundation for developing more sophisticated EEG-based ESP methods.

Recently, deep learning (DL) methods have achieved impressive performance in the EEG-based ESP field. Truong et al. used a Convolutional Neural Network (CNN) to classify the time-frequency features obtained by the Short Time Fourier Transform (STFT) on EEG signals [15]. Jana and Mukherjee presented an effective ESP approach using a CNN classifier with minimizing the channels of EEG signals [16]. Usman et al. proposed a DL-based ensemble learning approach to reduce the false positive rate of seizure prediction [17]. Rasheed et al. generated synthesized EEG samples by a deep convolutional generative adversarial network to solve the problem of scarcity of good-quality EEG data [18]. Zhao et al. developed an end-to-end adder network with supervised contrastive learning by considering the high computational complexity of deep learning methods [7]. Compared to traditional methods for ESP, DL-based methods could significantly improve the prediction performance.

However, existing deep learning models are often viewed as a "black box" due to their opaque reasoning process [19], [20]. The absence of interpretability for these models makes clinicians mistrust and question the prediction results, severely limiting its application to seizure prediction [21]. Several studies have noted this issue, caused by the "black box" nature of deep learning models. For example, Ozcan and Erturk applied an occlusion test approach to get the heat map for inputs of the CNN model, representing how important the patches from the input are in classification [22]. Dissanayake et al. employed the SHapley Additive exPlanations (SHAP) method to discuss how significant each channel contributes to the model's classification results [23]. Li et al. visualized feature maps learned from the neural network by a deconvolution scheme approach to explore the spatio-temporal-spectral dependencies for seizure prediction [24]. Jemal et al. visualized the learned filters in the first layer to interpret them as band-pass frequency filters and applied the layer-wise relevance propagation method to understand the decision process of their model [25].

Nevertheless, all these methods above focus on post hoc interpretation (i.e., the application of interpretation methods after model training) [26]. The post hoc explanation method is susceptible to signal noise interference, resulting in insufficient accuracy and reliability of the interpretation itself, which cannot deal with the "black box" problem fundamentally [27]. Unlike post hoc interpretation methods, intrinsically interpretable models have a transparent reasoning process, and the explanation obtained is more accurate and reliable. Hence, in this work, we propose an intrinsically interpretable deep learning model for epileptic seizure prediction.

Our work is inspired by prototypical part network (ProtoPNet), which was initially developed in the field of computer vision [28] and defines the prototype as the feature corresponding to a patch of a single fixed size learned from training samples. However, it is difficult to capture the multi-scale features with a patch of a single fixed-size, while EEG signals carry information among multiple time scales [29]. For example, spike wave and sharp wave are two classic waveforms related to epileptic EEG signals. The spike wave has a duration of 20 –70 ms, and the sharp wave has a duration of 70 – 200 ms [28]. Consequently, different waveforms in EEG signals may occur on different time scales. Depending on the single fixed-size of the patch, the prototypes cannot sufficiently capture the multi-scale features of EEG signals, leading to poor prediction performance. To address this issue, we proposed the Multi-Scale Prototypical Part Network (MSPPNet) for epileptic seizure prediction. Specifically, our model learns several prototypes at different scales during training and measures the similarity between the inputs and these prototypes as evidence to make final predictions. We evaluate our method on two public epileptic EEG datasets: the CHB-MIT database and the American Epilepsy Society Seizure Prediction Challenge (Kaggle) database. Despite strong constraints to make the network interpretable, MSPPNet achieves the current state-of-the-art prediction performance.

The main contributions in this work are as follows:

- To the best of our knowledge, we propose the first self-interpretable deep learning model for epileptic seizure prediction.
- Considering that EEG signals carry multi-scale information, we assign different sizes to the prototypes in latent space to improve the prediction performance and interpretability.
- We show that our model can achieve the current state-of-the-art performance with self-interpretable evidence on the CHB-MIT database and the American Epilepsy Society Seizure Prediction Challenge (Kaggle) database.

The rest paper is composed as follows. Section II describes the materials we used and our proposed methods. Section III presents the experimental results and comparison in this study. Section III-C further discusses our approach. Finally, we provide the conclusion for our work in Section IV.

II. Materials and Methods

A. Data Description

This work uses two public epileptic EEG datasets, the CHB-MIT [30], [31] and the Kaggle [32]. In the CHB-MIT dataset, long-term scalp (sEEG) EEG signals of 23 pediatric subjects with refractory epilepsy were recorded. These sEEG signals were collected at a sampling rate of 256 Hz. The electrodes were placed according to the 10-20 international system, using a bipolar montage. To guarantee the consistency of our approach, according to [28], we select 18 channels common to each patient in this study, including FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-C3, C3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, P8-O2, FZ-CZ, and CZ-PZ. In the Kaggle dataset, long-term intracranial EEG (iEEG) signals of 5 dogs and two patients were recorded. The iEEG signals from dog-1 to dog-4 were collected from 16 electrodes at 400 Hz, and 15 electrodes were used for dog-5. The iEEG signals from two patients were collected at a sampling rate of 5000 Hz, with 15 depth electrodes for patient-1 and 24 subdural electrodes for patient-2.

Before introducing our method, we define some related parameters which play a significant role in seizure prediction.

