EEG-Based Seizure Prediction via Model Uncertainty Learning

Chang Li, *Member, IEEE*, Zhiwei Deng, Rencheng Song, *Member, IEEE*, Xiang Liu, Ruobing Qian, and Xun Chen, *Senior Member, IEEE*

***Abstract*— Deep neural networks (DNNs) have the powerful ability to automatically extract efficient features, which makes them prominent in electroencephalogram (EEG) based seizure prediction tasks. However, current research in this field cannot take the model uncertainty into account, causing the prediction less credible. To this end, we introduce a novel end-to-end patient-specific seizure prediction framework via model uncertainty learning. Specifically, we propose a reparameterized EEG-based lightweight CNN architecture and a modified Monte Carlo dropout (RepNet-MMCD) strategy to improve the reliability of the DNNs-based model. In RepNet, we obtain multi-scale feature representations by applying depthwise separable convolutions of different kernels. After training, depthwise convolutions with different scales are equivalently converted into a single convolution layer, which can greatly reduce computational budgets without losing model performance. In addition, we propose a modified Monte Carlo (MMCD) strategy, leveraging the samples-based temporal information in EEG signals to simulate the Monte Carlo dropout sampling. Sensitivity, false-positive rate (FPR), and area under curve (AUC) of the proposed RepNet-MMCD achieve 93.1%, 0.033/h, 0.950 and 81.6%, 0.056/h, 0.903 on two public datasets, respectively. We further extend the MMCD strategy to the other baseline methods, which can improve the performance of seizure prediction by a clear margin.**

Manuscript received 1 May 2022; revised 3 October 2022; accepted 25 October 2022. Date of publication 28 October 2022; date of current version 30 January 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFF1200600; in part by the National Natural Science Foundation of China Grant 32271431, Grant 61922075, Grant 41901350, Grant 62271186, and Grant 32150017; in part by the Fundamental Research Funds for the Central Universities under Grant JZ2021HGTB0078; and in part by the Anhui Key Project of Research and Development Plan under Grant 202104d07020005. *(Corresponding author: Xun Chen.)*

Chang Li, Zhiwei Deng, and Rencheng Song are with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China, and also with the Anhui Province Key Laboratory of Meauring Theory and Precision Instrument, School of Instrument Science and Optoelectronics Engineering, Hefei University of Technology, Hefei, Anhui 230009, China (e-mail: changli@hfut.edu.cn; zhiweideng@mail.hfut.edu.cn; rcsong@hfut.edu.cn).

Xiang Liu and Ruobing Qian are with the Epilepsy Centre, Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230001, China (e-mail: ahslyyliuxiang@163.com; qianruobing@fsyy.ustc.edu.cn).

Xun Chen is with the Epilepsy Centre, Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230001, China, and also with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230001, China (e-mail: xunchen@ustc.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3217929

***Index Terms*— Electroencephalogram (EEG), seizure prediction, RepNet, modified Monte Carlo dropout (MMCD), model uncertainty learning.**

I. Introduction

EPILEPSY is a common chronic brain disease, which is caused by abnormal discharges of brain neurons. More than 1% people are suffering from this disease worldwide [1] and about one-third of them cannot be effectively controlled by surgery [2]. With the development of recording electroencephalogram (EEG) signals and continual exploration of brain science, studies [3], [4], [5] have demonstrated the possibility of using EEG to predict seizure onset. A timely manual intervention before seizures can greatly reduce anxiety in patients and improve treatment effectiveness. Therefore, it is imperative to develop a high-precision EEG-based epilepsy prediction system to predict the onset of seizures for patients.

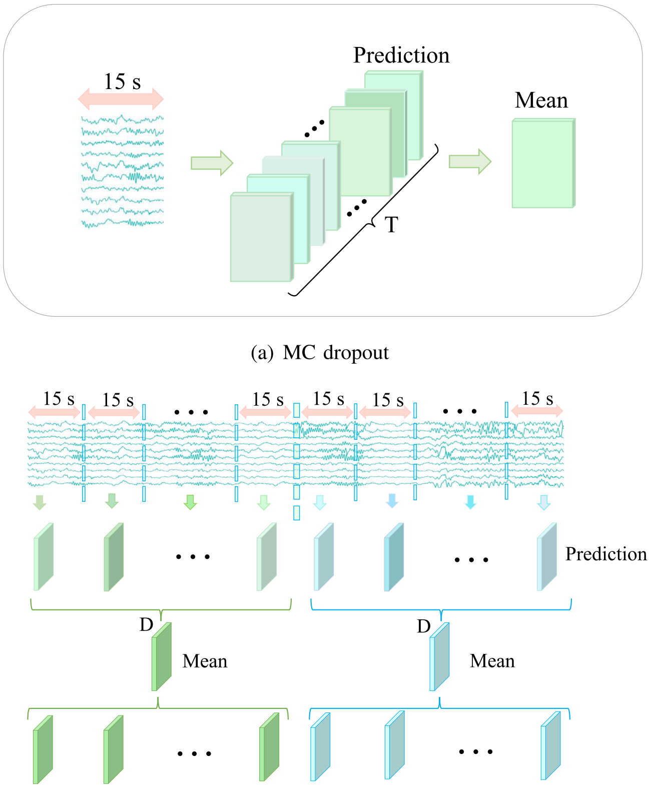
Linear or nonlinear features extracted using traditional machine learning methods, such as autoregressive coefficients [6] and Lyapunov exponent [7], are widely used to predict seizures. Then, the binary classification of EEG preictal and interictal states is implemented using classifiers, *e.g.* the k-nearest neighbor classifier [8] and the support vector machine (SVM) [4]. These traditional methods have achieved measurable improvements with well-handcrafted features. However, these hand-extracted features typically require extensive expertise and a lot of attempts. Besides, traditional classifiers with handcrafted features weaken the robustness in a more realistic setting with various artifacts affecting EEG recording.

Deep learning (DL) is gaining more attention for its excellent generalization and its powerful ability to automatically learn efficient features, encouraging its application in the field of epilepsy prediction. Many literatures [5], [9], [10], [11], [12], [13], [14] have shown that leading performance could be achieved with DNNs for seizure prediction, in contrast to traditional machine learning methods. Several studies [9], [12], [15], [16] have proposed methods to manually extract features from complex raw EEG, which are widely used to eliminate EEG artifacts. Khan et al. [15] processed the raw EEG signal using wavelet transform. Truong et al. [9] proposed to extract features from the raw EEG via the short-time Fourier transform (STFT). Li et al. [12] introduced a graph convolutional network that combines active learning and extracts information from the spatial-temporal spectrum for seizure prediction tasks. However, these DNNs designed using complex feature preprocessing of raw EEG signals typically require extra time and inevitably lead to information loss in the feature extraction process. Recently, studies [17], [18] have indicated that raw EEG signals could be used as inputs directly for seizure predictions as well, *e.g.*, 1D CNN [17] and the binary one-dimensional convolutional neural network (BSDCNN) [18]. Despite these networks being designed using 1D asymmetric convolutional layers or small kernel convolutional layers (*e.g.*, 3 × 3) to reduce the computational budget, the network overhead is still unsatisfactory. In this study, our reparameterized convolutional neural Network (RepNet) is an end-to-end lightweight network stacked by depthwise separable convolutions, which decomposes a standard convolution operation into two steps (*i.e.*, depthwise convolution and pointwise convolution operations) to reduce complexity and parameters budget. Inspired by the RepLKNet [19], we obtain much larger effective receptive fields via the 5 × 5 depthwise convolution. It helps to make up the optimization issue of large convolution kernels and enhances the feature representation capability by parallelizing with the 3 × 3 depthwise convolution. After training, depthwise convolutions with different kernels are fused into a single depthwise convolution layer, reducing the inference cost significantly.

Although deep neural networks (DNNs) can map input signals to a low-dimensional representation space by outputting a set of logits, these mappings are not unreservedly precise, and situations opposite to the truth can emerge unexpectedly. To overcome these challenges, researchers have proposed many uncertainty learning techniques in various fields such as hydrological forecasting [20], medical image analysis [21], [22], semantic segmentation and speech recognition [23], [24] over the last few years. However, an exploration of uncertainty in EEG-based models is currently lacking.

In general, the types of uncertainty could be roughly divided into data uncertainty and model uncertainty [25]. Data uncertainty (aleatoric uncertainty) describes the noise inherent in the EEG, such as muscle artifacts or electromagnetic interference [26]. Model uncertainty (epistemic uncertainty) captures the uncertainty of model parameters and can be reduced by increasing training samples. Modeling epistemic uncertainty has become attractive in improving the model performance of DNNs. However, standard neural networks can only provide deterministic values of weights instead of uncertainty estimates. Fortunately, Bayesian neural networks (BNN) [27], [28] provide a mathematically based framework to analyze uncertainty, which models uncertainty by considering model weights as probability distributions rather than point estimates. However, standard BNNs typically require a large number of calculations, which is a challenge for training. There are some Bayesian approximation techniques, and the MC dropout (MCD) [29] is considered to be one of the current mainstream approaches to capture the model uncertainty.

