# ScoreNet: A neural network-based post-processing model for identifying epileptic seizure onset and offset in EEGs

Poomipat Boonyakitanont[a], Apiwat Lek-uthai[a], Jitkomut Songsiri[a,1]

[a] *Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand*

**Abstract**

This article aims to design an automatic detection algorithm of epileptic seizure onsets and offsets in scalp EEGs. A proposed scheme consists of two sequential steps: the detection of seizure episodes, and the determination of seizure onsets and offsets in long EEG recordings. We introduce a neural network-based model called *ScoreNet* as a post-processing technique to determine the seizure onsets and offsets in EEGs. A cost function called a *log-dice loss* that has an analogous meaning to $F_1$ is proposed to handle an imbalanced data problem. In combination with several classifiers including random forest, CNN, and logistic regression, the ScoreNet is then verified on the CHB-MIT Scalp EEG database. As a result, in seizure detection, the ScoreNet can significantly improve $F_1$ to 70.15% and can considerably reduce false positive rate per hour to 0.05 on average. In addition, we propose detection delay metric, an *effective latency index* as a summation of the exponential of delays, that includes undetected events into account. The index can provide a better insight into onset and offset detection than conventional time-based metrics.

*Keywords:* Seizure onset and offset detection, EEG, ScoreNet, deep learning, EL-index

*Email addresses:* poomipatosk129@gmail.com (Poomipat Boonyakitanont), apiwat.l@chula.ac.th (Apiwat Lek-uthai), jitkomut.s@chula.ac.th (Jitkomut Songsiri)

[1] Corresponding author

## 1. Introduction

An epileptic seizure is defined by a transient event of a condition from an abnormal electrical activity in the brain [FAA+14]. Recently, epilepsy has affected approximately 65 million people around the world [TBB+11]. In the clinic, neurologists need to review long scalp EEGs to identify the seizure characteristics. However, the process of EEG visual examination is time-consuming, and the diagnosis can be inconsistent due to the fatigue of the neurologists. We, therefore, aim to develop an automated system to detect epileptic seizure events and to label the times of the events in the EEG signals.

Currently, many studies have been developed based on AI and machine learning to automatically detect epileptic seizures in EEG records [ASS+13, AFS+15, AESAA15, BLuCS20, HSZ20]. Some studies used features relevant to amplitude of EEGs such as aEEG [SLuC15] and energy [SG10, ATE10] to indicate ictal patterns. Statistical features were also extracted to distinguish ictal and normal patterns [SKG15, LCZ17]. Many studies made use of discriminative features involved with entropy [TYK16, LKY+18, GP19] and ictal patterns [KKP15, JB17, LCH+19] in the seizure detection. In addition, a combination of features was also considered to jointly distinguish ictal patterns and normal activities [MZCC17, AKS18, VI17, FHH+16]. Due to the current interest of deep learning, many deep learning models have been recently exploited to implicitly extract latent features and to classify seizure episodes [SSGMDEP19, SGK+20, TES+20, SGA+21]. Many studies mainly focused on designs and choices of deep learning architectures, for example, a long short-term memory model (LSTM) [HYX+20] and a convolutional neural network (CNN) [SSY+20], suitable for indicating seizures [AOH+18, TDY+19, HYX+20, OLBT20]. On the other hand, some work attempted to transform EEG segments into a specific representation of deep learning model input *e.g.* an EEG plot image [EKM+19] for a VGG16-based CNN and a three-dimensional EEG tensor [CKN+19] for a temporal graph CNN exploiting connections between each channel.

From the literature, we observed that results from epoch-based seizure de-

tection are not realistically suitable for inferring the seizure onset and offset even though those methods achieved promising results. First, false negatives that unfortunately occur in the middle of true positives lead to wrong inferences of seizure events. In this case, a single seizure is incorrectly predicted as many consecutive seizures due to the false negatives. Second, several isolated false positives cause too frequent false alarms. Both prediction outcomes are not clinically realistic because there is no abrupt change in EEG patterns during a seizure event. In addition, most of the previous studies intentionally selected data in their experiments to reduce imbalance in the data [OCDL16, TES+20]. However, inclusion and exclusion criteria were not given with exact details; it is not clear how well their methods perform on the excluded data set. It is evident that determining the seizure onset and offset as the first and last epochs of seizures predicted by existing epoch-based seizure classifiers is clinically inappropriate. We, hence, further focus on methods that can indicate the seizure onset and offset so that the seizure predictions are more clinically precise.

Recently, only a few researchers have mainly developed automatic seizure onset and offset detection. Shoeb et al. [SKS+11] proposed a method for detecting the ending of a seizure episode by a linear kernel support vector machine (linear SVM) using energies extracted from specific frequency bands of EEG epochs. However, this method failed to accurately determine the seizure offset when the seizure gradually changed, and it needed a powerful seizure onset detector since the seizure onset was required to be early identified. Orosco et al. [OCDL16] applied an artificial neural network (ANN) and linear discriminant analysis (LDA) with energy-based features calculated from stationary wavelet transform to determine seizures in EEG segments, and the seizure onset and offset were identified by the first and last positive predictions during actual seizures. With discriminative features selected by a feature selection method called Lambda of Wilks, LDA outperformed ANN in the seizure detection. Nevertheless, according to the reported performance metrics, there is no evidence that the proposed method could determine the duration of seizures accurately, so it is possible that the seizure episodes were correctly detected by the proposed method with only

short periods. Recently, a CNN model was designed to spatially and temporally capture ictal patterns in EEG epochs [BLuS20]. An additional post-processing method designed from clinical characteristics of seizures was added to recorrect false positives and false negatives in the prior classification. As a result, the additional post-processing could significantly improved $F_1$ in almost all cases. This model, yet, was rule-based; the model parameters were manually tuned to obtain the highest overall $F_1$. Since the characteristics of seizures are different across patients, the model parameters should be specific to each patient. It is time-consuming to specify the parameters for every incoming patient, so a self-learning scheme of customizing these parameters is needed.

To fill these gaps, we aim to design an automatic epileptic seizure onset and offset detection using multi-channel scalp EEG signals. The detection system is divided into two sub-tasks: epoch-based classification and onset and offset detection. The epoch-based classifier identifies seizures in small EEG segments independently, and the onset and offset detector estimates the starting and ending of the seizure event. Several existing models including logistic regression, SVM, and decision tree can be applied in the first stage. In the onset and offset detection, we establish a novel post-processing model called *ScoreNet* extending the work in [BLuS20]. The ScoreNet is a neural network-based model that automatically determines a group of seizure candidates from inputs, and then assigns a score to each individual candidate. Scores in each candidate group are accumulated to determine the possibility that the candidate group is regarded as a seizure episode. In addition, EEG seizure data are highly imbalanced, and outcomes of the detection method tend to be biased towards a normal class. We address this by establishing a cost function called *log-dice loss* based on a dice similarity coefficient. In addition, mean latencies used in [OCDL16, CUFK19] may be misleading for interpreting the seizure onset and offset detection results. In [OCDL16], positive and negative latencies were defined as early and late predictions of seizure onset/offset, respectively, and the mean latency was used as a performance index of delay detection. Cancellations in the calculation of the mean latencies possibly occur, making the average values low even though their

ranges could be particularly high. According to this misinterpretation, we introduce a new metric of time delay called an *effective latency index (EL-index)* that takes an undetected seizure onset/offset into consideration. Empirical results demonstrate that the ScoreNet can dramatically reduce false positives and false negatives, and indicate the seizure onset and offset in a long EEG more precisely. Additionally, it will be shown in the experiment that using the log-dice loss helps overcome the class-imbalanced problem, and the EL-index provides more meaningful interpretation than a latency.

In summary, the contributions of this article are:

1. A neural network-based seizure onset and offset detector called *ScoreNet* and a loss function named *log-dice loss*,

2. A metric of time delay termed *effective latency index*.

This article is organized as follows. A process of the seizure onset and offset detection is presented in Section 2, and the ScoreNet is explained in Section 3. Section 4 provides the problem formulation including the proposed loss function. In Section 5, the description of the EL-index is given. Furthermore, Section 6 demonstrates all experiments conducted to verify the ScoreNet. Finally, the results of seizure classification and seizure onset and offset determination are discussed with graphical illustrations in Section 7.