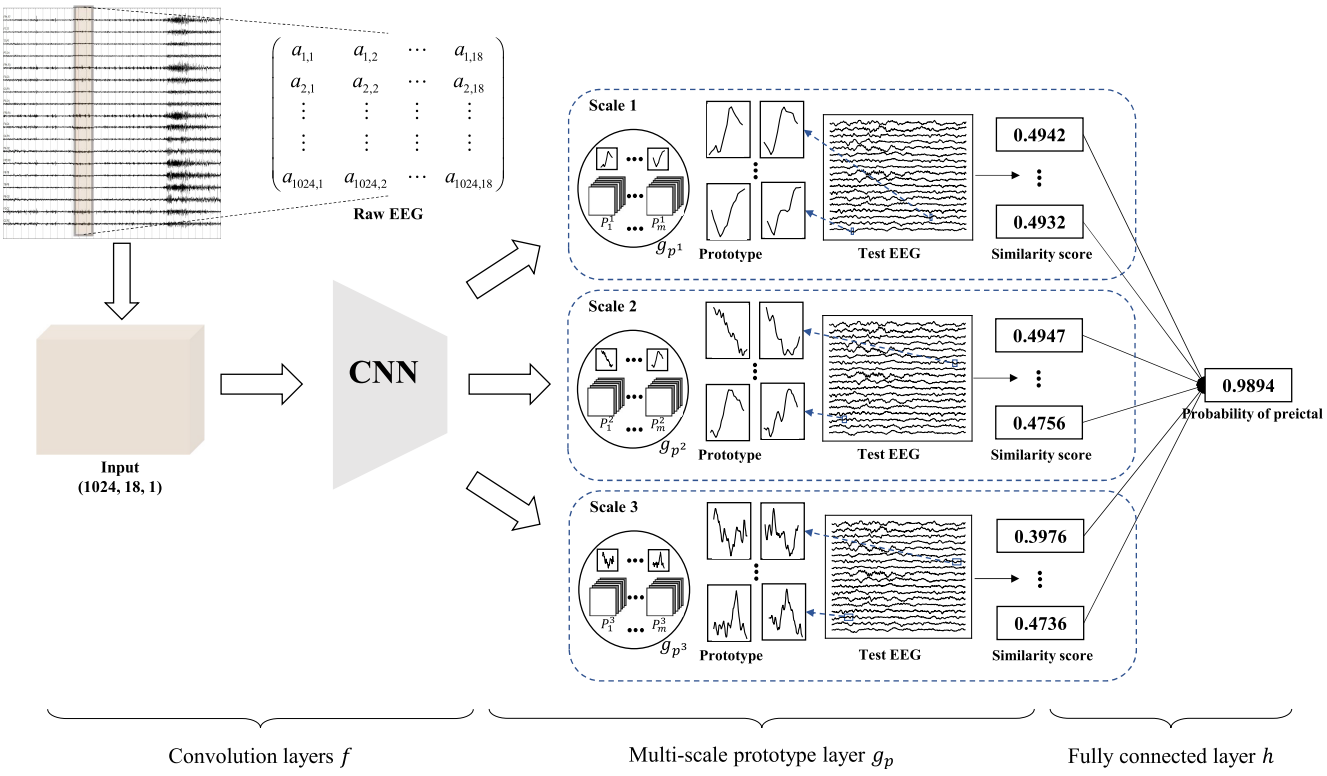

Fig. 1. The architecture of our proposed MSPPNet. Our model comprises three parts: a regular CNN module f, a multi-scale prototype layer g_p, and a fully connected layer h. The detailed structure of the CNN module is shown in Figure 2.

For the CHB-MIT dataset, we choose the value of these parameters according to [28]. Specifically, we select a preictal period of 30 minutes before the seizure, which is a commonly accepted choice for sEEG in the literature [15], [22], [28], [33], [34]. The intervention time for patients is defined as 1 minute period between the preictal and the seizure onset. The interictal period is defined as at least 1 hour before the seizure onset and at least 1 hour after the seizure. For cases with two consecutive seizures, if the interval period is less than 15 minutes, we consider them a single seizure due to a lack of preictal data. To avoid the overfitting problem, we select patients with no less than three seizures and interictal duration greater than three hours. In this situation, 16 subjects with a total of 85 seizures are used in this study. Finally, we divide the consecutive EEG signals into 4-second windows with a 2-second overlapping for subsequent classification. For the Kaggle dataset, we define these parameters following [7]. To be specific, the preictal period is defined as 1 hour before the seizure onset. This was defined by the sponsor of the Kaggle contest, so all researchers using the Kaggle data set for seizure prediction used this setting. The intervention period is defined as 5 minutes. We defined the minimum distance between the seizure and the interictal period as 4 hours. The minimum interval between seizures is also defined as 15 minutes. As for the subject selection, we excluded two human subjects due to the significant difference in sampling rate of the iEEG. Specifically, the iEEG sampling rate of dogs was 400 Hz, while that of human subjects was 5000 Hz. A higher iEEG sampling rate leads to an increase in data dimensionality, which can pose difficulties for neural networks in processing, particularly for end-to-end models [35]. Hence, five dogs with a total of 42 seizures are used for our experiments. The consecutive EEG signals are divided into 4-second windows without overlapping. The selected subject information from two datasets is presented in Table I and Table II.

B. MSPPNet

Figure 1 presents the overall architecture of our proposed model, MSPPNet. To provide intuitive interpretability, we use the raw EEG data without pre-processing as the model's input. A conventional CNN module is designed for feature extraction of EEG signals. Then we developed a multi-scale prototype layer to learn several prototypes at different scales from training samples and measure the similarity between the inputs and these prototypes as evidence to make final predictions. Intuitively, our model provides explanations in the form of *"this looks like that"* (i.e., this is a preictal sample because *this* part of the EEG signals *looks like that* prototypical part of a preictal training sample). The detailed architecture of MSPPNet is described in Section II-B.1, and the training algorithm is depicted in Section II-B.2. In Section II-B.3, we introduce the reasoning process for our model and present some examples.

1) MSPPNet Architecture: Our model comprises three parts: a regular CNN module f, a multi-scale prototype layer g_p, and a fully connected layer h.

a) The regular CNN module f: Given an input EEG sample x, the CNN module f extracts discriminative features $z = f(x)$ for the subsequent layer. We denote the shape of z as $C * H * W$, where C is the number of channels in

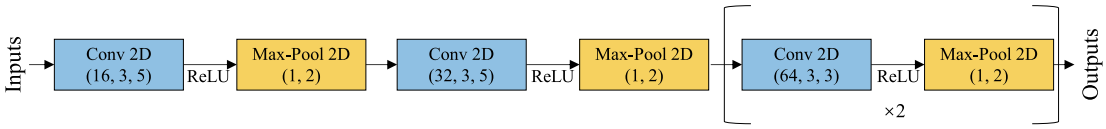

Fig. 2. The sturcture of the CNN module.