In the MCD method, a single sample is stochastically predicted T times by the dropout layer during testing, which usually increases the computational budget significantly. Generalization errors are reduced by using the average of the T models, as shown in Fig. 1 (a). We propose a modified MC dropout (MMCD) strategy, utilizing the coherence and the great informational similarity of the continuous EEG signals to predict EEG samples. Specifically, the MMCD predicts consecutive D samples during testing, and the calibrated probability is obtained by averaging the predictions of these D samples. Then, we perform D replications of the calibrated prediction to replace the predictions of the previous consecutive D samples (see Fig. 1 (b)). The proposed strategy requires only one prediction for each sample, which can greatly reduce the inference time and improve the performance of seizure prediction significantly.



(a) MC dropout

(b) modified MC dropout

Fig. 1. Comparison of prediction process between MC dropout and modified MC dropout in testing time. (a) The MC dropout samples T times, each for 15 s, and the calibrated predicted probability is obtained by the mean. (b) modified MC dropout (MMCD) aggregates D samples into the final prediction.

During the experiments, the Children's Hospital Boston and the Massachusetts Institute of Technology (CHB-MIT) [30] and the American Epilepsy Society Prediction Challenge (Kaggle) [31] databases are used as benchmarks to evaluate the model performance. The following are our main contributions:

1) We propose a novel re-parameterized lightweight end-to-end seizure prediction framework (RepNet-MCD) with uncertainty learning for multi-channel EEG-based seizure prediction. The proposed RepNet is able to share rich feature representations from different scales of depthwise convolutional layers with only the inference budget of a single convolutional layer during model deployment. Besides, we apply the MCD method for

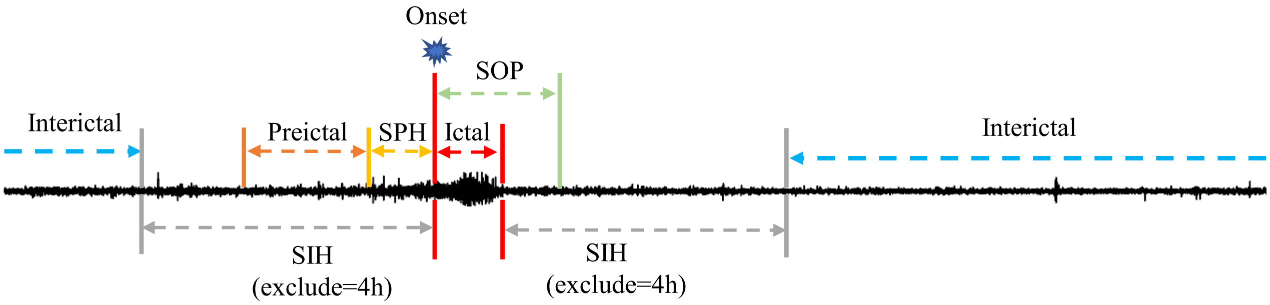


Fig. 2. Definition of interictal, preictal, SIH, SPH, SOP and seizure period (from the file *chb*01_03.edf).

uncertainty learning of the model and improve the model performance. To the best of our knowledge, this is the first time to combine DNNs with model uncertainty learning for seizure prediction.

2) We propose a modified EEG-based MC dropout strategy (RepNet-MMCD) that accelerates the inference speed during testing. Through extensive empirical experiments on two public datasets, our MMCD strategy significantly improves the performance and even surpasses the MCD approach in terms of effectiveness.
3) On 18 patients of the CHB-MIT database, RepNet-MMCD obtains 93.1%, 0.033/h, and 0.950 on sensitivity, FPR, and AUC respectively. In the Kaggle dataset, the proposed model reached 81.6%, 0.056/h, and 0.903 on sensitivity, FPR, and AUC, respectively.

The composition of the article is as follows. Section II introduces the details of the proposed method for seizure prediction. Experimental results on RepNet and other baseline models are provided in Section III. Section IV presents our discussion and Section V gives our conclusion.

II. Datasets and Methods

A. Datasets

In this work, we train and test on the CHB-MIT database [30] and the Kaggle database [31], respectively. The CHB-MIT dataset contains scalp electroencephalography (sEEG) signals from 23 pediatric subjects. These signals cover 844 hours of continuous EEG recordings with 182 seizure events, recorded using 22 electrodes with a sampling frequency of 256 Hz per second. In the Kaggle database, intracranial EEG (iEEG) signals are available for 5 dogs and 2 patients. There are 627.7 hours of interictal data with 48 seizure events recorded, supplied as 1 hour per sequence. Each one-hour sequence was divided into 10-minute segments, with the intervention period (SPH) being defined by the organizer as 5 minutes before seizure onset. Dogs 1-4 had iEEG signals collected with 16 electrodes at 400 Hz except Dog 5, which used 15 electrodes. Patient 1 collected iEEG data using 15 electrodes at 5000 Hz, while Patient 2 used 24 electrodes. We resample EEG signals in the Kaggle database to 200 Hz per second followed by [9].

B. Preprocessing

Different from numerous methods [5], [9], [16], [18] that take manually extracted features as input, our model automatically learns deep discriminative representations from raw EEG signals. It can reduce extra overhead in domain transformation and potential information degradation of the raw EEG signals. Besides, the sliding window analysis slices the long-range EEG signals into signal segments to yield sufficient samples available for training in deep neural networks. We adopt the definition of brain states (*i.e.*, preictal, interictal, seizure inter-ictal horizon (SIH), seizure prediction period (SPH), and seizure (SOP)) introduced by [9] and [32], as illustrated in Fig 2. The SPH [33] refers to the intervention period before seizures, where therapies (such as electrical stimulation) could be performed. The SOP indicates the period when seizures are anticipated to occur, which equals the preictal period in duration. For both publicly available datasets, we follow the 5 minutes SPH definition introduced by [5], [9], and [12]. The preictal phase implies a potential pattern of upcoming epilepsy, and the system alerts when patterns of interest are predicted during this phase. We define the preictal state as the 30 minutes preceding the onset of SPH, as in most works. For the CHB-MIT dataset, we used the setting of the 30 minutes SOP in [5], [9], and [34], while for the Kaggle dataset, the SOP is set to 1 hour introduced by [32] and [35]. EEG signals about 4 hours before seizure onset and 4 hours after the seizure ends are defined as SIH [10], where these signals are excluded to reduce the interference due to the adjacency of the ictal state. In cases where more than one seizure occurs in a short period, we assume only a leading seizure exists if the duration is less than 15 minutes after the last seizure [5]. Patients *chb*12 and *chb*15 are excluded from model training due to few available interictal signals and an average of one seizure per hour clinically. As in most studies, we also exclude patients *chb*04, *chb*06, and *chb*07 because of the heavy noise interference in the collected data. Based on the above definitions and considerations, a total of 87 epileptic events on 18 patients in the CHB-MIT database and 38 seizures from 4 dogs in the Kaggle database are assessed. We summarize these two datasets in Table I.

TABLE I
A Summary of the CHB-MIT and Kaggle Databases

Database	No. of patients	No. of Seizures	Interictal hours
CHB-MIT	18 patients	87	337.2
Kaggle	4 dogs	38	532.3

Seizure prediction tasks suffer from category imbalance of data. For most patients, interictal signals are much more than preictal signals, which greatly impact the final performance.

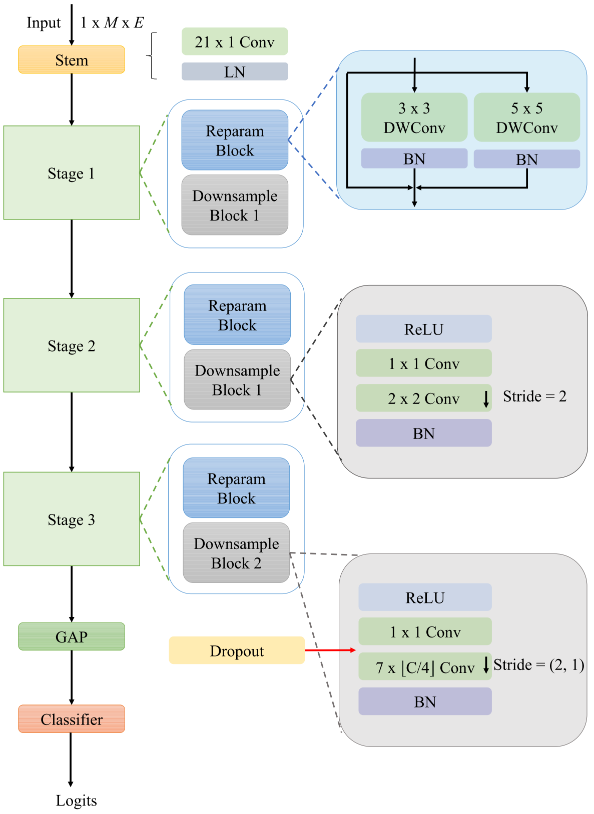


Fig. 3. The architecture of the proposed RepNet.