## 2. Problem Statement

This research aims to detect seizure episodes in long scalp EEGs and to determine the onsets and offsets of the seizures. Consider a process of detecting the seizure onset and offset in a long EEG signal shown in Figure 1. The process is divided into two independently sequential steps, namely, the classification of epileptic seizure epochs, and the detection of the seizure onset and offset. Firstly, long multi-channel EEG signals are segmented into small multi-channel EEG epochs, and the epochs are consequently used in the seizure classification. The classifier receives an input $x_i$ and produces an output $z_i$ for the epoch $i$. In this

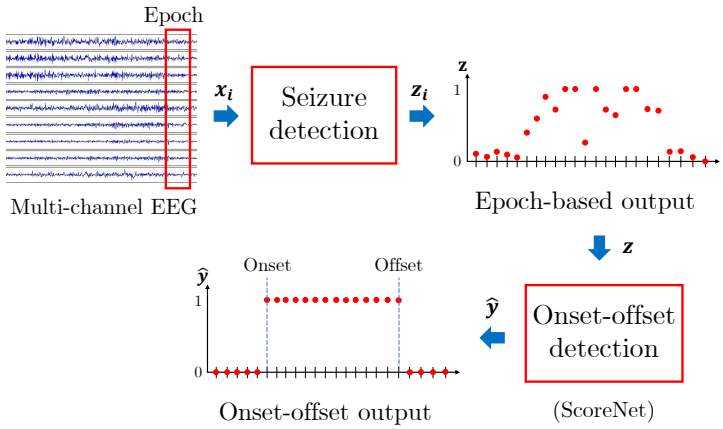



Figure 1: A scheme for determining the seizure onset and offset in long EEG signal. In this work, the onset and offset detection using ScoreNet is mainly focused

case, $x_i$ can be a multi-channel EEG epoch or a feature vector, and $z_i$ is either a seizure probability, *e.g.*, $z_i \in [0, 1]$, or a predicted value, *e.g.*, $z_i \in \{0, 1\}$. A collection of the predicted results of $N$ epochs $z = (z_1, z_2, \ldots, z_N)$ are then jointly analyzed to determine the beginnings and endings of a seizure activities in EEG signals. The ScoreNet converts the prediction sequence $z$ into the final sequence $\hat{y}$ as a screened prediction of seizures for inferring the seizure onset and offset. Finally, the onset and offset are indicated by the first and last indices of each predicted seizure group in $\hat{y}$, respectively.

## 3. ScoreNet

*ScoreNet* is a post-processing technique that takes a connection of adjacent epoch-based classification results into consideration. The ScoreNet chooses a seizure candidate and gives a score to each one. Adjacent seizure candidates are then collected into a group. The total score of the candidates in the group determines the possibility that the candidate group is a seizure. We first describe a counting-based method [BLuS20] in a general form of convolution. This form is then generalized to be the ScoreNet by allowing its parameters to be optimally tuned using an optimization algorithm such as a conjugate gradient method.

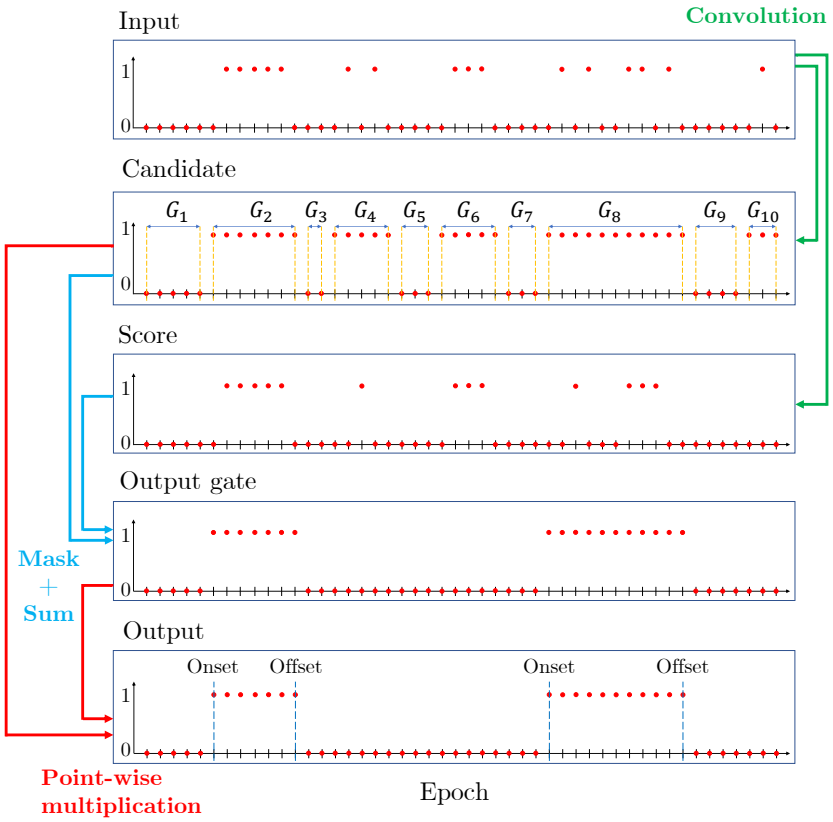



Figure 2: Illustration of counting-based approach [BLuS20] in general form of convolution when $w = 1$ and $p = 4$. Only two groups, *i.e.*, $G_2$ and $G_8$, having scores higher than $p$, are finally predicted events.

The counting-based approach in [BLuS20] first considered epoch-based prediction outcomes as inputs, and the inputs were either 0 or 1. In the neighborhood of the epoch $i$ having the width $2w + 1$ epochs, if there was at least one predicted positive, a *seizure candidate* $c_i$ was counted as one. All adjacent candidates were then gathered into groups. At the $i$-th input epoch, a *positive prediction pair* was a combination of any two predicted positives in the neighborhood. A candidate group was regarded as a predicted seizure episode (output) if the input epochs corresponding to the candidate group had at least

$p$ pairs of positive predictions. These processes are visualized in Figure 2 where $G_2, G_4, G_6, G_8$, and $G_{10}$ are groups of seizure candidates, and $G_2$ and $G_8$ are predicted seizure events.

We now formalize the above description in [BLuS20] in an analytic form as follows. The seizure candidate can be expressed by applying a filter of 1's with length $2w+1$ followed by setting one as a threshold.

$$c_i = \Theta\left(\mathbf{z}_i^T \mathbf{1} - 1\right), \quad i = 1, 2, \ldots, N, \tag{1}$$

where $\Theta$ is a Heaviside step function, $\mathbf{1}$ indicates the vector of ones with a compatible size, $\mathbf{z}_i = (z_{i+w}, \ldots, z_i, \ldots, z_{i-w})$, and $z_j = 0$ for $j \leq 0$ and $j > N$. A group of seizure candidates is then formed by adjacent seizure candidates that have the same value as shown in Figure 2. Unlike the above description that $c_i = 1$ was only used for grouping, $c_i = 0$ and $c_i = 1$ are both utilized to form groups of seizure candidates. Subsequently, a *score* $s_i$ indicates the existence of a positive prediction pair at the $i$-th input epoch. Similar to (1), $s_i$ can be mathematically expressed as

$$s_i = \Theta\left(\mathbf{z}_i^T \mathbf{1} - 2\right), \quad i = 1, 2, \ldots, N.$$

The value of two as the bias term refers to the existence of the predicted pair. We then define an *output gate* $o_i$ which automatically allows a seizure candidate to be an epoch of a seizure when the total score of the group is higher than $p$:

$$o_i = \Theta\left(\sum_{j \in G_l} s_j - p\right), \quad \forall i \in G_l,$$

when $G_l$ denotes the set of indices of the group $l$, and $p$ is the required number of adjacent prediction pairs. Furthermore, for any $i, k \in G_l$, we always have $o_i = o_k$ because the scores are from the same candidate group. Hence, we denote an output gate of the group $l$ as

$$\tilde{o}_l = \Theta\left(\sum_{j \in G_l} s_j - p\right), \quad l = 1, 2, \ldots, m, \tag{2}$$

when $m$ is a number of candidate groups. Intuitively, $\tilde{o}_l$ can be interpreted similarly to an output gate in LSTM; $\tilde{o}_l$ indicates whether a candidate group $l$ is chosen. Finally, for $i \in G_l$ and $l = 1, \ldots, m$, the output $\hat{y}_i$ is

$$\hat{y}_i = c_i \tilde{o}_l. \tag{3}$$

The output $\hat{y}_i = 1$ when the epoch has at least one adjacent detected epoch ($c_i = 1$), and the number of prediction pairs meets the requirement ($\tilde{o}_l = 1$).