TABLE I
SUBJECT INFORMATION OF THE CHB-MIT DATASET

Subject	Gender	Age (years)	No. of seizures	Interictal time (h)
Chb-01	F	11	7	28.6
Chb-02	M	11	3	30.2
Chb-03	F	14	6	30.4
Chb-05	F	7	5	29.8
Chb-07	F	14.5	3	61.8
Chb-09	F	10	4	61.0
Chb-10	M	3	7	36.9
Chb-13	F	3	6	22.3
Chb-14	F	9	8	16.4
Chb-16	F	7	8	9.8
Chb-17	F	12	3	16.1
Chb-18	F	18	4	29.9
Chb-19	F	19	2	27.0
Chb-20	F	6	8	20.4
Chb-21	F	13	4	28.3
Chb-23	F	6	7	17.2
Total	\	\	85	466.1

TABLE II
SUBJECT INFORMATION OF THE KAGGLE DATASET

Subject	No. of seizures	Interictal time (h)
Dog-1	4	80.0
Dog-2	7	83.3
Dog-3	12	240.0
Dog-4	14	134.0
Dog-5	5	75.0
Total	42	612.3

CNN, and we set C to 64 in this study. In this case, for the CHB-MIT dataset with input size $1 * 18 * 1024$, the shape of z is $64 * 18 * 64$. For the Kaggle dataset with input size $1*16*1600$ (or $1*15*1600$), the shape of z is $64*16*100$ (or $64 * 15 * 100$). The detailed structure of the CNN is presented in Figure 2.

b) The multi-scale prototype layer g_p: In the multi-scale prototype layer, our model learns m prototype at each scale for preictal. The number of scales is set to 3 according to work in [28]. Hence, the prototypes at scale s can be indicated as $P^s = \left\{p_j^s\right\}_{j=1}^{m}$, where $s = 1, 2, 3$. The shape of each prototype p_j^s is $C * H^s * W^s$ with $H^s \leq H$ and $W^s \leq W$. It is worth mentioning that in ProtoPNet, all prototypes have the same shape [36]. However, capturing the multi-scale features with a patch of a single shape is difficult, while EEG signals carry information among multiple scales [28]. To address this issue, we assign different shapes to the prototypes in latent space so that the multi-scale features can be extracted. Expressly, we set $W^1 = 1, W^2 = 2$ and $W^3 = 4$ for three scales and set $H^1 = H^2 = H^3 = 1$. Every prototype presents some prototypical activation pattern in a patch of the convolutional output, which corresponds to some prototypical EEG segment in the original pixel space. Hence, each prototype p_j^s can be understood as the latent representation of some prototypical part of some preictal EEG samples in our model.

As an intuitive illustration, each prototype p_j^s in Figure 1 corresponds to a specific waveform in raw EEG signals at different scales. For a given output of CNN module $z = f(x)$, the j-th prototype at s-th scale $g_{p_j^s}$ in the multi-scale prototype layer calculates the squared L^2 distances d_j^s between p_j^s and all patches of z, and converts distances d_j^s to similarity scores. The outcome is a similarity map whose value denotes how strong a prototypical part appears in the EEG sample. This similarity map keeps the spatial relation of the convolutional output and can identify which portion of the input EEG sample is most like the learned prototype. Then a global max-pooling layer is applied to the similarity map to obtain a single similarity score reflecting how strongly a prototypical part appears in some segment of the input EEG sample. Mathematically, the process by which a prototype p_j^s converts $z = f(x)$ to a single similarity score can be described as follows:

$$g_{p_j^s} = \max_{z' \in \text{patches}(z)} \left(\frac{1}{1 + \exp\left(\left\| z' - p_j^s \right\|_2^2 \right)} \right)$$

According to the formula above, we can find that the function $g_{p_j^s}$ is decreasing monotonically with regard to L^2 distance $\left\| z' - p_j^s \right\|_2$ (while z' is the latent patch closet to p_j^s), and the upper bound is 0.5.

In Figure 1, the similarity score between the first prototype at the first scale p_1^1 and the most activated segment of the input preictal EEG sample is 0.4942. In the same way, the similarity score between the first prototype at the second scale p_1^2 and the most activated segment of the input is 0.4947. Since we set the number of prototypes in each scale $m = 20$, there is a total of 60 prototypes in our model, and the output shape of the multi-scale prototype layer is $60 * 1$.

c) The fully connected layer h: Finally, a fully connected layer h without adding bias is applied to the output of the multi-scale prototype layer to obtain the final prediction

results. The purpose of this layer without adding bias is to make the model have a transparent reasoning process during decision-making. In Figure 1, the 60 similar scores pass through the fully connected layer h and a sigmoid activation function, resulting in a probability of 0.9894 that the input EEG sample is preictal.

2) Training Process: The training process of our MSPPNet consists of two stages: (1) Adam optimization for all layers; (2) projection of prototypes. In this study, we set the number of training epochs to 60. When the epoch is greater than or equal to 10 and divisible by 5, we perform a projection of prototypes; In other cases, we perform Adam optimization for all layers.

a) Adam optimization for all layers: In this stage, we optimize the weights of all layers by the Adam optimizer to learn a latent space, where the most significant segments for identifying the input as preictal are clustered around semantically similar prototypes of preictal. To this end, the weights of all layers (i.e., the CNN module's weights w_{conv}, multi-scale prototype layer's weights P, and the last fully connected layer's weights w_h) are jointly optimized using the Adam optimizer. Let $D_t = [X, Y] = \{(x_i, y_i)\}_{i=1}^{n}$ denote the EEG training set. We can formula the optimization problem as follows:

$$\min_{P, w_{conv}, w_h} L_a = \frac{1}{n}\sum_{i=1}^{n} BCELoss\left(h \circ g_p \circ f\left(x_i\right), y_i\right) - \mu_1 * Clst + \mu_2 * \|w_h\|_1$$

where Clst is defined as:

$$Clst = \frac{1}{n}\sum_{i=1}^{n} \max_{j,s} \max_{z \in patches(f(x_i))} \frac{y_i}{1 + \exp\left(\left\|z - p_j^s\right\|_2^2\right)}$$

The binary cross entropy loss (BCELoss) reduces misclassification on the training set, and the maximization of the cluster cost (Clst) encourages each preictal training sample to have some latent patch that is close to at least one prototype, while the L^2 regularization on the weights of the last layer makes the contribution of prototypes to the prediction probabilities sparse and facilitates the search and visualization for significant prototypes. In this work, we set the coefficient of the cluster cost $\mu_1 = 0.8$, and the coefficient of the L^1 regularization $\mu_2 = 10^{-4}$.

b) Projection of prototypes: In our MSPPNet, each prototype is conceptually equivalent to some fragment of a preictal training sample since we project each prototype p_j^s for preictal onto the nearest latent training patch. Specifically, we will replace each prototype p_j^s with the nearest training patch from learned features z for every five epochs. Mathematically, we make the following update for the prototype p_j^s:

$$p_j^s \Leftarrow \arg\min_{z \in Z} \left\|z - p_j^s\right\|_2^2,$$

where $Z = \left\{z' : z' \in \text{patches}\left(f\left(x_i\right)\right) \forall i \text{ s.t. } y_i = 1\right\}$.