TABLE II
ARCHITECTURES OF THE REPNET ON *chb*_01 IN THE CHB-MIT DATASET

Block Name	Output Size	RepNet
Stem	$16 \times 382 \times 22$	$k = (21, 1)$, $s = (10, 1)$, $p = (0, 0)$
Reparam	$16 \times 382 \times 22$	$k_1 = (3, 3)$, $s_1 = (1, 1)$, $p_1 = (1, 1)$ $k_2 = (5, 5)$, $s_2 = (1, 1)$, $p_2 = (2, 2)$
Downsample1	$32 \times 191 \times 11$	$k_1 = (1, 1)$, $s_1 = (1, 1)$, bias = False $k_2 = (2, 2)$, $s_2 = (2, 2)$, bias = False
Reparam	$32 \times 191 \times 11$	$k_1 = (3, 3)$, $s_1 = (1, 1)$, $p_1 = (1, 1)$ $k_2 = (5, 5)$, $s_2 = (1, 1)$, $p_2 = (2, 2)$
Downsample1	$64 \times 95 \times 5$	$k_1 = (1, 1)$, $s_1 = (1, 1)$, bias = False $k_2 = (2, 2)$, $s_2 = (2, 2)$, bias = False
Reparam	$64 \times 95 \times 5$	$k_1 = (3, 3)$, $s_1 = (1, 1)$, $p_1 = (1, 1)$ $k_2 = (5, 5)$, $s_2 = (1, 1)$, $p_2 = (2, 2)$
Downsample2	$64 \times 45 \times 1$	$k_1 = (1, 1)$, $s_1 = (1, 1)$, bias = False $k_2 = (7, 5)$, $s_2 = (2, 1)$, bias = False
GAP	$64 \times 1 \times 1$	$k = (1, 1)$
Linear	2	-

Where k, s, p refers to the kernel, stride, and padding of conv, respectively. The input size of *chb*_01 is $1 \times (15 \times 256) \times 22$ for example.

To balance the inconsistent classes in the training set, we employ an overlapping sliding technique with a moving step S to obtain extra preictal signals, as in [9] and [34]. Specifically, let the window size of the sliding window analysis as W, the total length of the preictal signals is noted as P, and the length of the interictal signals is denoted I, then the ratio R is computed as P/I. The number of extra preictal segments after oversampling is calculated as $\frac{(P-W)}{S}+1$, where $S = W \times R$.

C. RepNet

Inspired by the RepLKNet [19], we use a stack of 3×3 and 5×5 depthwise convolutions to build the re-parameterized block, where the kernel size of 3×3 is employed to enhance the representational power of the 5×5 depthwise convolution. Fig. 3 illustrates an overview of the proposed RepNet architecture. The proposed RepNet contains an asymmetric Stem block, several Reparam blocks and downsample blocks, a global average pooling (GAP) layer, and a classification layer.

The proposed RepNet takes raw EEG slices as input directly. We convert the 2D multichannel EEG segments into the 3D tensor with a channel dimension of 1 ($X \in \mathbb{R}^{1\times M\times E}$), which allows for the usage of the 2D convolution layer. M indicates the number of signal points sampled within a time window (equals to $F \times W$, where W refers to the window size, and F denotes the sampling frequency, *e.g.*, $\mathbb{R}^{1\times(256\times15)\times22}$). E represents the number of electrodes used for sampling. The asymmetric stem block is designed to mitigate the extreme asymmetry in the size of the input tensor ($M \gg E$) and extract the initial features of EEG signals. The asymmetric stem block consists of a standard convolution without padding and a layer normalization (LN) layer. It is an efficient block that effectively improves the representation of raw signals at a low computational cost. Our model has three stages to generate different hierarchical representations, each stage consists of a Reparam block and downsampled block. In the Reparam block, we construct a 3×3 depthwise convolution parallel to the 5×5 one, then add up their outputs with the identity shortcut after the batch normalization (BN) layer. The downsample block contains a ReLU activation function, a point convolution, a standard convolution with the kernel size of 2 and the stride of 2, and a BN layer. The model ends with a global average pooling (GAP), and a classification layer. More details of the architecture settings can be viewed in Table II.

Structural reparameterization of convolutions is a technique of equivalently transforming model structures by utilizing the additivity of weights and biases in the convolution operation. In this work, we adopt this methodology to merge the weights and bias of the 3×3 kernel and BN in the Reparam block into the parallel 5×5 kernel during testing, as illustrated in Fig. 4. In this way, we enable the larger kernel capable of capturing small-scale patterns.

D. MC Dropout

Assume the model weights are abstracted as $\mathcal{W}$ after training, and then the posterior probability $P(\mathcal{W}|\mathcal{D})$ on the train set $\mathcal{D}$ is required. The posterior distribution $P(\mathcal{W}|\mathcal{D})$ describes a set of credible model parameters, which can be formulated by Bayesian inference as: $P(\mathcal{W}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{W})P(\mathcal{W})}{P(\mathcal{D})}$. However, the posterior probability in Bayesian inference is difficult to compute by integrating all model parameters in practice. Bayesian approximation techniques provide an accessible method, which is essentially an approximation to fit the posterior distribution with the simple distribution. The MC dropout (MCD) is a popular Bayesian approximation technique which utilizes a simple Bernoulli (B) distribution to approximate $P(\mathcal{W}|\mathcal{D})$.

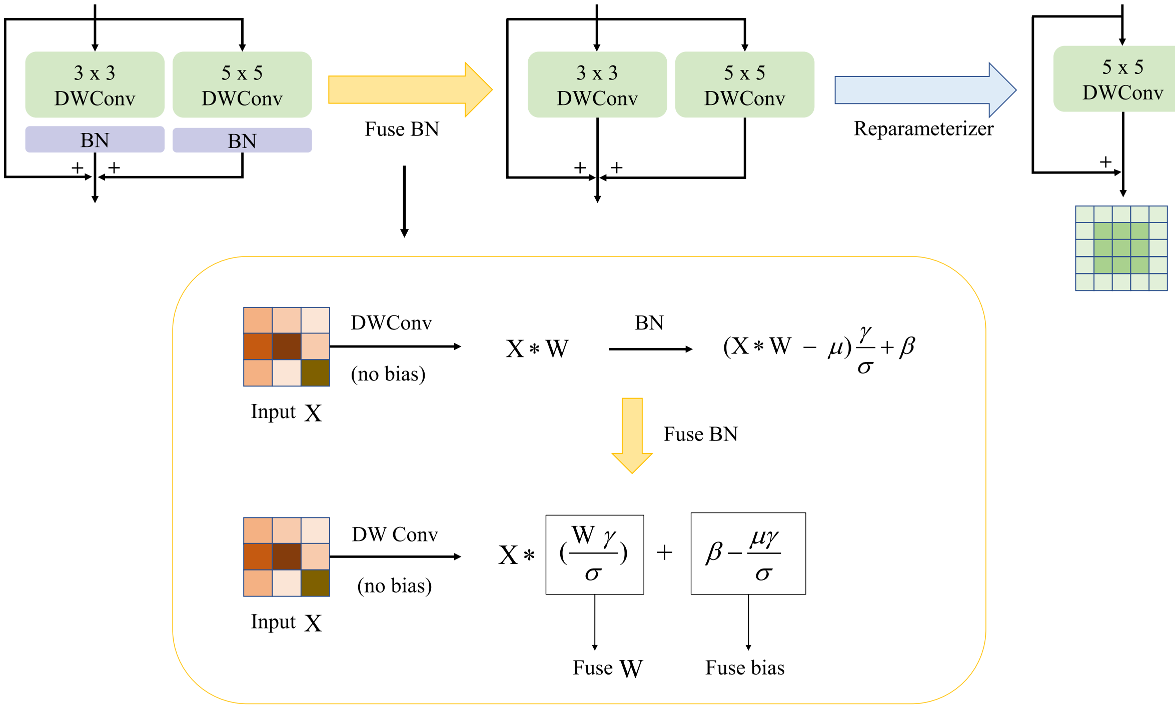


Fig. 4. Structural reparameterization of depthwise convolutions. The BN layer is first fused to the depthwise convolution, and then the small kernel is merged into the large one. The yellow areas depict the pipeline for fusing the BN into the deep convolution.