### *3.2. ScoreNet architecture*

To add more flexibility to the scheme presented in Section 3.1, we generalize the term $z_i^T \mathbf{1} - 1$ to $z_i^T a + b$ since different epochs may contribute significance differently. With this expression, $a$ and $b$ are allowed to vary upon data. Furthermore, the Heaviside step function can be changed to any differentiable function with a similar meaning. In this case, a sigmoid function ($\sigma$) is used to determine the candidate $c_i$. On the other hand, the hyperbolic tangent function (tanh) is applied to calculate the score $s_i$ because its larger output range helps distinguish a group of candidates. Therefore, we propose the formulae of the candidate and score as

$$c_i = \sigma\left(\mathbf{z}_i^T a_1 + b_1\right), \quad i = 1, 2, \ldots, N, \tag{4}$$

and

$$s_i = \tanh\left(\mathbf{z}_i^T a_2 + b_2\right), \quad i = 1, 2, \ldots, N, \tag{5}$$

where $a_1, a_2, b_1$, and $b_2$ are model parameters. In this case, candidates are then separated into groups using a threshold $\gamma$. Since the sigmoid function is a one-to-one function, the bias $b_1$ is uniquely adapted for a chosen $\gamma$ during training. Hence, $\gamma$ can be arbitrarily set to any real value in $(0, 1)$; here, we set $\gamma = 0.5$. Moreover, we use an average of the group scores in the calculation of the output gate $\tilde{o}_i$ of the group $l$ in (2) instead of the summation to eliminate the effect of the candidate group size. From (2), adding parameters $a_3$ and $b_3$ as scaling and bias terms, the output gate of the group $l$ is proposed as

$$\tilde{o}_l = \sigma\left(\frac{a_3}{N_l} \sum_{j \in G_l} s_j + b_3\right), \tag{6}$$

where $N_l$ is the group size. Finally, the output form in (3) is also modified because, in some cases, $\hat{y}_i$ could be low even though $c_i$ and $\tilde{o}_l$ are both sufficiently high. For instance, when $c_i = 0.7$ and $\tilde{o}_l = 0.7$, this epoch should have been classified as abnormal, but it is decided to be normal since $c_i\tilde{o}_l$ is 0.49, which is normally classified as normal. Similar to (6), we add a weight $a_4$ and a bias $b_4$ to scale the value of $c_i\tilde{o}_l$. For any $i \in G_l$, the output $\hat{y}_i$ is given by

$$\hat{y}_i = \sigma\left(a_4 c_i \tilde{o}_l + b_4\right). \tag{7}$$

As a result, we have proposed a more general model, described by (4) - (7), of seizure onset and offset detection. The model parameters are $a_1, b_1, a_2, b_2, a_3, b_3, a_4$, and $b_4$ and can be chosen to minimize a loss function of $(y, \hat{y})$.

## 4. Problem Formulation and Optimization

Let $y = (y_1, y_2, \ldots, y_N)$ be a target, and $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ be an output of the ScoreNet, respectively. The formulation of this problem is to minimize a loss function

$$\text{minimize} \quad \mathcal{L}\left(y, \hat{y}\right) \tag{8}$$

over the ScoreNet parameters. In a binary classification problem, many loss functions indicate a similarity between the output $\hat{y}_i$ and the label $y$. Here, we propose a cost function called *log-dice loss* that is motivated by a dice similarity coefficient. Moreover, three existing loss functions for the classification, namely, binary cross-entropy, soft-dice loss [LSF+18], and squared-dice loss [MNA16] are discussed. Note that both dice losses were previously called *soft-dice loss*; here, we call them differently to avoid confusion.

### *4.1. Binary cross-entropy*

The binary cross-entropy is a common loss in classification problem which is derived from the log-likelihood function of the Bernoulli distribution:

$$\mathcal{L}_{\text{ent}}\left(y, \hat{y}\right) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right]. \tag{9}$$

We observe that $\sum_i (1-y_i)\log(1-\hat{y}_i)$ and $\sum_i y_i \log \hat{y}_i$ intuitively reflect penalties to false positives and false negatives. This means that minimizing the cross-entropy is equivalent to maximizing classification accuracy. However, this loss penalizes all incorrect samples equally; a majority class does provide more contribution than a minority class when the cross-entropy is employed, so this loss is typically not suitable when the data are highly imbalanced.

### *4.2. Soft-dice loss*

A dice similarity coefficient (DSC) is a measure for determining an analogy of two sets of samples. In a classification problem, the DSC, also known as $\mathrm{F}_1$, that indicates the overlap between true and predicted values, is defined as

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{10}$$

when TP, FP, and FN stand for numbers of true positives, false positives, and false negatives, respectively. The DSC does not directly measure majority samples, *i.e.*, true negative, so this coefficient is an appropriate evaluation metric in imbalanced data classification.

Closedly related to DSC, the soft-dice loss ($\mathcal{L}_{\mathrm{softDL}}$) [LSF+18] is defined as

$$\mathcal{L}_{\mathrm{softDL}}\left(y, \hat{y}\right) = 1 - \frac{2\sum\limits_{i=1}^{N} y_i \hat{y}_i}{\sum\limits_{i=1}^{N} \left(y_i + \hat{y}_i\right)}. \tag{11}$$

If we substitute $\sum_i y_i\hat{y}_i, \sum_i (1-y_i)\hat{y}_i$, and $\sum_i y_i(1-\hat{y}_i)$ for TP, FP, and FN, respectively, in DSC, we obtain the second term of RHS of (11). Therefore, minimizing $\mathcal{L}_{\mathrm{softDL}}$ is similar to maximizing $\mathrm{F}_1$. This loss is conceptually suitable for a problem of imbalanced data.

### *4.3. Squared-dice loss*

The squared-dice loss ($\mathcal{L}_{\mathrm{sqDL}}$) was originally inspired by the DSC for an imbalanced data problem [MNA16]. In contrast to $\mathcal{L}_{\mathrm{softDL}}$, $\mathcal{L}_{\mathrm{sqDL}}$ uses the sum

of squares of the actual and predicted labels in the denominator:

$$\mathcal{L}_{\mathrm{sqDL}}(y, \hat{y}) = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} (y_i^2 + \hat{y}_i^2)} = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i^2 + \hat{y}_i^2)}. \tag{12}$$

We can see that $\mathcal{L}_{\mathrm{sqDL}}$ can be alternatively reformulated as the sum squared error, normalized by the sum of the squares.

### *4.4. Log-dice loss*

We establish the log-dice loss ($\mathcal{L}_{\mathrm{logDL}}$) to overcome an imbalanced data problem by conceptually maximizing $\mathrm{F}_1$. According to a soft-dice loss [LSF+18], penalties of wrongly classifying hard and easy samples, *i.e.,* samples that are incorrectly classified by large and small classification errors respectively, are linearly proportional. This means that costs in incorrectly predicting hard samples and easy samples are slightly distinguished. We then propose the log-dice loss that includes different penalties on the hard and easy samples using the dissimilarity of $y_i$ and $\hat{y}_i$ in the log-scale.