By doing so, each prototype corresponds to a patch of the feature map of a training sample, and hence, corresponds to a local region of an EEG training sample. It is worth noting that convolutional neural networks (CNNs) have a crucial property in that there is a spatial correspondence between the feature maps and the original input sample. In other words, every element of a feature map maps to a local region of the input sample. Therefore, each prototype corresponds to a local region of a raw EEG signal that represents a specific waveform pattern. It is worth noting that the prediction result will not change for samples in which the model prediction is correct and with some confidence before the projection if the "projection process" does not move the prototypes too much (guaranteed by the maximization of Clst loss). The proof of this statement can be found in [36].

3) Reasoning Process for Our Model: Figure 3 presents the reasoning process of our proposed MSPPNet in making a classification decision on a preictal test sample (from the CHB-MIT dataset) at the top of the figure. For a given test sample x, MSPPNet measures the similarity between the latent features $z = f(x)$ and the learned prototypes as evidence to make a final prediction. Since each prototype of trained MSPPNet represents some patch of the convolutional output, and this patch corresponds to a segment of raw preictal EEG signals, we can visualize the prototype by presenting this corresponding segment of training preictal EEG samples. For example, in Figure 3, our model attempts to obtain evidence for preictal by comparing its latent features with each prototype at different scales (shown in column "Prototype"). This comparison results in a similarity score for each prototype used for subsequent classification. The EEG waveforms in the second column shows the fragments of the test EEG sample, which corresponds to the most significantly activated patch by the prototype. The fourth column shows where the prototypes come from. Finally, a total of 60 similarity scores are weighted and summed together to provide a prediction probability of 0.9894 for this test sample belonging to the preictal class. More examples about the reasoning process of our model is presented in the Supplementary Material.

According to these examples, we can see that our model can learn highly effective prototypes for preictal and can capture similar EEG fragments to these prototypes in test samples. This further shows that our proposed model has a transparent reasoning process, which could provide a decision basis for seizure prediction.

C. Postprocess

In this study, we perform post-processing according to [28]. Specifically, a 60-second causal moving average filter is applied to the output of our model, and the prediction alarm threshold is set to 0.5 for all patients. Besides, a refractory period is determined to be 30 mins to avoid consecutive prediction alarms from occurring for a short period.

D. Comparative Methods

To the best of our knowledge, we are the first study that develops a self-interpretable deep learning model for epileptic seizure prediction. In order to evaluate the effectiveness of our model, we compare our model with several state-of-the-art non-interpretable methods.

CNN with Short-time Fourier transform (STFT) method [15] analyzed the time-frequency characteristic of

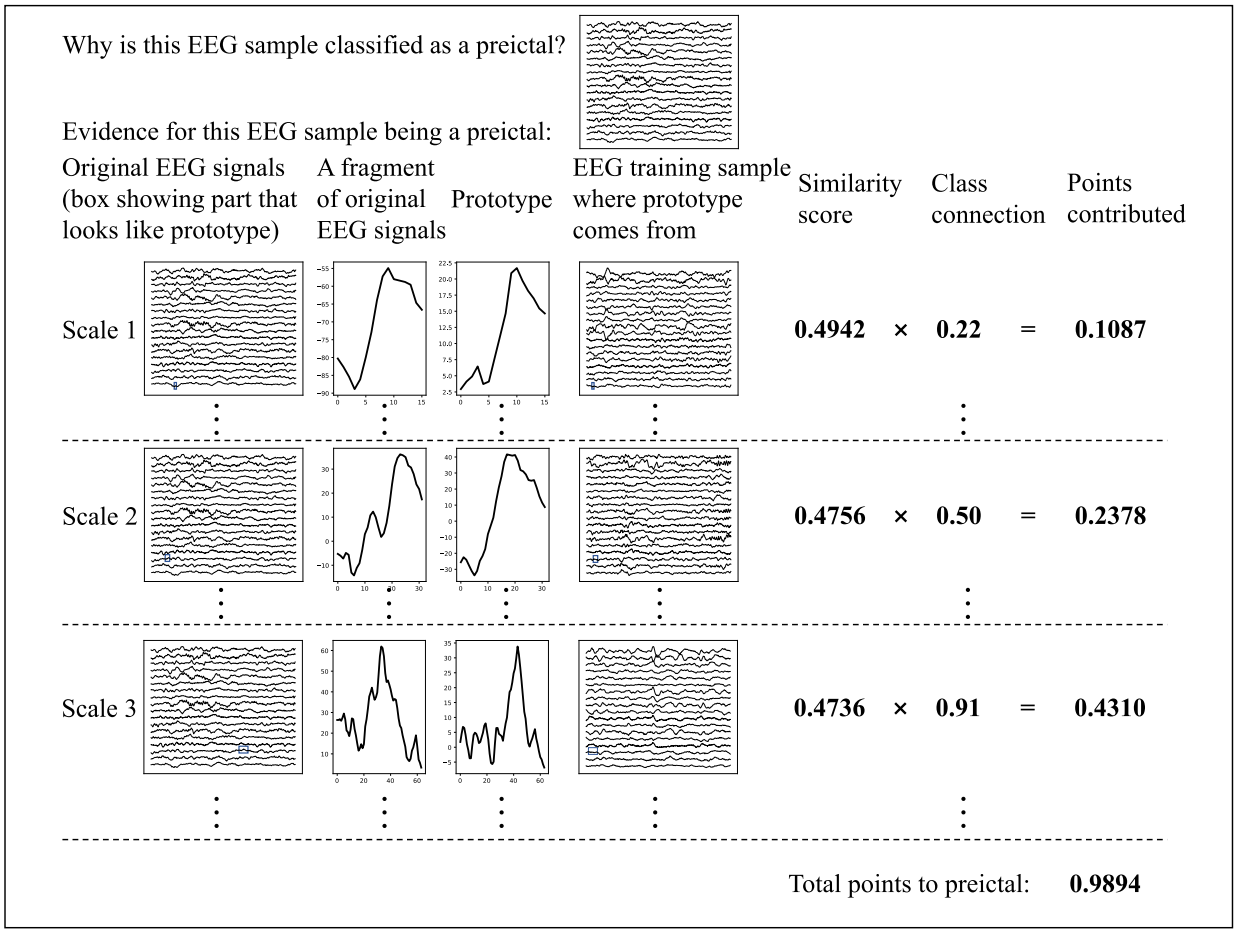

Fig. 3. The reasoning process of our model in deciding the class of a EEG sample (top).