For a neural network with n dropout layers, the i-th dropout layer randomly deactivates some neurons of the model with a dropout rate p. Then, the weights W_i of the i-th layer in the model can be regarded as a Bernoulli distribution with the parameter p (*i.e.*, $W_i \sim \text{Bernoulli}(p)$). The posterior distribution of the model weights can be approximated as:

$$P(\mathcal{W}|\mathcal{D}) \approx \textstyle\prod_{i=1}^{n} B(W_i; p). \tag{1}$$

Kullback-Leibler (KL) divergence measures the distance between two distributions. It has been proved in [29] that the T predictions using the MCD essentially minimize $\text{KL}(B(\mathcal{W}; p) \parallel P(\mathcal{W}|\mathcal{D}))$, which is an optimization of the weights in the Bernoulli distribution. The cross-entropy loss is used in model uncertainty learning to optimize the neural network classifier, which minimizes the difference between our label distribution and the predicted distribution. The neurons for T sampling are random, which inevitably leads to fluctuations of predictions for the same sample. The mean of probability can be obtained by averaging the fluctuations of these T predictions, formally as:

$$P(y = c|x, \mathcal{D}) \approx \frac{1}{T}\sum_{t=1}^{T} \text{Softmax}(f^{\hat{\omega}_t}(x)), \tag{2}$$

where $\hat{\omega}_t$ denotes the t-th model weight predicted on input sample x, and $f^{\hat{\omega}_t}(x)$ indicates the logits of the model under the t-th model weight. In supervised learning, y is the target output and c is the category predicted from the sample x.

The uncertainty indicates the degree of confidence that the model predicts the EEG sample, which can be quantified by functions such as entropy:

$$H(y|x, \mathcal{D}) = -\sum_{c=1}^{C} P(y = c|x, \mathcal{D}) \log P(y = c|x, \mathcal{D})), \tag{3}$$

where C denotes the number of categories. The model is confident in its prediction when the value of entropy is low. Conversely, the model is uncertain about the prediction of the EEG sample.

E. Modefied MC Dropout

The MCD method performs T forward mappings on a single sample, where each mapping follows the Bernoulli distribution with minor differences. It allows for more reliable predictions than the baseline model due to the calibrated prediction probabilities. However, T (*e.g.*, $T = 5$) stochastic forward passes of a single sample lead to T times slower than the standard network prediction in testing time. Besides, it cannot explore the information relationship among successive samples. We propose a modified MC dropout (MMCD) strategy, which utilizes continuous samples-based temporal information to speed up the sampling process and obtain high-accuracy reliable prediction performance. The EEG signals of each state show tremendous information similarity in adjacent samples. Concretely, several consecutive EEG samples from a single patient within the same state contain minor information differences over a short period of time, which can be used to simulate the process of MC dropout sampling. In MMCD, each of the D following EEG segments is predicted only once, a total of D times (while the MCD needs $D \times T$ times for D samples). Notably, consecutive D EEG samples should maintain continuity in the time dimension, ensuring proper simulation of MC sampling during testing.

We assume that the test data set is composed of N consecutive EEG samples, *i.e.*, $X = \{x_1, x_2, \dots, x_N\}$, $Y = \{y_1, y_2, \dots, y_N\}$. Suppose that N is divisible by D, and the marginalization of the MMCD is performed to obtain the calibrated predictions:

$$P_i \approx \frac{1}{D} \sum_{d=i}^{i+D-1} \text{Softmax}(f(x_d)), \tag{4}$$

$$i = 1 + (n-1)D, \; n = \left\{1, 2, \dots, (\frac{N}{D} - 1)\right\}, \tag{5}$$

$$P_i = P_{i+1} = \dots = P_{i+(D-1)}, \tag{6}$$

where P_i is used to simplify $P(y_i = c|x_i, \mathcal{D})$, which refers to the probability predicted for the i-th test sample. $f(x_d)$ denotes the logits obtained from the d-th sample through the last layer of the network before the softmax, and D represents the number of samples aggregated. The MMCD strategy aggregates consecutive D samples in the test signals into the final prediction, and the entropy can be denoted as:

$$H_i = -\sum_{c=1}^{C} P_i(c) \log P_i(c), \tag{7}$$

$$i = 1 + (n-1)D, \; n = \left\{1, 2, \dots, (\frac{N}{D} - 1)\right\}, \tag{8}$$

$$H_i = H_{i+1} = \dots = H_{i+(D-1)}, \tag{9}$$

where H_i signifies the entropy of i-th test sample x_i, and $P_i(c)$ represents the c-th dimension of the P_i vector.

F. Postprocessing

The continuous event-based alerting proposal introduced by [9] is adopted for our seizure predictor. Specifically, it is considered a positive event if at least 120 s of 150 s consecutive signals are predicted as positive. A seizure with a positive event predicted in the preictal period is accepted as one successful prediction. Conversely, positive events predicted in the interictal period are taken to be false predictions. In addition, we adopt the 30-minute refractory period suggested in the literature [5], [9] to avoid frequent forecasts in a short time.

III. Results

In this section, we introduce the details of the experimental setting and explore the optimal window length for model performance. In addition, we conduct extensive ablation experiments to evaluate the effectiveness and efficiency of the proposed MMCD in improving the performance of DNNs on two public widely-used datasets and several architectures.

A. Experimental Settings

The evaluation metrics we chose were sensitivity (S_n) [36], false prediction rate (FPR) [9], [37], AUC, and p-value [38], which are widely utilized in event-based evaluation of seizure prediction. Sensitivity denotes the ratio of seizures correctly predicted to all seizures. FPR refers to the number of false predictions per hour, with a refractory period of 30 minutes. AUC is used to evaluate the classification performance of a model. A random classification model can reach an AUC value of 0.5, while a model with an AUC value of 1 is considered perfect. The p-value indicates the significance of the model prediction from a statistical perspective, which is considered significant over a random predictor when the p-value is lower than 0.05.

TABLE III
Test Performance Comparison on CHB-MIT Dataset Using Different Window Lengths

Model	Window	$S_n(\%)$	FPR/h	AUC	p-value	Training time (s)
RepNet	5 s	86.2	0.047	0.907	17/18	384.49
	10 s	87.4	0.062	0.916	18/18	221.35
	15 s	89.7	0.047	0.932	18/18	231.35
	30 s	88.5	0.077	0.931	16/18	229.42

Where S_n, FPR and AUC are reported as the mean of 18 patients in the CHB-MIT dataset, and the training time is tested on chb01. The ratio of the number of patients with p-values less than 0.05 to the total evaluated patients is recorded. The evaluation metrics for the experimental postprocessing are based on the same k-of-n strategy (*i.e.*, at least 120 seconds out of 150 seconds).

Our experiments are based on PyTorch 1.11.0, which is implemented using Python 3.7. The leave-one-out cross-validation method is adopted to train and evaluate the model. We optimize the loss using the AdamW [39] optimizer (lr $=$ 0.004, $\beta_1 = 0.9$, $\beta_2 = 0.999$), set the batch size as 128, and the epoch of training as 40. The classification layer is fine-tuned using a learning rate of 0.0003. Besides, the patience of 10 in the early stopping is used to reduce overfitting on training signals. All models are trained on NVIDIA Titan XP GP102.

B. Effects of Different Window Lengths of EEG Signals

Continuous long-range EEG signals are divided into small segments of seconds by sliding window analysis, and these segments are used as training data for the deep neural network. Current studies have various window lengths of EEG signals ranging from 4 to 30 seconds. An appropriate window length is expected to obtain better generalization performance. To this end, we evaluate the effect of different window lengths using the proposed RepNet. The post-processing of experiments is implemented based on the same k-of-n strategy (*i.e.*, at least 120 s out of 150 s) and the results are shown in Table III. Within 15 s, it contains more discriminative feature information as the window length increases, which improves the performance significantly. The AUC increases slightly when the EEG window length exceeds 15 s, indicating that window lengths over 15 s contain sufficient discriminatory information, and the classification performance learned by the RepNet eventually reaches a bottleneck. Besides, Table III shows the strong correlation between the training time and the window length. Within 10 s, the sliding window analysis using a longer window length results in a shorter model training time due to fewer generated samples. However, a longer window length implies a larger resolution as well, which leads to an increase in training time due to higher flops. We aim to strike a balance between classification performance and training speed and ultimately choose an EEG window of 15 s.