For this purpose with the aim of conceptually maximizing $\mathrm{F}_1$, the terms TP, FP, and FN in (10) are substituted by $\sum_i y_i \log(1-\hat{y}_i)$, $\sum_i (1-y_i)\log(1-\hat{y}_i)$, and $\sum_i y_i \log \hat{y}_i$, respectively, as the proposed $\mathcal{L}_{\mathrm{logDL}}$. Those three terms are large when $y_i$ and $\hat{y}_i$ agree with their original definitions, and small otherwise. For instance, when $y_i = 1$ and $\hat{y}_i \approx 1$, which is a true positive case, $y_i \log(1-y_i)$ approaches infinity, $(1-y_i)\log(1-\hat{y}_i) = 0$, and $y_i \log \hat{y}_i \approx 0$. With these definitions, $\mathcal{L}_{\mathrm{logDL}}$ is expressed as

$$\mathcal{L}_{\mathrm{logDL}}(y, \hat{y}) = 1 - \frac{2\sum_{i=1}^{N} y_i \log(1-\hat{y}_i)}{\sum_{i=1}^{N} \left[(1+y_i)\log(1-\hat{y}_i) + y_i \log \hat{y}_i\right]} \tag{13}$$

The value of $\mathcal{L}_{\mathrm{logDL}}$ is in the range of $(0, 1]$; $\mathcal{L}_{\mathrm{logDL}}(y, \hat{y}) = 1$ when $y = 0$. This loss decreases when $y$ and $\hat{y}$ are similar, and increases when $y$ and $\hat{y}$ are different.

## 5. Effective Latency index (EL-index)

An *effective latency index (EL-index)* is an indicator for measuring delays of correct event detection while undetected events are also taken into account. The motivation is to give a positive score to any event which is correctly detected, and zero for any undetected event. This score should be low when the delay of the detected event is high, and vice versa. Normally, a latency is the time difference between a predicted event and an actual event. There is currently no consensus of latency determination when situations such as late or early detection and multiple predictions happen. We denote $d_i > 0$ and $d_i < 0$ the latencies of early and late detection of seizure onset and offset, respectively, and there are $n$ seizure activities. Note that $d_i$ is not defined for an undetected seizure event. Moreover, let $k_i$ be an indicator of the event $i$ being detected: $k_i = 1$ when the event is correctly detected, and $k_i = 0$ otherwise. With these notations, the *mean latency* is given by $\bar{d} = \left(\sum_{i=1}^{n} k_i d_i\right) \Big/ \left(\sum_{i=1}^{n} k_i\right)$. Moreover, the *mean absolute latency* is similarly expressed as $\hat{d} = \left(\sum_{i=1}^{n} k_i |d_i|\right) \Big/ \left(\sum_{i=1}^{n} k_i\right)$.

In general, $\bar{d}$ and $\hat{d}$ are particularly low when delays are small, which commonly refer to high precision in the detection. However, both of them ignore all undetected events from the calculation, leading to misunderstanding when their values are promising. For instance, they are low if a few events are almost perfectly detected while the other events are all overlooked. Moreover, $\bar{d}$ is possibly low because of the cancellation of positive and negative latency. Therefore, we propose the EL-index to solve these problems. The EL-index is defined as

$$\text{EL-index} = \frac{1}{n} \sum_{i=1}^{n} k_i r^{|d_i|}, \tag{14}$$

where $0 < r < 1$ and $n$ is the number of true seizure events. This decay rate $r$ can be arbitrarily specified by the user according to an impact degree of the delay in each application. For example, in the case of detecting life-threatening abnormalities in the ECG, $r$ should be low because the delay impact is considerable.

The EL-index values range from zero (missing all events) to one (perfectly detecting all events). Given a correct detection ($k_i = 1$), the value $r^{|d_i|}$ exponentially decreases when $|d_i|$ increases, meaning that a higher latency dramatically causes a smaller EL-index. On the other hand, when an event is undetected ($k_i = 0$), the term $r^{|d_i|}$ is ignored from the EL-index. This is similar to assigning $d_i = \infty$ for the undetected event, making $r^{|d_i|}$ approaches zero. Hence, undetected events always provide the least contribution to the EL-index.

If we denote GDR (a good detection rate) a portion of correctly detected seizure events in one record and hence given by $(1/n)\sum_{i=1}^{n} k_i$, we can see that the EL-index given in (14) can be regarded as an exponentially weighted GDR. Moreover, an upper limit and a lower limit of the EL-index can be derived as

$$\text{GDR} \cdot r^{|d|_{\max}} \leq \text{EL-index} \leq \text{GDR}, \tag{15}$$

where $|d|_{\max}$ is the maximum value of absolute time delays. It is evident that the EL-index cannot be higher than GDR, and it is bounded below by the function $\text{GDR} \cdot r^{|d|}$. Moreover, the EL-index can distinguish distributions of collected time delays. Suppose we have two cases of time delay samples from test results. The distributions of these two cases are narrowly spread and highly varied, but the samples of two cases have the same mean absolute latency $\hat{d}$ and the same GDR. When all time delays are similar, *i.e.*, $d_i \approx \pm\hat{d}$, for any detected event $i$, we can derive that the EL-index is close to the bound

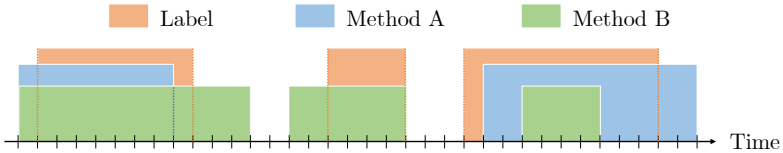
$$\text{EL-index} \approx \text{GDR} \cdot r^{\hat{d}}. \tag{16}$$

In contrast, when the time delays are highly varied, the EL-index is far from the bound (16), and the EL-index is always higher than that of the first case:

$$\text{GDR} \cdot r^{\hat{d}} \leq \text{GDR} \cdot \frac{\sum_{i=1}^{n} k_i r^{|d_i|}}{\sum_{i=1}^{n} k_i} = \frac{1}{n}\sum_{i=1}^{n} k_i r^{|d_i|}.$$

Figure 3 shows comparisons of these three metrics when $r = 0.9$. It is obvious that the mean latency is not an appropriate metric because the mean

latency of onset latencies from both methods are zero, but both methods do not perfectly detect seizure onsets. Furthermore, method A obtains less mean absolute latency of onset than method B but method A cannot detect the second seizure event ($k_2 = 0$). The EL-index of detecting seizure onset of method B is, therefore, higher than that of method A. Moreover, the seizure onsets are more accurately detected more than the seizure offsets when method A is considered; thus, the EL-index of detecting the onset is higher. In addition, for method B, latencies of detecting the event terminations are relatively diverse compared to those of detecting the starting points, so the EL-index of detecting offset is higher than that of detecting onset given the same mean absolute latency.




| | | Mean latency | Mean absolute latency | EL-index |
|---|---|---|---|---|
| Onset | Method A | 0.00 | 1.00 | 0.60 |
| | Method B | 0.00 | 2.00 | 0.81 |
| Offset | Method A | 0.50 | 1.50 | 0.57 |
| | Method B | 0.00 | 2.00 | 0.82 |

Figure 3: Example of time-based measurements. Mean of onset latencies from methods A and B are zero, but both methods do not perfectly detect seizure onsets. Method A obtains a lower mean absolute latency of onset than method B, but the EL-indices of detecting the seizure onset and offset of method B are higher than that of method A because the method A does not detect the second seizure.

## 6. Experiment

We conducted experiments to ensure that the ScoreNet can provide reasonable outcomes, and significantly improve epoch-based detection performances. The experiments were performed on *all* records in the CHB-MIT Scalp EEG database [GAG+00]. Many commonly used classifiers such as logistic regression, SVM, random forest, and deep learning models were applied to determine seizure episodes. Moreover, binary cross-entropy, soft-dice loss [LSF+18], and

square-dice loss [MNA16] were also used to compare their effectiveness in dealing with an imbalanced data problem. Finally, we evaluated the methods using three types of measurement: event-based metrics, epoch-based metrics, and time-based metrics because these metrics indicate the detection performances in different aspects. Codes of running the ScoreNet and experimental results are available at https://github.com/Siyaosk129/ScoreNet.

### *6.1. Data preparation*

The CHB-MIT Scalp EEG database is a collection of EEG signals observed from 24 pediatric cases at the Children's Hospital Boston [GAG+00, Sho09]. The data were sampled at the sampling rate of 256 Hz with a resolution of 16 bits, and they were recorded with the international 10-20 electrode system. Most records consist of 23 channels collected from either referential or bipolar montages. See more details of the number of records in Appendix B. This data set is available on PhysioNet (https://physionet.org/content/chbmit/1.0.0/).