EEG signals by the STFT approach, and the spectral images were used as input for the CNN model.

3D-CNN with Manual Features [22] developed a model by considering the location of the electrodes in EEG signals. A 3D CNN was applied to classify the extracted features with an image-based method.

CNN with Common Spatial Pattern Methods [33] applied a common spatial pattern approach to EEG signals to design spatial filters and obtain the discriminative features. Then a CNN was performed on these features to make final predictions.

GCN with Active Preictal Interval Learning Scheme [24] developed a spatio-temporal-spectral hierarchical graph convolutional network (GCN) to capture preictal biomarkers for predicting seizures. Besides, a semi-supervised active preictal interval learning method is employed to find the optimal patient-specific preictal interval.

Multi-scale CNN with Dilated Convolution [28] proposed a multi-scale CNN with dilated convolution model for end-to-end classification by considering that EEG signals carry information on multiple scales.

Adder Network with Supervised Contrastive Learning [7] proposed an end-to-end adder network and supervised contrastive learning to solve the problem of high complexity of deep learning networks.

Deep Learning with Semi-supervised Transfer Learning [10] designed four deep learning models for end-to-end seizure prediction, and a semi-supervised transfer learning method is used to improve performance.

CNN with two different dimensional kernels [37] adopts one- and two-dimensional kernels for end-to-end seizure prediction.

Among the above comparative methods,the first six methods are used for comparison in the CHB-MIT database, and the last three methods are used for comparison in the Kaggle database.

III. Results and Discussion

This section first describes the details of our experiments and the evaluation metrics. Then we present the experimental results on both datasets and the performance improvement of the multi-scale prototype layer compared to the existing single-scale prototype layer. Furthermore, the performance comparisons of our model with several state-of-the-art methods are presented. Finally, we provide analysis and discussion for our methods.

A. Experimental Settings and Evaluation Metrics

In this work, we use the leave-*one*-out cross validation (LOOCV) strategy to evaluate the effectiveness of our model

TABLE III
PREDICTION PERFORMANCE OF OUR PROPOSED MODEL AND THE SINGLE-SCALE PROTOPNET METHOD ON BOTH DATASETS

Subjects	Single-scale ProtoPNet				**Multi-scale ProtoPNet**			
	Sens (%)	Fpr/h	AUC	p-value	Sens(%)	Fpr/h	AUC	p-value
CHB-MIT:								
1	100.0	0.000	0.961	<0.001	100.0	0.000	0.965	<0.001
2	66.7	0.000	0.748	<0.001	100.0	0.000	0.771	<0.001
3	100.0	0.000	0.909	<0.001	100.0	0.000	0.890	<0.001
5	80.0	0.000	0.639	<0.001	100.0	0.101	0.728	<0.001
7	100.0	0.032	0.692	<0.001	100.0	0.129	0.788	<0.001
9	50.0	0.000	0.612	<0.001	50.0	0.000	0.636	<0.001
10	85.7	0.107	0.677	<0.001	100.0	0.294	0.765	<0.001
13	83.3	0.353	0.926	0.002	100.0	0.044	0.899	<0.001
14	87.5	0.178	0.566	<0.001	87.5	0.059	0.705	<0.001
16	87.5	0.486	0.756	<0.001	87.5	0.000	0.795	<0.001
17	100.0	0.000	0.905	<0.001	100.0	0.058	0.904	0.001
18	75.0	0.033	0.826	<0.001	75.0	0.000	0.811	<0.001
19	100.0	0.000	0.990	<0.001	100.0	0.000	0.990	0.001
20	100.0	0.000	0.981	<0.001	100.0	0.000	0.996	<0.001
21	100.0	0.000	0.844	<0.001	100.0	0.000	0.867	<0.001
23	100.0	0.000	0.931	<0.001	100.0	0.172	0.932	<0.001
Ave	88.5	0.074	0.810	\	**93.8**	**0.054**	**0.840**	\
Kaggle:								
Dog-1	50.0	0.164	0.582	<0.001	50.0	0.164	0.568	<0.001
Dog-2	100.0	0.157	0.844	<0.001	100.0	0.048	0.870	<0.001
Dog-3	100.0	0.210	0.755	<0.001	100.0	0.180	0.811	<0.001
Dog-4	57.1	0.267	0.633	<0.001	92.9	0.244	0.700	<0.001
Dog-5	100.0	0.188	0.892	<0.001	100.0	0.094	0.872	<0.001
Ave	81.4	0.197	0.741	\	**88.6**	**0.146**	**0.764**	\

according to work in [28], and the *"one"* indicates one seizure. For example, given an epileptic patient with a total of N seizures, there would be N corresponding preictal period. We then randomly separate this patient's interictal data into N equal parts and combine them with N preictal periods to be N pairs. As the LOOCV strategy shows, we take one of the pairs as the test set and the remaining N-1 pairs as the training set in each round. Therefore, our model is trained and tested N times for this patient. During training, our model is performed in Python 3.7.11 environment and Pytorch 1.11.0.

For evaluation metrics, we adopt four parameters to measure the performance of our model: the sensitivity (Sens, correctly predicted seizures as a proportion of total seizures), false prediction rate (Fpr, the number of false alarms per hour), AUC (area under the ROC Curve), and p-value.

As stated in [22], we calculate the p-value to measure the significance of our proposed model's improvement over the chance level. To do so, we assume that the interval between two successive alarms follows an exponentially distributed Poisson process, and we calculate the probability of at least one alarm rising randomly in an interval of Δt. This probability is independent of t and is approximately equal to $\lambda_w \Delta t$, where λ_w is the Poisson rate parameter.