TABLE IV
PERFORMANCE COMPARISON OF ALL MODELS ON THE CHB-MIT DATASET

Patient	No.of seizures	Interical hours	CNN [17]				AdderNet [40]				RepNet-MMCD			
			S_n (%)	FPR/h	AUC	p-value	S_n (%)	FPR/h	AUC	p-value	S_n (%)	FPR/h	AUC	p-value
1	7	17	100.0	0.000	0.992	<0.001	100.0	0.000	1.0	<0.001	100.0	0.000	1.0	<0.001
2	3	23	33.3	0.000	0.725	<0.001	66.7	0.000	0.815	<0.001	66.7	0.000	0.846	<0.001
3	6	25	83.3	0.200	0.873	<0.001	83.3	0.040	0.922	<0.001	100.0	0.040	0.999	<0.001
5	5	14	100.0	0.071	0.968	<0.001	100.0	0.000	0.990	<0.001	100.0	0.000	1.0	<0.001
8	5	5	100.0	0.000	0.987	<0.001	100.0	0.200	0.981	<0.001	100.0	0.000	1.0	<0.001
9	4	46.3	50.0	0.043	0.678	0.003	25.0	0.043	0.596	**0.082**	75.0	0.043	0.809	<0.001
10	7	24	85.7	0.042	0.886	<0.001	85.7	0.000	0.916	<0.001	85.7	0.000	0.909	<0.001
11	2	32	100.0	0.031	0.983	<0.001	100.0	0.000	0.937	<0.001	100.0	0.000	0.997	<0.001
13	5	14	100.0	0.143	0.969	<0.001	100.0	0.143	0.996	<0.001	100.0	0.143	0.999	<0.001
14	7	5	85.7	0.800	0.748	0.006	71.4	0.400	0.764	0.003	85.7	0.200	0.777	<0.001
16	6	7	83.3	0.429	0.801	0.001	83.3	0.143	0.888	<0.001	83.3	0.000	0.924	<0.001
17	3	6	100.0	0.000	0.993	<0.001	100.0	0.000	1.0	<0.001	100.0	0.000	1.0	<0.001
18	6	24	100.0	0.083	0.957	<0.001	100.0	0.000	1.0	<0.001	100.0	0.000	1.0	<0.001
19	3	25	100.0	0.000	0.964	<0.001	100.0	0.040	0.999	<0.001	100.0	0.000	1.0	<0.001
20	6	20	83.3	0.050	0.996	<0.001	83.3	0.050	0.979	<0.001	100.0	0.000	0.997	<0.001
21	4	24	100.0	0.250	0.900	<0.001	100.0	0.125	0.969	<0.001	100.0	0.083	0.959	<0.001
22	3	13	100.0	0.385	0.849	0.005	66.7	0.385	0.724	**0.081**	66.7	0.231	0.883	0.033
23	5	12.9	100.0	0.078	0.990	<0.001	100.0	0.078	0.992	<0.001	100.0	0.000	1.0	<0.001
Average	87	337.2	89.7	0.098	0.903	-	87.4	0.063	0.915	-	93.1	0.033	0.950	-

where SPH is 5 min, SOP is 30 min, interictal-ictal distance is 4 hours, and p-value above 0.05 are marked in bold.

TABLE V
PERFORMANCE COMPARISON OF ALL MODELS ON THE KAGGLE DATASET

Patient	No.of seizures	Interical hours	CNN [17]				AdderNet [40]				RepNet-MMCD			
			S_n (%)	FPR/h	AUC	p-value	S_n (%)	FPR/h	AUC	p-value	S_n (%)	FPR/h	AUC	p-value
dog_2	7	83.3	71.4	0.048	0.885	<0.001	85.7	0.024	0.917	<0.001	85.7	0.012	0.971	<0.001
dog_3	12	240	75.0	0.075	0.864	<0.001	75.0	0.050	0.830	<0.001	75.0	0.042	0.873	<0.001
dog_4	14	134	92.9	0.396	0.701	<0.001	92.9	0.313	0.726	<0.001	85.7	0.127	0.829	<0.001
dog_5	5	75	60.0	0.107	0.916	0.001	80.0	0.040	0.892	<0.001	80.0	0.040	0.939	<0.001
Average	38	532.3	76.3	0.109	0.855	-	78.9	0.058	0.851	-	81.6	0.056	0.903	-

where SPH is 5 min, SOP is 60 min, the p-value is calculated as in [9].

TABLE VI
COMPARISON OF COMPUTATIONAL COMPLEXITY, NUMBER OF PARAMETERS, AND INFERENCE TIME OF ALL MODELS ON CHB-MIT AND KAGGLE DATASETS

Model	CHB-MIT			Kaggle		
	#Params (M)	MAdds (G)	Inference time	#Params (M)	MAdds (G)	Inference time
CNN [17]	1.072	0.175	0.172 s	0.679	0.098	0.514 s
AdderNet [40]	**0.116**	0.116	0.205 s	**0.116**	0.065	0.671 s
RepNet-MMCD	0.163	**0.030**	**0.126 s**	0.134	**0.018**	**0.315 s**

where the inference time in the CHB-MIT and Kaggle datasets are tested using NVIDIA Titan XP GP102 on *chb*_01 and dog_5, respectively. The best performance is marked in bold.

C. Peformance Comparison

To demonstrate the effectiveness of the proposed RepNet-MMCD, a representative CNN [17] and AdderNet [40] are employed as baselines for evaluation and comparison. CNN [17] first used the end-to-end paradigm of stacking max-pooling layers and standard convolutional to consistently downsample the features and extract higher semantic feature information, which obtains promising performance. AdderNet [40] introduced a lightweight addition convolutional network, which replaces the multiplication operation in traditional convolution with the addition operation to significantly reduce the computational cost. For fair comparisons, all models use the same configuration, *e.g.*, window length, and batch size.

Table IV and Table V demonstrate the performance of these baseline models on the CHB-MIT and Kaggle datasets, respectively. In terms of overall performance, the RepNet-MMCD shows a superior classification result, which supports the effectiveness of the proposed model. On 18 patients of the CHB-MIT database, the RepNet-MMCD achieves a performance of 93.1%, 0.033/h, and 0.950 on sensitivity, FPR, and AUC, respectively. In the Kaggle dataset, the proposed model reaches 81.6%, 0.056/h, and 0.903 on sensitivity, FPR, and AUC, respectively. Additionally, the computational cost (flops), the number of model parameters, and the inference time of these models are also measured, as shown in Table VI. As can be observed from Table VI, the proposed model exceeds AdderNet by more than 70% in computational power, and the model inference is nearly 40% faster than AdderNet with only about 40 KB more model size. We also investigate the impact of the computational budget due to structural reparameterization. As shown in

TABLE VII
COMPARISON OF COMPUTATION AND PARAMETERS BURDEN OF REPNET WITH STRUCTURAL REPARAMETERIZATION

Model	Structural re-parameterization	CHB-MIT			Kaggle		
		#Params (M)	MAdds (G)	Inference time	#Params (M)	MAdds (G)	Inference time
RepNet	×	0.164	0.034	0.172 s	0.135	0.020	0.408 s
	✓	0.163	0.030	0.122 s	0.134	0.018	0.306 s

where the inference time in the CHB-MIT and Kaggle datasets are tested on *chb*_01 and dog_5, respectively.

Table VII, we observe a significant reduction in computation and parameters overhead, which is quite useful for model deployment.

D. Ablation Studies

In this subsection, we experimentally investigate the effectiveness of the proposed model uncertainty learning through extensive empirical experiments on the CHB-MIT and Kaggle datasets.

The implementation of the MCD approach typically requires embedding at least one dropout layer into the model, which introduces two major factors that affect the performance of the MCD technique, *i.e.*, the dropout position, and the dropout rate. Appropriate usage of dropout benefits our baseline models. Kong et al. [41] suggests that a dropout layer embedded before the last-conv of the end of the network would be better. Inspired by this, we only embed a dropout layer before the last-conv in the downsample module, as illustrated by the red arrow in Fig. 3. Note that the dropout layer in the MCD method is required to be turned on in both training and testing. The MCD method performs T stochastic forward passes for a single EEG sample in the testing phase. These T models are fused in an average fashion, which typically captures more reliable individual predictions. However, the inference time of models increases linearly with the number of predictions, which limits the number of stochastic forward passes. It is sufficient to perform 5 stochastic forward passes through extensive experiments. In addition, different dropout rate tends to affect the performance varyingly, and the dropout mask in a classification network should not exceed 50%. Table VIII demonstrates the effects of different dropout rates on model performance. As observed from Table VIII, the proposed model can obtain the highest performance when the dropout rate is 0.1. On the CHB-MIT dataset, the MCD improves the sensitivity and AUC of the baseline by 1.28% and 1.5%, respectively, and reduces the FPR by 20.0%. On the Kaggle dataset, the MCD increases the sensitivity and AUC by 3.0% and 1.9%, respectively, and decreases the FPR by 12.1%.