In this research, we used *all records* in this database. Since both referential and bipolar montages were employed to record the data, the EEG records were firstly rearranged so that the montages were bipolar. The sequential order of the modified 18 channels were *FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-T3, T3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, P8-O2, FZ-CZ*, and *CZ-PZ*. The long EEG records were segmented into non-overlapped epochs of one-second period.

### *6.2. Evaluation*

In this study, seizure detection methods were evaluated using a patient-specific leave-one-record-out cross validation (LOOCV) scheme. Training and test data were taken from the same case but different records. Results from all validations were used for the assessment using event-based, epoch-based, and time-based metrics. The event-based metrics quantify correctness of detecting a seizure occurrence in an EEG recording, whereas epoch-based metrics consider

an EEG epoch as a sample [TTM+11, BLuCS20]. In addition, the time-based metrics indicate precision of identifying the onset and offset. Finally, all metrics from all validations were used to present the overall detection performances.

In this research, we considered both good detection rate (GDR) and false positive rate per hour (FPR/h) as the event-based metrics. GDR is defined by a portion of correctly detected seizure events in one record. On the other hand, FPR/h that reveals incorrect seizure prediction is defined as a number of false positive events in one hour. To indicate the performance of imbalanced-class classification problem, we mainly focused on $F_1$ [Pow11]. To assess the performance of seizure onset and offset detection, we illustrate the usage of the EL-index explained in Section 5 and compare it with other latency indices. If more than one positive event occurred during one actual event, the onset of the *first* predicted event and the offset of the *last* positive activity were used to calculate the time-based metrics. In addition, a time gap between any two actual seizures is usually substantial, so it is unlikely that our predicted seizure could overlap with two actual events.

### *6.3. Experimental Setup*

In the experiments, five classifiers, namely, CNN, logistic regression, linear SVM, decision tree, and random forest were applied to detect seizure epochs. In this work, we exploited the CNN structure in [BLuS20]. The CNN input was a raw EEG segment since meaningful features can be implicitly extracted by the model while the other classifiers require hand-crafted features to characterize ictal and normal patterns. Selected features in this experiment have been widely used and able to discriminate epileptic event from normal epochs [BLuCS20, SG10, AFS+15, ASS+13]. The *time-domain features* computed from a raw EEG epoch were variance, energy, nonlinear energy, Shannon entropy, sample entropy, and approximate entropy. The *frequency-domain features* calculated from the power spectral densities were energies from eight sub-bands ranging 0 - 25 Hz. In addition, discrete wavelet transform coefficients from five decomposition levels with the Daubechies 4 tap wavelet were used to extract the *time-frequency-*

*domain features*, and the features were mean absolute value, variance, energy, maximum, minimum, and line length. Each feature was extracted from each channel in an EEG epoch, and was normalized to a z-score. Totally, 900 features were combined into a feature vector for the classification of EEG epoch. In addition, we used a grid search for tuning hyperparameters of the other epoch-based classifiers. The hyperparameters that yielded at least one true positive in every patient and achieved high overall $F_1$ from test records were selected. For ScoreNet, we set the size of vectors $a_1$ and $a_2$ as 13 ($w = 6$) in order to compare the performances with the best counting-based method in [BLuS20].

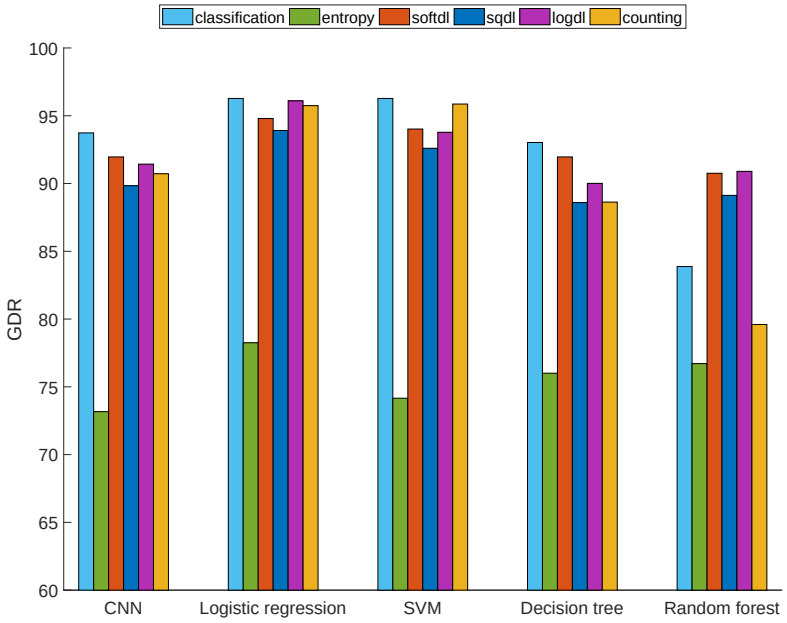
## 7. Result and Discussion

First, epoch-based classification results performed by CNN, logistic regression, linear SVM, decision tree, and random forest were reported and labeled by `classification`. After the ScoreNet was applied with different loss functions *i.e.*, entropy, soft-dice loss, squared-dice loss, and log-dice loss, the performances are compared with those from the prior classification step, and are labeled by `entropy, softdl, sqdl, logdl` in Figures 4 and 5. When the counting-based method was applied, we refer to this result as `counting`.

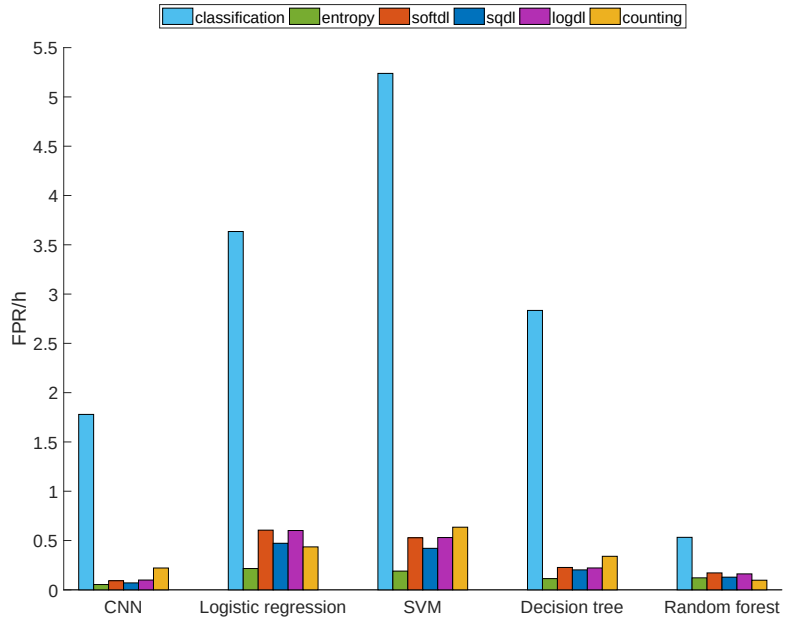
### *7.1. Seizure detection*

Figures 4 and 5 show the performances of applying the ScoreNet to improve seizure detection results from using only the classifiers in terms of $F_1$, GDR and FPR/h, respectively. In overall, the seizure classifiers in the prior step detected seizure events with GDR of more than 80%, obtained $F_1$ of less than 40%, but FPR varied drastically from 0.53 per hour to 5.24 per hour on average upon the classifiers. It was shown that the random forest and CNN can be compromised classifiers if we consider all $F_1$, GDR and FPR/h as metrics but their wrong inference results of seizure events could have been improved.

After applying the ScoreNet, regardless of the classification methods, $F_1$ scores increased at least 18%, FPRs were significantly reduced at least 0.36




(a) GDR.




(b) FPR/h.

Figure 4: Comparisons of averaged GDR and FPR/h obtained from test cases using different epoch-based seizure detection methods.