In this context, the sensitivity of the chance predictor, S_{nc}, is defined as follows:

$$S_{nc} = 1 - \exp\left(-\lambda_w \tau_w + \left(1 - e^{-\lambda_w \tau_{w0}}\right)\right)$$

where the detection interval τ_{w0} denotes the seizure prediction horizon (SPH), while τ_w represents the sum of SPH and the seizure occurrence period (SOP). The difference between observed and chance sensitivity depend on ρ_w is a strong measure of predictability [38]. For a seizure prediction method with sensitivity S_n and proportion of time-in-warning ρ_w, the sensitivity improvement-over-chance metric is given as below:

$$S_n - S_{nc} = S_n - 1 + \exp\left(-\lambda_w \tau_w + \left(1 - e^{-\lambda_w \tau_{w0}}\right)\right)$$

where

$$\lambda_w = -\frac{1}{\tau_w} \ln\left(1 - \rho_w\right)$$

To calculate the significance of an improvement over chance, we assume that our proposed prediction method correctly identifies n out of N seizures for an individual subject. The one-sided p-values are calculated as follows:

$$p = 1 - \sum_{i=0}^{n-1} \binom{N}{i} S_{nc}^{i} \left(1 - S_{nc}\right)^{N-i}, \quad \text{for } \frac{n}{N} \geq S_{nc}$$

In this case, the corresponding hypotheses can be described as follows:

$$H_0 : \text{median } ((S_n - S_{nc}) \text{ for algorithm}) = 0$$
$$H_1 : \text{median } ((S_n - S_{nc}) \text{ for algorithm}) \neq 0$$

When we calculate the p-value, the significant level p is set to 0.05.

B. Results and Comparison

In Table III, we show the seizure prediction performance of our model for both two datasets: CHB-MIT and Kaggle. To further evaluate the effectiveness of our proposed

TABLE IV
COMPARISON TO RECENT EPILEPTIZ SEIZURE PREDICTION METHODS ON BOTH DATASETS

Authors	Dataset	Method	NO. of common subjects	Sens (%)	Fpr (/h)	Our Sens (%)	Our Fpr (/h)	Preictal length (mins)	Interictal distance (mins)	Intervention time (mins)
Truong et al. 2018 [15]	CHB-MIT	STFT+CNN	13	80.2	0.182	**93.3**	**0.046**	30	240	5
Ozcan et al. 2019 [22]	CHB-MIT	Manual features+3D-CNN	16	79.2	0.202	**93.8**	**0.054**	30	60	1
Zhang et al. 2020 [33]	CHB-MIT	CSP+CNN	16	93.1	0.103	**93.8**	**0.048**	30	240	\
Li et al. 2021 [24]	CHB-MIT	ICs from ICA+GCN	15	94.3	0.119	**93.3**	**0.049**	Adaptive	Adaptive	1
Gao et al. 2022 [28]	CHB-MIT	Multi-scale CNN	16	93.3	0.007	**93.8**	**0.054**	30	60	1
Zhao et al. 2022 [7]	CHB-MIT	Adder Network	15	91.8	0.067	**93.3**	**0.049**	30	240	1
Daoud et al. 2019 [10]	Kaggle	DCNN+Bi-LSTM	5	81.5	0.167	**88.6**	**0.146**	60	240	5
Xu et al. 2020 [37]	Kaggle	CNN	5	80.9	0.134	**88.6**	**0.146**	60	240	5
Zhao et al. 2022 [7]	Kaggle	ResCNN	5	81.2	0.161	**88.6**	**0.146**	60	240	5
Zhao et al. 2022 [7]	Kaggle	Adder Network	5	89.1	0.120	**88.6**	**0.146**	60	240	5

Fpr: false prediction rat; Sens: sensitivity; STFT: short-time Fourier transform; ICs: independent component.

multi-scale prototype layer, we also present the performance comparison of multi-scale ProtoPNet with single-scale ProtoPNet in Table III. Our MSPPNet achieves an average sensitivity of 93.8%, an average false prediction rate of 0.054/h, an average AUC of 0.840 on 16 patients in the CHB-MIT dataset, and an average sensitivity of 88.6%, an average false prediction rate of 0.146/h, and an average AUC of 0.764 in the Kaggle dataset. Besides, 14 out of 16 patients have a p-value less than 0.001 in the CHB-MIT dataset, and all subjects have a p-value less than 0.001 in the Kaggle dataset (significant level p is set to 0.05).

Compared with single-scale ProtoPNet, our model improves the prediction performance significantly, according to Table III above. Specifically, our model significantly increases the sensitivity, reduces the false prediction rate, and increases the AUC value on both datasets. This demonstrates the effectiveness of our proposed multi-scale prototype layer for epileptic seizure prediction.

Furthermore, we compare the performance of our model with several state-of-the-art non-interpretable methods, and the comparison results are listed in Table IV. To make a fair comparison, for each study to be compared, we use their performance obtained when they select the common subjects with us since different studies on the CHB-MIT dataset used different subjects. For example, in the first row of Table IV, Truong et al. study have 13 common subjects with us. On these 13 patients, they achieve an average Sens of 80.2% and an average Fpr of 0.182/h, and we achieve an average Sens of 93.3% and an average Fpr of 0.052/h. The performance of [10], [37] on the Kaggle dataset is derived from [7]. Besides, given that it is unfair to compare the Fpr if the interictal distance are different, we also perform the experiments on the CHB-MIT dataset at an interictal distance of 240 minutes. Hence, the performance in Table IV is compared when the preictal length and interictal distance are equal (except Li et al. 2021 [24], because their preictal length and interictal distance are adaptively inferred). According to Table IV, we can see that our proposed model can achieve comparable performance to the state-of-the-art non-interpretable methods.

C. Analysis and Discussion

The experimental results above show that our proposed model can provide both excellent interpretability and prediction performance. Nevertheless, some details of our model are worth discussing for further development. For example, the number of prototypes m for each scale in our model is predefined; how much does m affect the performance? To address this problem, we try different m in the model to consider the contribution of the number of prototypes. Specifically, we conducted experiments on the CHB-MIT dataset with m values ranging from 10 to 25, and the results are shown in Figure 4.