Although the MCD technique is effective, it typically requires multiple stochastic forward passes, which inevitably leads to a linear increase in the inference latency. Besides, models with the MCD can only test each EEG sample independently, ignoring the large informational similarity between consecutive EEG samples. We propose the MMCD strategy, which utilizes the fact that EEG samples from a single patient in the same state tend to be consistent over short timescales. Note that the MMCD does not require dropout to get minor variations across samples. The number of sample aggregations is a crucial hyperparameter that affects the performance of the MMCD strategy. We conduct empirical experiments on the CHB-MIT and Kaggle datasets to evaluate the effect of different numbers of sample aggregations, as shown in Table IX. From Table IX, as the number of aggregates increases, the FPR and AUC consistently improve but exist a significant decrease in sensitivity. Thus, the aggregation of 2 samples (30 seconds) is the appropriate choice. For the CHB-MIT dataset, the MMCD technique raises the sensitivity and AUC by 3.8% and 1.9%, respectively, and reduces the FPR by 45%. For the Kaggle dataset, the MMCD improves the sensitivity and AUC by 3.8% and 1.9%, respectively, and reduces the FPR by 45%. More details about the performance and inference time for RepNet, RepNet-MCD, and RepNet-MMCD on the two public datasets are demonstrated in Table X and Table XI. As can be observed, the results demonstrate the effectiveness of MMCD, and the MMCD strategy consistently outperforms the MCD method. Moreover, comparisons of the average inference time reveal that the MMCD strategy is at least 5x (*i.e.*, $T = 5$) faster than MCD in terms of inference speed.

TABLE VIII
PERFORMANCE COMPARISON OF REPNET-MCD WITH DIFFERENT DROPOUT RATES ON CHB-MIT DATASET

Database	Dropout rate	S_n (%)	FPR/h	AUC	p-value (<0.05)
CHB-MIT	0.05	89.7	0.056	0.938	18/18
	0.1	**90.8**	**0.047**	**0.946**	18/18
	0.2	88.5	0.059	0.942	18/18
	0.3	89.7	0.056	0.944	17/18
	0.5	88.5	0.053	0.942	17/18
Kaggle	0.05	81.6	0.096	0.862	4/4
	0.1	**86.8**	**0.086**	**0.886**	4/4
	0.2	84.2	0.101	0.876	4/4
	0.3	78.9	0.103	0.866	4/4
	0.5	84.2	0.092	0.871	4/4

where bold font represents the best performance.

TABLE IX
PERFORMANCE COMPARISON WITH DIFFERENT NUMBERS OF SAMPLE AGGREGATIONS ON CHB-MIT AND KAGGLE DATASETS

Database	No. of samples (D)	S_n (%)	FPR/h	AUC	p-value (<0.05)
CHB-MIT	2	**93.1**	0.033	0.950	18/18
	3	92.0	0.033	0.952	18/18
	4	89.7	0.027	0.952	18/18
	5	87.4	**0.021**	**0.958**	18/18
Kaggle	2	**81.6**	0.056	0.903	4/4
	3	78.9	0.060	0.910	4/4
	4	73.7	**0.054**	0.920	4/4
	5	76.3	**0.054**	**0.922**	4/4

where the mean value for S_n, FPR, and AUC is taken from 18 patients and 4 dogs in the CHB-MIT and Kaggle dataset.

TABLE X
PERFORMANCE AND INFERENCE TIME COMPARISON OF THE REPNET, REPNET-MCD, REPNET-MMCD ON THE CHB-MIT DATABASE

Patient	RepNet					RepNet-MCD					RepNet-MMCD				
	S_n (%)	FPR/h	AUC	p-value	Inference	S_n (%)	FPR/h	AUC	p-value	Inference	S_n (%)	FPR/h	AUC	p-value	Inference
1	100.0	0.000	1.0	<0.001	0.1215 s	100.0	0.000	1.0	<0.001	0.6897 s	100.0	0.000	1.0	<0.001	0.1258 s
2	66.7	0.000	0.771	<0.001	0.3389 s	66.7	0.000	0.836	<0.001	1.9321 s	66.7	0.000	0.846	<0.001	0.3424 s
3	100.0	0.080	0.998	<0.001	0.1888 s	100.0	0.040	0.998	<0.001	1.0878 s	100.0	0.040	0.999	<0.001	0.1928 s
5	100.0	0.000	0.993	<0.001	0.1370 s	80.0	0.00	0.992	<0.001	0.7186 s	100.0	0.000	1.0	<0.001	0.1523 s
8	100.0	0.000	0.992	<0.001	0.0636 s	100.0	0.000	0.996	<0.001	0.3283 s	100.0	0.000	1.0	<0.001	0.0640 s
9	50.0	0.065	0.767	0.006	0.3983 s	50.0	0.043	0.813	<0.001	2.0789 s	75.0	0.043	0.809	0.001	0.4248 s
10	71.4	0.042	0.906	<0.001	0.1513 s	85.7	0.083	0.922	<0.001	0.8148 s	85.7	0.000	0.909	<0.001	0.1680 s
11	100.0	0.031	0.995	<0.001	0.6871 s	100.0	0.031	0.996	<0.001	3.6051 s	100.0	0.000	0.997	<0.001	0.7264 s
13	100.0	0.143	0.999	<0.001	0.1012 s	100.0	0.143	0.999	<0.001	0.5917 s	100.0	0.143	0.999	<0.001	0.1021 s
14	85.7	0.600	0.747	0.002	0.0480 s	85.7	0.20	0.773	<0.001	0.2672 s	85.7	0.200	0.777	<0.001	0.0481 s
16	83.3	0.000	0.896	<0.001	0.0520 s	83.3	0.000	0.915	<0.001	0.2908 s	83.3	0.000	0.924	<0.001	0.0601 s
17	100.0	0.000	1.0	<0.001	0.1026 s	100.0	0.000	1.0	<0.001	0.5864 s	100.0	0.000	1.0	<0.001	0.1031 s
18	100.0	0.042	0.998	<0.001	0.1816 s	100.0	0.000	0.998	<0.001	0.9246 s	100.0	0.000	1.0	<0.001	0.1852 s
19	66.7	0.000	0.999	<0.001	0.3592 s	100.0	0.000	0.999	<0.001	1.7303 s	100.0	0.000	1.0	<0.001	0.3693 s
20	83.3	0.100	0.931	<0.001	0.1555 s	100.0	0.050	0.991	<0.001	0.8563 s	100.0	0.000	0.997	<0.001	0.1646 s
21	100.0	0.167	0.927	<0.001	0.2645 s	100.0	0.125	0.939	<0.001	1.3037 s	100.0	0.083	0.959	<0.001	0.2645 s
22	100.0	0.308	0.849	0.003	0.1536 s	66.7	0.231	0.865	0.033	0.7103 s	66.7	0.231	0.883	<0.001	0.1611 s
23	100.0	0.000	1.0	<0.001	0.1267 s	100.0	0.000	1.0	<0.001	0.6838 s	100.0	0.000	1.0	<0.001	0.1292 s
Average	89.7	0.059	0.932	-	0.2017 s	90.8	0.047	0.946	-	1.0667 s	93.1	0.033	0.950	-	0.2102 s

TABLE XI
PERFORMANCE AND INFERENCE TIME COMPARISON OF THE REPNET, REPNET-MCD, REPNET-MMCD ON THE KAGGLE DATABASE

Patient	RepNet					RepNet-MCD					RepNet-MMCD				
	S_n (%)	FPR/h	AUC	p-value	Inference	S_n (%)	FPR/h	AUC	p-value	Inference	S_n (%)	FPR/h	AUC	p-value	Inference
dog_2	85.7	0.036	0.947	<0.001	0.2695 s	85.7	0.024	0.959	<0.001	1.0654 s	85.7	0.012	0.971	<0.001	0.2855 s
dog_3	75.0	0.046	0.836	<0.001	0.3558 s	83.3	0.042	0.865	<0.001	0.9306 s	75.0	0.042	0.873	<0.001	0.3995 s
dog_4	85.7	0.246	0.782	<0.001	0.2323 s	92.9	0.224	0.808	<0.001	0.8677 s	85.7	0.127	0.829	<0.001	0.2433 s
dog_5	60.0	0.067	0.905	<0.001	0.3057 s	80.0	0.067	0.913	<0.001	1.3122 s	80.0	0.040	0.939	<0.001	0.3157 s
Average	78.9	0.098	0.868	-	0.2908 s	86.8	0.086	0.886	-	1.0440 s	81.6	0.056	0.903	-	0.3110 s

TABLE XII
EFFECTS OF DIFFERENT CHANNEL SELECTIONS ON PERFORMANCE, COMPUTATION AND PARAMETERS BURDEN

Method	CHB-MIT	#Params (M)	MAdds (G)	Performance	Inference time
RepNet-MMCD	22 channels	0.163	0.030	93.1%-0.033-0.950	0.126 s
	18 channels	0.134	0.024	90.8%-0.044-0.937	0.109 s
	10 channels	0.077	0.013	88.5%-0.056-0.934	0.061 s
	6 channels	0.048	0.007	80.5%-0.083-0.909	0.045 s

where the inference time in the CHB-MIT and Kaggle datasets are tested on chb_01 and dog_5, respectively.