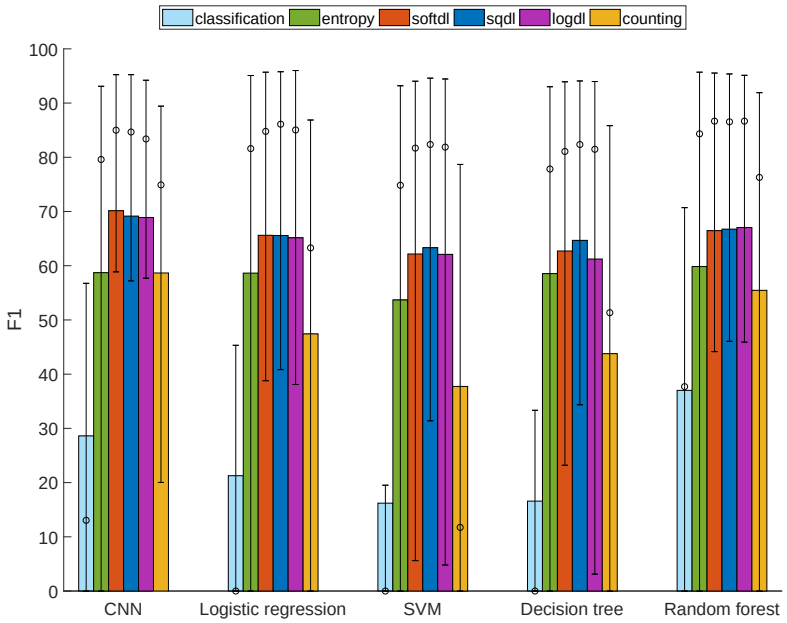



Figure 5: Comparisons of $F_1$ obtained from test cases using different epoch-based seizure detection methods. Color bars indicate the average values of all test records, the circle markers present the median, and the vertical bars show the interquartiles.

times per hour but GDRs were slightly dropped. On the other hand, GDR obtained from the case of the random forest increased up to 7%. As shown in Figure 6, we found that the random forest can produce a sequence of small seizure probabilities distinguishable from the background. In this case, the ScoreNet could identify the seizure from the seizure probabilities. This means that using the ScoreNet generally improves epoch-based classification performances. Specifically, a combination of the ScoreNet and any dice loss function can also indicate some seizures from events that are not originally detected in the classification process, and the dice loss functions are, thus, appropriate for handling an imbalanced-class data problem.

From the reports of $F_1$, GDR, and FPR/h, we can generally interpret that some isolated predictions, *i.e.*, true positives, false positives, and false negatives, are eliminated by using the proposed seizure onset and offset determination methods, and the result of eliminating these predictions depends on the em-

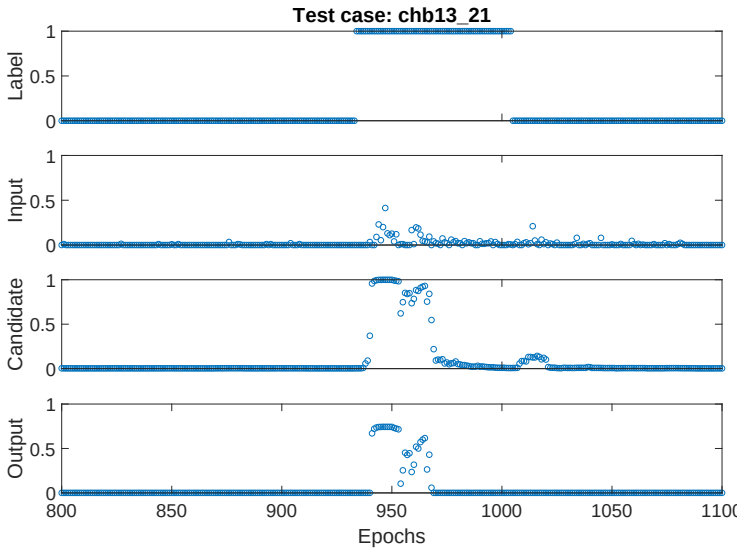



Figure 6: Results of the ScoreNet when the log-dice loss and the random forest are used, and the test case is `chb13_21`.

ployed cost function. According to a large reduction of FPR/h, this indicates that using the cross-entropy with the ScoreNet potentially reduces several isolated false positives. However, since using the cross-entropy has a bias towards the normal class, and only a few epochs are correctly classified as seizures by epoch-based classifiers, the number of predicted positives is not sufficiently high and hence, those predicted positives are eliminated, resulting in a decrease in the true positives. In the case of the dice loss functions, the ScoreNet generally yields similar results of seizure detection performance across any seizure detection methods. Specifically, a combination of the ScoreNet and the soft-dice loss generally provides the best results of detecting seizure activities, and the ScoreNet with the squared-dice loss can better reduce FPR/h. Since only a few segments during some seizure episodes were detected, the model recognized them as artifacts. As a result, GDR also was relatively lower compared to that of the other loss functions. In addition, using the log-dice loss can better improve the classification performance when classification errors are large.

For instance, the results of using random forest and log-dice loss are illustrated in Figure 6. Seizure probabilities from the random forest (input) during the actual seizure are not sufficiently high to be detected, but they are visually distinguishable from the background. We can see that these predictions are then jointly boosted by the ScoreNet so that they are potential enough to represent the seizure episode. GDR and $F_1$ are, thus, improved the most when using the log-dice loss in the case of random forest.

Table 1: Comparison of seizure detection methods using CHB-MIT database.

| Study | Detection type | Data specification | Validation | Method | Acc | Sen | Spec | $F_1$ | GDR | FPR/h |
|---|---|---|---|---|---|---|---|---|---|---|
| [SLuC15] | Event | Long seizures | No CV | aEEG + Adaptive threshold | NR | NR | NR | NR | 88.50 | 0.18 |
| [YXJZ17] | Event | NR | NR | Spectrogram + mSSDA | 93.82 | NR | NR | 96.05 | NR | NR |
| [VI17] | Event | Specific records (22 cases) | LOOCV | Energy and fractal dimension + RVM | NR | NR | NR | NR | 97.00 | 0.10 |
| [GGC+20] | Event | 166 mins (11 cases) 70% training data | No CV | PSD + CNN-based ensemble | 92.60 | 92.30 | 97.00 | NR | NR | NR |
| [HYX+20] | Event | All Balanced training data | No CV | Statistical features + Bi-LSTM | 92.66 | 93.61 | 91.85 | NR | NR | NR |
| [SG10]* | Onset | NR | LOOCV | Energy + RBF SVM | NR | NR | NR | NR | 96.00 | 0.08 |
| [TLLP20] | Onset | Seizure in record | LOOCV | Unified multi-level spectral-temporal feature + RBF SVM | 97.80 | NR | NR | 78.00 | 97.20 | 0.64 |
| [OCDL16] | Onset/offset | Bipolar montage (18 cases) Balanced training data | 10-fold | Relative band energy + LDA | NR | NR | 99.99 | NR | 92.60 | 0.30 |
| [CUFK19] | Onset/offset | 397 hrs (18 cases) 60% training data | 5-fold | Statistical features + LDA | 98.00 | NR | 98.05 | NR | 100.00 | 4.02 |
| [BLuS20] | Onset/offset | All | LOOCV | CNN + counting-based method | 99.69 | 78.18 | 99.77 | 58.66 | 90.72 | 0.22 |
| **Proposed method** | Onset/offset | All | LOOCV | CNN+ ScoreNet (soft-dice loss) | 99.83 | 76.54 | 99.92 | 70.15 | 91.96 | 0.09 |

NR = no report, LOOCV = leave-one-record-out cross validation, All = use full data set, * Use median instead of mean to report the performance

Finally, Table 1 summarizes a comparison of seizure detection performance noted with an amount of data and a validation scheme used in each study; the method achieving the best $F_1$ is chosen to compare the performance. It was found that several studies specifically selected records from the database, and a few applied LOOCV as a validation scheme. Work in [TLLP20] used only records containing seizure in the experiment, and neither data specification nor validation scheme was reported in [YXJZ17]. Moreover, some studies performed data sampling approaches to create balanced training data [HSZ20]. In practice,

characteristics of EEGs such as rarity and types cannot be accurately selected. It is more clinically challenging to use all data and apply LOOCV to verify the detection performance. Our proposed method, therefore, yielded a competing performance reported in a very fair setting.