As the correct prediction of epileptic seizures is of utmost importance for epileptic patients, we prioritized the Sens metric when selecting the m value. As shown in the Figure 4 above, the Sens metric achieves its highest value (93.8%) when

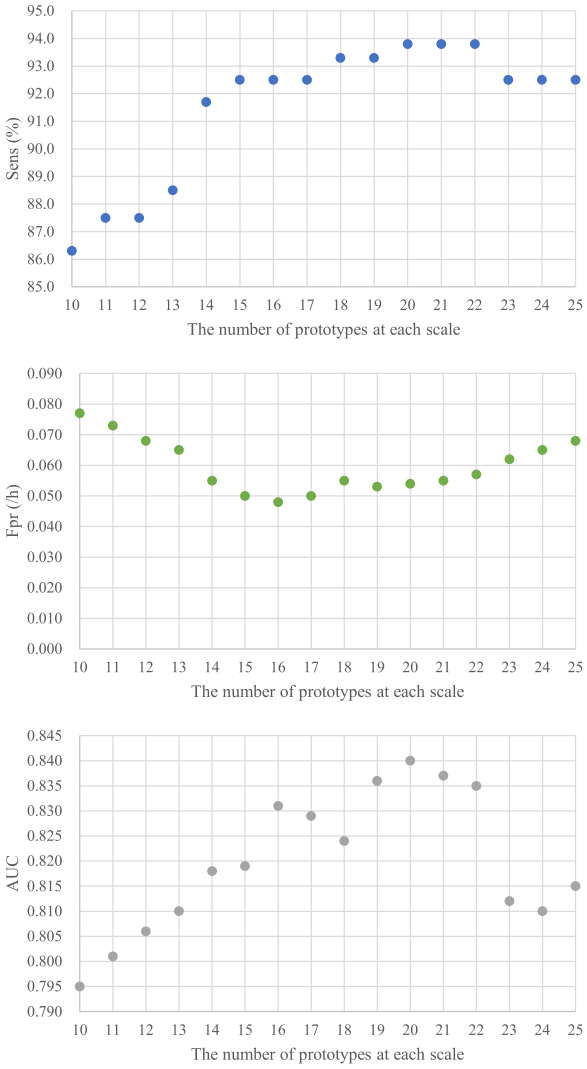

Fig. 4. Prediction performance with different number of prototypes on the CHB-MIT database.

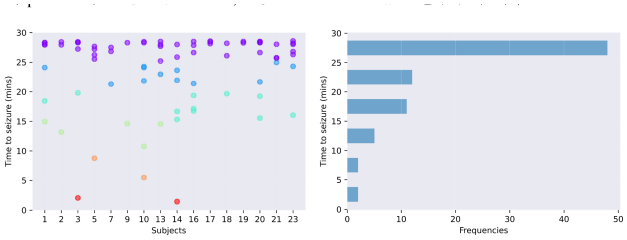

Fig. 5. The prediction time of our model for each patient in the CHB-MIT database.

m is at 20, 21, and 22. Therefore, we restricted the range of m values to between 20 and 22. For the Fpr, we found that the lowest value (0.054/h) was obtained when m was equal to 20 within this range. Furthermore, we observed that the AUC value was highest when m was equal to 20. Hence, we choose the number of prototypes at each scale, m, to be 20.

Besides, to further evaluate whether our proposed model can predict seizures in time, another indicator in terms of average prediction time (APT) is calculated, formulated by the average interval between the time of the alarm and the onset of the corresponding seizure. An effective seizure prediction model should have enough APT to give patients sufficient time to take preventive measures. The prediction time of our model for each patient in the CHB-MIT database is presented in Figure 5. Since we set the preictal length to 30 minutes, the false alarm longer than 30 minutes is not shown in Figure 5.

In Figure 5, each point on the left represents a patient's prediction time for a seizure. The distribution of prediction time for all seizures is shown on the right. We can see that all successful seizure prediction times in subjects 7, 17, 19, and 21 are greater than 20 minutes. In particular, for subjects 17 and 19, all successful seizure prediction times are greater than 25 minutes. According to the right picture, more than half of all seizures could be predicted at least 25 minutes before their onset. The APT of our model is 23.5 minutes, which could give patients sufficient time to take preventive measures.

Furthermore, given the critical importance of computational efficiency in online seizure prediction, we carefully measured the computation time of our proposed MSPPNet model. During training, our model consists of two stages. In the "Adam optimization for all layers" stage, our model takes about 80 milliseconds on the CHB-MIT database and 110 milliseconds on the Kaggle database to process a batch of samples (i.e., one gradient update, including 256 4-second EEG samples). For a specific subject, there are about 50000 training samples in the CHB-MIT dataset (e.g., patient-1) and about 16000 training samples in the Kaggle dataset (e.g., dog-1). Thus, it takes about 16 seconds on the CHB-MIT dataset and 7 seconds on the Kaggle dataset for our model to train an epoch. In the "Projection of prototypes" stage, our model takes about 1 second for 10000 training samples on the CHB-MIT dataset and 3 seconds for 10000 training samples on the Kaggle dataset. In our study, we set the number of training epochs to 60, and we perform a projection of prototypes when the epoch is greater than or equal to 10 and divisible by 5; otherwise, we perform Adam optimization for all layers. Therefore, the total training time on one subject is about 839 s on the CHB-MIT database and 409 s on the Kaggle database (using one single RTX 3080 GPU).

During the testing phase, we evaluated one hour of EEG data with our proposed MSPPNet model, which takes about 0.26 seconds on the CHB-MIT database and 0.19 seconds on the Kaggle database. These results suggest that our current model is fast enough for real-time prediction in the clinical setting.

In recent years, most deep learning-based seizure prediction studies focus on improving the prediction performance, while the deep learning model is often treated as a "black box". Recent work indicates that the absence of an explanation for the deep learning model seriously limits the clinical acceptance of EEG-based seizure prediction [21]. Hence, building an intrinsically interpretable deep learning model is crucial for developing seizure prediction. To this end, we construct MSPPNet with built-in interpretability. On the other hand, as far as we know, EEG biomarkers for preictal are scarce, which is the second reason hindering seizure prediction development. The EEG biomarkers for preictal are challenging to capture due to the complex spatial-temporal dynamics in the epileptic brain [24]. In our proposed MSPPNet, we can capture

several prototypes at multiple scales during training, which correspond to some fragments of raw EEG signals and can be viewed as potential biomarkers. Experimental results also show the efficiency of these potential biomarkers. Our study provides a novel way to search the bio-markers for preictal. In our future studies, we will think about how to make better use of these prototypes, and we believe that the effective potential biomarkers will further improve seizure prediction performance.

IV. Conclusion

In this study, we propose a self-interpretable deep learning model for patient-specific epileptic seizure prediction, which could provide both a transparent reasoning process and excellent prediction performance. Our model is evaluated on two public epileptic EEG datasets: CHB-MIT and Kaggle. After the LOOCV strategy, we obtain an average sensitivity of 93.8%, an average false prediction rate of 0.054/h in the CHB-MIT database, an average sensitivity of 88.6%, and an average false prediction rate of 0.146/h in the Kaggle dataset. The experimental results show that our model can achieve the current state-of-the-art performance with self-interpretable evidence. This work provides a promising solution for EEG-based epileptic seizure prediction.