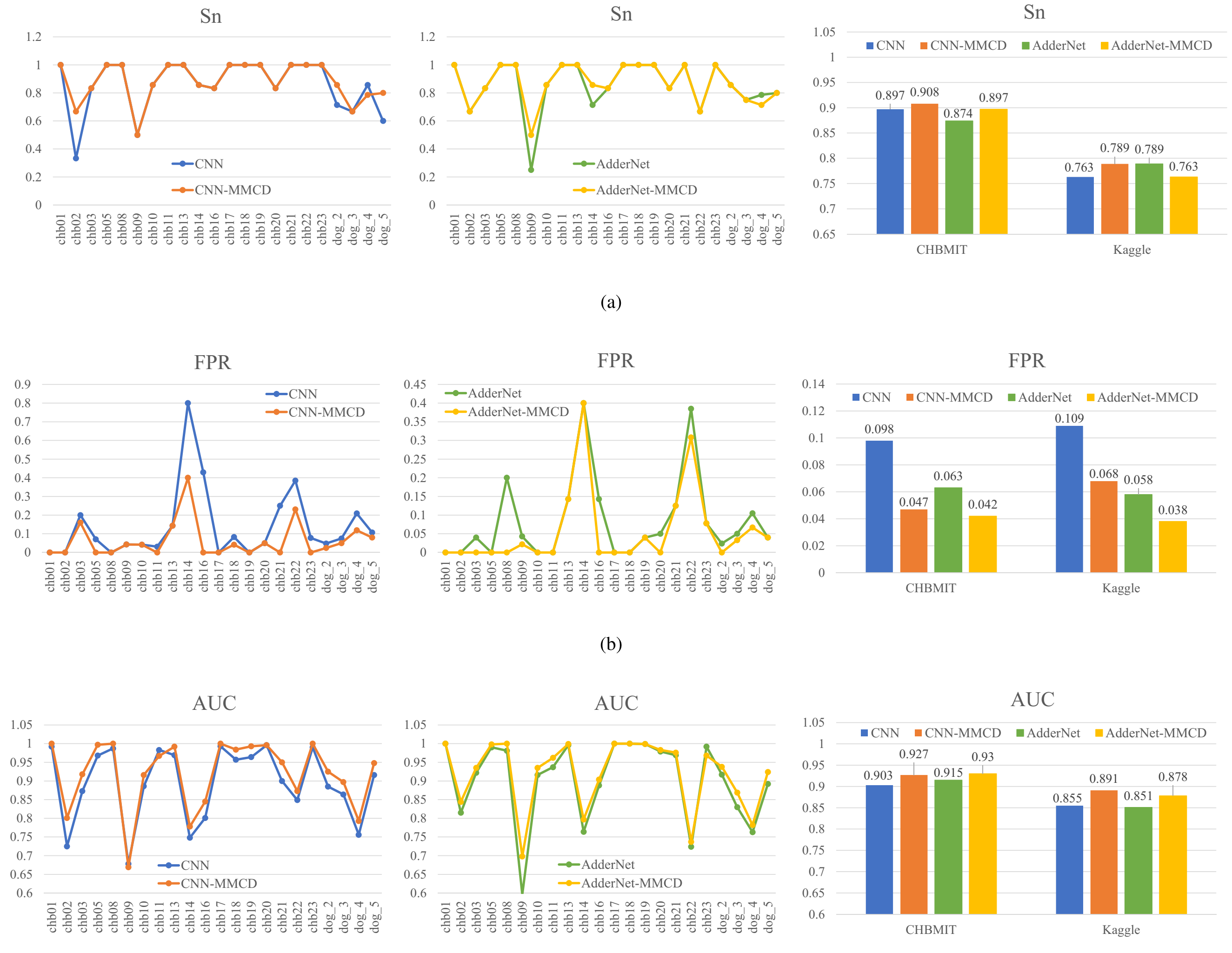
E. Effects of Channel Selection and Training Data

Channel selection has been an interesting issue, and different channel selections can affect the generalization capability of the model. Fewer and more important channels can save a large amount of model computational overhead with slight performance degradation, which is beneficial for the promotion of portable head-mounted EEG devices. We followed the setup of the studies [34] and [42] and conducted empirical experiments for several scenarios with different channel selections on the CHB-MIT dataset. Table XII shows the comparison of performance, parameters, and computational burden for the 22 channels (ours), 18 channels [34], 10 channels [42], and 6 channels [42] settings. For the 10- and 6-channel scenarios, their channel selections are focused on the temporal areas of the brain. See [34] and [42] for more details on specific channel selections. Generally, the model delivers superior performance with the increasing number of EEG channels, but with higher parameter and computational costs. We also observe that several patients perform better with fewer channels. This may be attributed to some electrodes being heavily contaminated during signal acquisition, and removing these electrodes would lead to better model predictions.

In the leave-one-out cross-validation strategy, the ratio of training data to validation data usually varies across methods, which may lead to different model generalizability. We also conduct experiments on several common scenarios, and the average performances are shown in Table XIII. As can be observed, the ratio of training data to validation data of 80% to 20% is a good choice.

F. Apply the MMCD Strategy to Other Baseline Models

In this subsection, we integrated the MMCD into other DNNs-based models such as CNN [9] and AdderNet [40]. Fig. 5 shows the performance of CNN, CNN-MMCD, AdderNet and AdderNet-MMCD methods. In the CHB-MIT database, the MMCD strategy improves the sensitivity of baseline CNN and AdderNet by 1.28% and 2.63%, respectively, and improves the AUC by 2.66% and 1.64%. The FPR of



(a)

(b)

(c)

Fig. 5. Performance evaluated on the CHB-MIT and Kaggle datasets after applying the MMCD strategy to the baseline CNN and AdderNet. (a) The two left graphs show the sensitivity before and after applying the MMCD strategy to the baseline CNN and AdderNet, respectively, while the right graph reports the average sensitivity of the two datasets. (b) The effect of MMCD strategy on the FPR of these baseline models. (c) Comparison of AUC before and after applying MMCD strategy to these baseline models.

TABLE XIII
Comparison of Different Rates Between the Training Data and Validation Data

Method	train:validation	S_n (%)	FPR (/h)	AUC
RepNet-MMCD	90%:10%	93.1%	0.036	0.946
	* 80%:20%	93.1%	0.033	0.950
	70%:30%	90.8%	0.033	0.938
	50%:50%	90.8%	0.039	0.939

where * represents the setting used in this study.

these baseline models is reduced by 51.5% and 33.3%. On the Kaggle database, the sensitivity, and the AUC of the baseline CNN are lifted by 3.45% and 4.21%, respectively, and the FPR decreased by 37.9%. We also observe that the MMCD strategy improves the AUC of AdderNet by 3.17% and the FPR effectively reduces by 35.5%, but the sensitivity decreases slightly by 3.33%. The effectiveness of the proposed MMCD is further validated by these empirical experiments.

IV. Discussion

In this section, some state-of-the-art seizure prediction methods in recent years are summarized in Table XIV. Notably, it is challenging to compare directly between different methods due to the diversity of selected patient data, postprocessing strategies, and preprocessing. For instance, Jemel et al. [37] employed 5-fold cross-validation instead of leave-one-out cross-validation to evaluate performance, and the SOP in the literature is not available. Zhao et al. [32] evaluated the performance of the CHB-MIT dataset with only 10 subjects selected, which could not sufficiently validate the generalizability of their model.

In addition, Li et al. [43] used the serial paradigm of CNN with the transformer to get outstanding results, but the multi-headed attention mechanism (MHSA) of quadratic

TABLE XIV
Performance of Existing Methods

Author	Dataset	Features	Classfier	SPH-SOP (min)	No. of seizures	Intericatal-Preictal Intervals (min)	Evaluated hours	Average S_n(%)-FPR(/h)-AUC
Khan *et al.* 2017 [15]	MIT, 15 patients	Wavelet transform coefficient	CNN	NA-10	18	10	70.5	87.8-0.147-0.866

Truong *et al.* 2018 [9]	FB, 13 patients; MIT, 13 patients; Kaggle, 5 dogs, 2 patients	STFT spectrograms	CNN	5-30	59; 64; 48	240	311.4;
							209;
							627.7
								81.4-0.06-NA
								81.2-0.16-NA
								75.0-0.21-NA

Ozcan *et al.* 2019 [5]	MIT, 16 patients	Spectral power, statistical moments, Hjorth	3D CNN	1-30
				5-30
					77	60; 120; 240	466.1;
							419.4;
							353.5
								86.8-0.292-NA
								87.0-0.186-NA
								85.7-0.096-NA