### *7.2. Onset and offset determination*

GDR, $|d|$, and the corresponding EL-indices collected from the test results are displayed in Figure 7, for each value of GDR shown in different colors. When GDR $= 0$, we set $d$ to zero for the purpose of visualization in the plot. The test cases of 0% GDR shown in yellow markers imply that there is a portion of undetected events. Hence, if we use only the mean absolute latency index, these detection failures are ignored whereas EL-index can capture this since zero GDR is mapped to zero score of EL-index. As shown in Figure 7, at low GDR (about 40 - 50%), it is possible that seizure events were randomly detected but low mean absolute latency could be obtained whereas the EL-index is dropped (indicating worse performance). So the proposed EL-index is more suitable for being a time-based index than the mean absolute latency.

For the cases of non-zero GDR, the EL-index is dominantly high when the detection delay is insignificant and many focused events are detected. From Figure 7, the relationship of EL-index and mean absolute latency mostly satisfies the exponential bound, GDR $\cdot\, r^{|d|}$ represented in dashed lines. As analyzed in Section 5, this means that latencies from detecting seizure onsets/offsets mostly have low variation. In addition, consider markers above the dashed line for a specific GDR. The mean absolute latency definitely cannot differentiate cases of similar and different onset/offset latencies, whereas the EL-index does. Therefore, the EL-index provides not only a meaning of accurate seizure onset/offset detection but also the accuracy of seizure event detection and the interpretation of latency distributions when considered jointly with GDR.

Figure 8 shows comparisons of EL-indices of detecting seizure onsets and offsets using various methods when $r = 0.9$, and Table 2 concludes EL-indices obtained from all methods. In the case of seizure onset detection, the mean

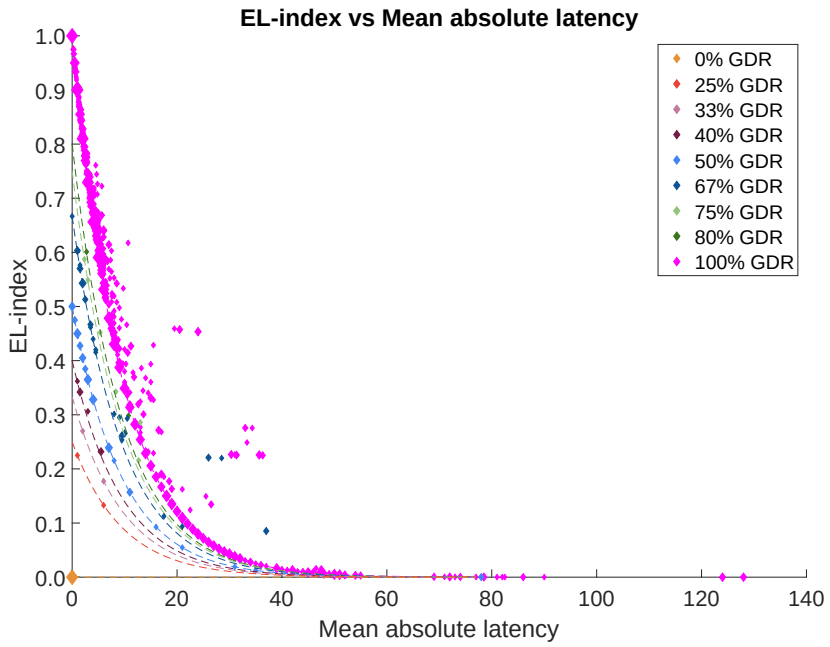
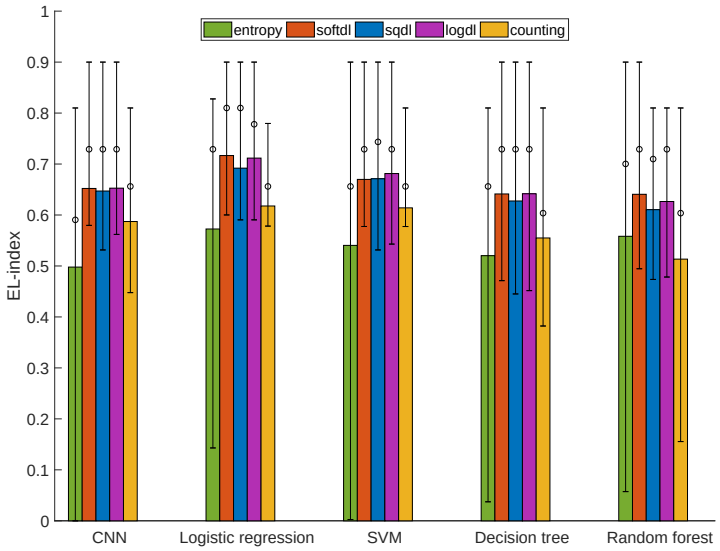
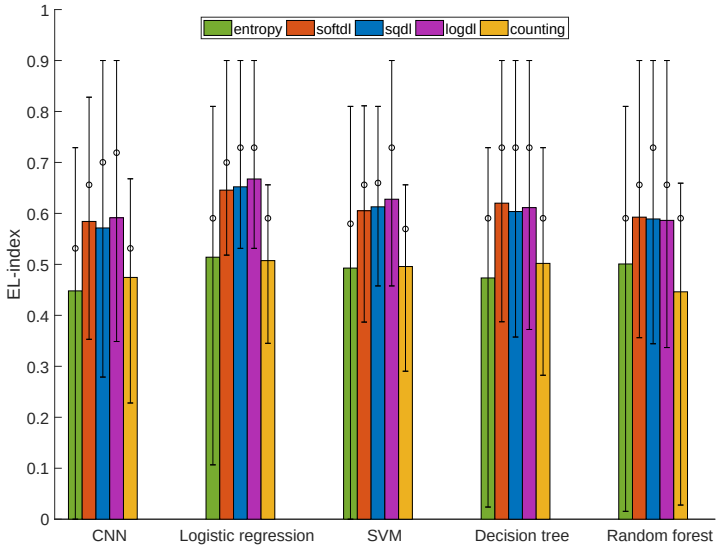



Figure 7: Relation of EL-index and average of absolute latency from test data given $r = 0.9$. The marker size is proportional to the number of samples in a log scale and a dashed line illustrates GDR $\cdot\, r^{|d|}$. We set $d = 0$ when no seizure events are detected (GDR $= 0$) for the purpose of visualization.

values of EL-indices ranged from 0.50 to 0.71, and the medians ranged from 0.59 to 0.81. For detecting seizure offset, means and medians of EL-indices ranged from 0.45 to 0.67 and from 0.53 to 0.73, respectively. As shown in Figure 7 and from the minimum median of 0.53, this implies that seizure onsets and offsets are typically detected with mean absolute errors less than 10 seconds, which is clinically acceptable. In particular, EL-indices of indicating seizure onset and offset using the ScoreNet with dice loss functions were similarly high compared to those of the other methods, and applying the log-dice loss with the ScoreNet achieved slightly better EL-indices of detecting seizure offset among the other dice loss functions. It is evident that all seizure onset and offset detection methods could better indicate seizure onsets than seizure offsets. This is due to epileptic seizure characteristics that ictal patterns establish at the end of an




(a) Seizure onset.




(b) Seizure offset.

Figure 8: Comparisons of EL-indices of detecting seizure onset and offset. Color bars indicate the average values. The circle markers present the median, and the vertical bars show the interquartiles.

Table 2: Summary of mean EL-index of seizure onset and offset determination. The maximum EL-index from each classifier is given in boldface.

| | | CNN | Logistic regression | SVM | Decision tree | Random forest |
|---|---|---|---|---|---|---|
| Onset | entropy | 0.50 | 0.57 | 0.54 | 0.52 | 0.56 |
| | softdl | **0.65** | **0.72** | 0.67 | **0.64** | **0.64** |
| | sqdl | **0.65** | 0.69 | 0.67 | 0.63 | 0.61 |
| | logdl | **0.65** | 0.71 | **0.68** | **0.64** | 0.63 |
| | counting | 0.59 | 0.62 | 0.61 | 0.55 | 0.51 |
| Offset | entropy | 0.45 | 0.51 | 0.49 | 0.47 | 0.50 |
| | softdl | 0.58 | 0.65 | 0.61 | **0.62** | **0.59** |
| | sqdl | 0.57 | 0.65 | 0.61 | 0.60 | **0.59** |
| | logdl | **0.59** | **0.67** | **0.63** | 0.61 | **0.59** |
| | counting | 0.47 | 0.51 | 0.50 | 0.50 | 0.45 |

event less dominantly than at the beginning, so predictions of seizure epochs tend to appear less at the event termination. This also means that the ScoreNet with the log-dice loss can better fix incorrect classification outcomes which have low seizure probabilities at the ending of the event. Therefore, exploiting the log-dice loss is more beneficial when a large margin of classification error occurs.