References

[1] *Epilepsy: A Public Health Imperative*, World Health Org., Geneva, Switzerland, 2019.

[2] P. Kwan and M. J. Brodie, "Early identification of refractory epilepsy," *New England J. Med.*, vol. 342, no. 5, pp. 314–319, Feb. 2000.

[3] X. Tian et al., "Deep multi-view feature learning for EEG-based epileptic seizure detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 1962–1972, Oct. 2019.

[4] Y. Li, Y. Liu, W.-G. Cui, Y.-Z. Guo, H. Huang, and Z.-Y. Hu, "Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 4, pp. 782–794, Apr. 2020.

[5] C. Li et al., "Seizure onset detection using empirical mode decomposition and common spatial pattern," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 458–467, 2021.

[6] G. Wang et al., "Seizure prediction using directed transfer function and convolution neural network on intracranial EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2711–2720, Dec. 2020.

[7] Y. Zhao, C. Li, X. Liu, R. Qian, R. Song, and X. Chen, "Patient-specific seizure prediction via adder network and supervised contrastive learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1536–1547, 2022.

[8] L. Kuhlmann, K. Lehnertz, M. P. Richardson, B. Schelter, and H. P. Zaveri, "Seizure prediction—Ready for a new era," *Nature Rev. Neurol.*, vol. 14, no. 10, pp. 618–630, Oct. 2018.

[9] Y. Gao, A. Liu, X. Cui, R. Qian, and X. Chen, "A general sample-weighted framework for epileptic seizure prediction," *Comput. Biol. Med.*, vol. 150, Nov. 2022, Art. no. 106169.

[10] H. Daoud and M. A. Bayoumi, "Efficient epileptic seizure prediction based on deep learning," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 5, pp. 804–813, Oct. 2019.

[11] L. Chisci et al., "Real-time epileptic seizure prediction using AR models and support vector machines," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 5, pp. 1124–1132, May 2010.

[12] M. Bedeeuzzaman, T. Fathima, Y. U. Khan, and O. Farooq, "Seizure prediction using statistical dispersion measures of intracranial EEG," *Biomed. Signal Process. Control*, vol. 10, pp. 338–341, Mar. 2014.

[13] Z. Zhang and K. K. Parhi, "Low-complexity seizure prediction from iEEG/sEEG using spectral power and ratios of spectral power," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 3, pp. 693–706, Jun. 2016.

[14] S. M. Usman, M. Usman, and S. Fong, "Epileptic seizures prediction using machine learning methods," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–10, Dec. 2017.

[15] N. D. Truong et al., "Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram," *Neural Netw.*, vol. 105, pp. 104–111, Sep. 2018.

[16] R. Jana and I. Mukherjee, "Deep learning based efficient epileptic seizure prediction with EEG channel optimization," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102767.

[17] S. M. Usman, S. Khalid, and S. Bashir, "A deep learning based ensemble learning method for epileptic seizure prediction," *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104710.

[18] K. Rasheed, J. Qadir, T. J. O'Brien, L. Kuhlmann, and A. Razi, "A generative model to synthesize EEG data for epileptic seizure prediction," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2322–2332, 2021.

[19] D. Lei, X. Chen, and J. Zhao, "Opening the black box of deep learning," 2018, *arXiv:1805.08355*.

[20] C. Rudin, "Please stop explaining black box models for high stakes decisions," *Stat*, vol. 1050, p. 26, Nov. 2018.

[21] M. F. Pinto et al., "On the clinical acceptance of black-box systems for EEG seizure prediction," *Epilepsia Open*, vol. 7, no. 2, pp. 247–259, 2022.

[22] A. R. Ozcan and S. Erturk, "Seizure prediction in scalp EEG using 3D convolutional neural networks with an image-based approach," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 11, pp. 2284–2293, Nov. 2019.

[23] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for patient-independent epileptic seizure prediction using scalp EEG signals," *IEEE Sensors J.*, vol. 21, no. 7, pp. 9377–9388, Apr. 2021.

[24] Y. Li, Y. Liu, Y.-Z. Guo, X.-F. Liao, B. Hu, and T. Yu, "Spatio-temporal-spectral hierarchical graph convolutional network with semisupervised active learning for patient-specific seizure prediction," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12189–12204, Nov. 2022.

[25] I. Jemal, N. Mezghani, L. Abou-Abbas, and A. Mitiche, "An interpretable deep learning classifier for epileptic seizure prediction using EEG data," *IEEE Access*, vol. 10, pp. 60141–60150, 2022.

[26] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc interpretability for neural NLP: A survey," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–42, Aug. 2023.

[27] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 180–186.

[28] Y. Gao et al., "Pediatric seizure prediction in scalp EEG using a multi-scale neural network with dilated convolutions," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–9, 2022.

[29] H. Jaseja and B. Jaseja, "EEG spike versus EEG sharp wave: Differential clinical significance in epilepsy," *Epilepsy Behav.*, vol. 25, no. 1, p. 137, Sep. 2012.

[30] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[31] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[32] B. H. Brinkmann et al., "Crowdsourcing reproducible seizure forecasting in human and canine epilepsy," *Brain*, vol. 139, no. 6, pp. 1713–1722, 2016.

[33] Y. Zhang, Y. Guo, P. Yang, W. Chen, and B. Lo, "Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 465–474, Feb. 2020.

[34] S. Zhao, J. Yang, and M. Sawan, "Energy-efficient neural network for epileptic seizure prediction," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 1, pp. 401–411, Jan. 2022.

[35] C. Li, Y. Zhao, R. Song, X. Liu, R. Qian, and X. Chen, "Patient-specific seizure prediction from electroencephalogram signal via multi-channel feedback capsule network," *IEEE Trans. Cogn. Develop. Syst.*, early access, Oct. 5, 2022, doi: 10.1109/TCDS.2022.3212019.

[36] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[37] Y. Xu, J. Yang, S. Zhao, H. Wu, and M. Sawan, "An end-to-end deep learning approach for epileptic seizure prediction," in *Proc. 2nd IEEE Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Aug. 2020, pp. 266–270.

[38] D. E. Snyder, J. Echauz, D. B. Grimes, and B. Litt, "The statistics of a practical seizure warning system," *J. Neural Eng.*, vol. 5, no. 4, p. 392, 2008.