Daoud *et al.* 2019 [11]	MIT, 8 patients	Raw data	DACE+Bi-LSTM	NA-60	43	240	NA	99.72-0.004-NA
Zhang *et al.* 2019 [10]	MIT, 23 patients	Wavelet	CSP-CNN	30-NA	156	240	NA	92.2-0.12-NA
Xu *et al.* 2020 [17]	MIT, 7 patients; Kaggle, 5 dogs	Raw data	1D CNN	5-30	27; 44	240	NA	98.8-0.074-0.988
Wang *et al.* 2020 [44]	FB, 19 patients	Channel-frequency maps	CNN	5-30	82	NA-30	459.1	90.8-0.08-NA
Yang *et al.* 2021 [13]	MIT, 13 patients	STFT spectral images	RDANet	5-30	64	240	268.6	89.25-NA-0.913
Zhao *et al.* 2021 [32]	MIT, 10 patients; Kaggle, 5 dogs, 2 patients	Raw data	CNN	5-60	NA	NA	NA	99.8-0.005-1.000 93.5-0.063-0.977
Chen *et al.* 2021 [45]	Kaggle, 5 dogs, 2 patients	STFT	STCNN	5-30	64	NA	627.6	82.0-0.380-0.746
Li *et al.* 2022 [46]	MIT, 19 patients; Kaggle, 5 dogs	Raw	FB-CapsNet	1-30 5-30	105-42	240	375.9; 612.3	95.7-0.087-0.948 88.6-0.127-0.837
Jemal *et al.* 2022 [37]	MIT, 23 patients	Raw data	CNN	5-30 5-30	166	NA	940	96.1-0.040-0.918
Li *et al.* 2022 [43]	MIT, 16 patients; Kaggle, 5 dogs	STFT	TGCNN	5-30 5-30	82-45	240	296.8; 612.3	91.5-0.145-0.935 82.2-0.060-0.835
This work	MIT, 18 patients; Kaggle, 4 dogs	Raw data	RepNet-MMCD	5-30 5-60	87-38	240	337.2; 532.3	93.1-0.033-0.950 81.6-0.056-0.903

where NA means not applicable in the relevant work, and No. of seizures is the number of seizures participating in the assessment.

complexity in the transformer resulted in a large parameter and computational budget. We propose a novel lightweight seizure prediction framework. It uses depthwise separable convolutions to reduce parameters and computational overhead, and structural reparameterization is employed to further reduce computational costs during deployment. For the first time, we propose a method (MCD) to reduce model uncertainty from the perspective of uncertainty, which can be easily integrated into a single deterministic network but often requires T times of forwarding pass. We propose the MMCD strategy to simulate the process of MCD sampling based on the consistency of adjacent EEG samples, which can further improve the reliability of the EEG-based models while overcoming the drawbacks of the MCD in terms of prediction speed. Empirical experiments on multiple baselines demonstrate the effectiveness of our model uncertainty techniques.

V. Conclusion

This paper seeks to explore a more credible prediction to improve the performance of seizure prediction tasks. We propose a novel end-to-end and EEG-based patient-specific seizure prediction framework (ReptNet-MMCD) from the perspective of model uncertainty. For RepNet, it is a lightweight network stacked using depthwise convolutions, and we use structural reparameterization to further reduce the computation and parameters overhead during model deployment. For the MCD method, we demonstrate that the proper usage of the dropout layer offers modest performance improvements to the architecture. We also propose the MMCD strategy to simulate the MCD sampling based on the similarity between consecutive samples in a short time. Empirical experiments demonstrate that the proposed MMCD strategy outperforms the MCD in terms of performance and achieves 5x faster than MCD (*e.g.*, $T = 5$) in terms of inference speed. In addition, the MMCD strategy is further extended to the baseline CNN and AdderNet, which significantly improves the performance of these architectures. We hope the MMCD strategy could help the DNNs-based architectures for more reliable seizure prediction.

References

[1] M. J. Cook et al., "Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: A first-in-man study," *Lancet Neurol.*, vol. 12, no. 6, pp. 563–571, 2013.

[2] D. J. Thurman et al., "Standards for epidemiologic studies and surveillance of epilepsy," *Epilepsia*, vol. 52, pp. 2–26, Sep. 2011.

[3] A. Yadollahpour and M. Jalilifar, "Seizure prediction methods: A review of the current predicting techniques," *Biomed. Pharmacol. J.*, vol. 7, no. 1, pp. 153–162, 2015.

[4] H.-T. Shiao et al., "SVM-based system for prediction of epileptic seizures from iEEG signal," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 5, pp. 1011–1022, May 2017.

[5] 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.

[6] 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.

[7] L. D. Iasemidis et al., “Adaptive epileptic seizure prediction system,” *IEEE Trans. Biomed. Eng.*, vol. 50, no. 5, pp. 616–627, May 2003.

[8] S. Lahmiri and A. Shmuel, “Accurate classification of seizure and seizure-free intervals of intracranial EEG signals from epileptic patients,” *IEEE Trans. Instrum. Meas.*, vol. 68, no. 3, pp. 791–796, Mar. 2019.

[9] 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.

[10] 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.

[11] 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.

[12] 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. 2021.

[13] X. Yang, J. Zhao, Q. Sun, J. Lu, and X. Ma, “An effective dual self-attention residual network for seizure prediction,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1604–1613, 2021.

[14] Q. Xin, S. Hu, S. Liu, L. Zhao, and S. Wang, “WTRPNet: An explainable graph feature convolutional neural network for epileptic EEG classification,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 3s, pp. 1–18, Oct. 2021.

[15] H. Khan, L. Marcuse, M. Fields, K. Swann, and B. Yener, “Focal onset seizure prediction using convolutional networks,” *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 2109–2118, Sep. 2017.

[16] Q. Xin, S. Hu, S. Liu, L. Zhao, and Y.-D. Zhang, “An attention-based wavelet convolution neural network for epilepsy EEG classification,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 957–966, 2022.

[17] 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.

[18] S. Zhao, J. Yang, Y. Xu, and M. Sawan, “Binary single-dimensional convolutional neural network for seizure prediction,” in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Oct. 2020, pp. 1–5.

[19] X. Ding, X. Zhang, J. Han, and G. Ding, “Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11963–11975.

[20] D. Klotz et al., “Uncertainty estimation with deep learning for rainfall–runoff modelling,” *Hydrol. Earth Syst. Sci. Discuss.*, vol. 26, pp. 1–32, Mar. 2021.

[21] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, “Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation,” *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101557.

[22] J. C. Reinhold et al., “Validating uncertainty in medical image translation,” in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 95–98.

[23] S. Däubener, L. Schönherr, A. Fischer, and D. Kolossa, “Detecting adversarial examples for speech recognition via uncertainty quantification,” 2020, *arXiv:2005.14611*.

[24] C. J. Holder and M. Shafique, “Efficient uncertainty estimation in semantic segmentation via distillation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3087–3094.

[25] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5574–5584.

[26] X. Chen, C. Li, A. Liu, M. J. McKeown, R. Qian, and Z. J. Wang, “Toward open-world electroencephalogram decoding via deep learning: A comprehensive survey,” *IEEE Signal Process. Mag.*, vol. 39, no. 2, pp. 117–134, Mar. 2022.

[27] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural network,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[28] D. Krueger, C.-W. Huang, R. Islam, R. Turner, A. Lacoste, and A. Courville, “Bayesian hypernetworks,” 2017, *arXiv:1710.04759*.

[29] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

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

[32] 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.

[33] T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H. U. Voss, A. Schulze-Bonhage, and J. Timmer, “Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic,” *Phys. D, Nonlinear Phenomena*, vol. 194, nos. 3–4, pp. 357–368, 2004.

[34] A. Bhattacharya, T. Baweja, and S. P. K. Karri, “Epileptic seizure prediction using deep transformer model,” *Int. J. Neural Syst.*, vol. 32, no. 2, Feb. 2022, Art. no. 2150058.

[35] S. M. Usman, S. Khalid, and M. H. Aslam, “Epileptic seizures prediction using deep learning techniques,” *IEEE Access*, vol. 8, pp. 39998–40007, 2020.

[36] 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.

[37] 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.

[38] B. Schelter et al., “Testing statistical significance of multivariate time series analysis techniques for epileptic seizure prediction,” *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 16, no. 1, 2006, Art. no. 013108.

[39] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017, *arXiv:1711.05101*.

[40] 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.

[41] X. Kong, X. Liu, J. Gu, Y. Qiao, and C. Dong, “Reflash dropout in image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6002–6012.

[42] R. Asif, S. Saleem, S. A. Hassan, S. A. Alharbi, and A. M. Kamboh, “Epileptic seizure detection with a reduced montage: A way forward for ambulatory EEG devices,” *IEEE Access*, vol. 8, pp. 65880–65890, 2020.

[43] C. Li, X. Huang, R. Song, R. Qian, X. Liu, and X. Chen, “EEG-based seizure prediction via transformer guided CNN,” *Measurement*, vol. 203, Nov. 2022, Art. no. 111948.

[44] 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.

[45] R. Chen and K. K. Parhi, “Seizure prediction using convolutional neural networks and sequence transformer networks,” in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 6483–6486.

[46] 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. Cognit. Develop. Syst.*, early access, Oct. 5, 2022, doi: 10.1109/TCDS.2022.3212019.