## 8. Conclusion

This research established an automatic epileptic seizure onset and offset detection scheme composed of two processes: detecting seizures in epochs and determining the beginning and ending of a seizure event. A neural network-based model named ScoreNet was designed to detect epileptic seizure onset and offset from epoch-based classification results by incorporating a log-dice loss function to alleviate the issue of class-imbalanced data inherited in EEG classification. An EL-index was presented to demonstrate the ability of a seizure onset and offset detection method. The proposed scheme was evaluated on the CHB-MIT Scalp EEG database. As a result, the ScoreNet can improve $F_1$ up to 70.15% on epoch-based seizure classification, dramatically reduced false

alarms to 0.05 times per hour, and yielded onset and offset detection errors in a range that is clinically acceptable. The improved performance from the use of log-dice loss was gained the most when prediction errors from the epoch-based seizure detection are large. In addition, the EL-index is suitable for measuring a detection delay and also provides information about the correct detection of events, as well as more details about latency distribution.

## 9. Acknowledgment

This work is financially supported by the 100th Anniversary Chulalongkorn University Fund for Doctoral Scholarship, the 90th Anniversary of Chulalongkorn University Fund (Ratchada-phiseksomphot Endowment Fund), and the Chula Engineering Research grant.

## References

[AESAA15] T. Alotaiby, F.E.A. El-Samie, S.A. Alshebeili, and I. Ahmad. A review of channel selection algorithms for EEG signal processing. *EURASIP Journal on Advances in Signal Processing*, 2015(1):66–86, 2015.

[AFS+15] U.R. Acharya, H. Fujita, V.K. Sudarshan, S. Bhat, and J.E.W. Koh. Application of entropies for automated diagnosis of epilepsy using EEG signals: A review. *Knowledge-Based Systems*, 88:85–96, 2015.

[AKS18] E. Alickovic, J. Kevric, and A. Subasi. Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. *Biomedical Signal Processing and Control*, 39:94–102, 2018.

[AOH+18] U.R. Acharya, S.L. Oh, Y. Hagiwara, J.H. Tan, and H. Adeli. Deep convolutional neural network for the automated detection

and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, 100(1):270–278, 2018.

[ASS+13] U.R. Acharya, S.V. Sree, G. Swapna, R.J. Martis, and J.S. Suri. Automated EEG analysis of epilepsy: A review. *Knowledge-Based Systems*, 45:147–165, 2013.

[ATE10] S. Altunay, Z. Telatar, and O. Erogul. Epileptic EEG detection using the linear prediction error energy. *Expert Systems with Applications*, 37(8):5661–5665, 2010.

[BLuCS20] P. Boonyakitanont, A. Lek-uthai, K. Chomtho, and J. Songsiri. A review of feature extraction and performance evaluation in epileptic seizure detection using EEG. *Biomedical Signal Processing and Control*, 57:101702, 2020.

[BLuS20] P. Boonyakitanont, A. Lek-uthai, and J. Songsiri. Automatic epileptic seizure onset-offset detection based on CNN in scalp EEG. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[CKN+19] I. Covert, B. Krishnan, I. Najm, J. Zhan, M. Shore, J. Hixson, and M. J. Po. Temporal graph convolutional networks for automatic seizure detection. *arXiv preprint arXiv:1905.01375*, 2019.

[CUFK19] G. Chandel, P. Upadhyaya, O. Farooq, and Y.U. Khan. Detection of seizure event and its onset/offset using orthonormal triadic wavelet based features. *IRBM*, 40(2):103–112, 2019.

[EKM+19] A. Emami, N. Kunii, T. Matsuo, T. Shinozaki, K. Kawai, and H. Takahashi. Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images. *NeuroImage: Clinical*, 22:101684, 2019.

[FAA+14] R.S. Fisher, C. Acevedo, A. Arzimanoglou, A. Bogacz, J.H. Cross, C.E. Elger, J. Jr Engel, L. Forsgren, J.A. French, M. Glynn, et al. ILAE official report: a practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482, 2014.

[FHH+16] P. Fergus, A. Hussain, D. Hignett, D. Al-Jumeily, K. Abdel-Aziz, and H. Hamdan. A machine learning system for automated whole-brain seizure detection. *Applied Computing and Informatics*, 12(1):70–89, 2016.

[GAG+00] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C. Peng, and H.E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23):e215–e220, 2000.

[GGC+20] Y. Gao, B. Gao, Q. Chen, J. Liu, and Y. Zhang. Deep convolutional neural network-based epileptic electroencephalogram (EEG) signal classification. *Frontiers in Neurology*, 11, 2020.

[GP19] V. Gupta and R. B. Pachori. Epileptic seizure identification using entropy of FBSE based EEG rhythms. *Biomedical Signal Processing and Control*, 53:101569, 2019.

[HSZ20] A. R. Hassan, A. Subasi, and Y. Zhang. Epilepsy seizure detection using complete ensemble empirical mode decomposition with adaptive noise. *Knowledge-Based Systems*, 191:105333, 2020.

[HYX+20] X. Hu, S. Yuan, F. Xu, Y. Leng, K. Yuan, and Q. Yuan. Scalp EEG classification using deep Bi-LSTM network for seizure detection. *Computers in Biology and Medicine*, 124:103919, 2020.

[JB17] A.K. Jaiswal and H. Banka. Local pattern transformation based feature extraction techniques for classification of epileptic EEG

signals. *Biomedical Signal Processing and Control*, 34:81–92, 2017.

[KKP15] T.S. Kumar, V. Kanhangad, and R.B. Pachori. Classification of seizure and seizure-free EEG signals using local binary patterns. *Biomedical Signal Processing and Control*, 15:33–40, 2015.

[LCH+19] Y. Li, W. Cui, H. Huang, Y. Guo, K. Li, and T. Tan. Epileptic seizure detection in EEG signals using sparse multiscale radial basis function networks and the fisher vector approach. *Knowledge-Based Systems*, 164:96–106, 2019.

[LCZ17] M. Li, W. Chen, and T. Zhang. Classification of epilepsy EEG signals using DWT-based envelope analysis and neural network ensemble. *Biomedical Signal Processing and Control*, 31:357–365, 2017.

[LKY+18] P. Li, C. Karmakar, J. Yearwood, S. Venkatesh, M. Palaniswami, and C. Liu. Detection of epileptic seizure based on entropy analysis of short-term EEG. *PloS ONE*, 13(3):e0193691, 2018.

[LSF+18] R. Lguensat, M. Sun, R. Fablet, P. Tandeo, E. Mason, and G. Chen. EddyNet: A deep neural network for pixel-wise classification of oceanic eddies. In *Proceedings of the 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1764–1767. IEEE, 2018.

[MNA16] F. Milletari, N. Navab, and S.A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[MZCC17] M. Mursalin, Y. Zhang, Y. Chen, and N.V. Chawla. Automated epileptic seizure detection using improved correlation-based fea-

ture selection with random forest classifier. *Neurocomputing*, 241:204–214, 2017.

[OCDL16] L. Orosco, A.G. Correa, P. Diez, and E. Laciar. Patient non-specific algorithm for seizures detection in scalp EEG. *Computers in Biology and Medicine*, 71:128–134, 2016.

[OLBT20] A. O'Shea, G. Lightbody, G. Boylan, and A. Temko. Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *Neural Networks*, 123:12–25, 2020.

[Pow11] D.M. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Techonologies*, 2:37–63, 2011.

[SG10] A.H. Shoeb and J. Guttag. Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th International Conference on Machine Learning*, pages 975–982, 2010.

[SGA+21] A. Shoeibi, N. Ghassemi, R. Alizadehsani, M. Rouhani, H. Hosseini-Nejad, A. Khosravi, M. Panahiazar, and S. Nahavandi. A comprehensive comparison of handcrafted features and convolutional autoencoders for epileptic seizures detection in EEG signals. *Expert Systems with Applications*, 163:113788, 2021.

[SGK+20] A. Shoeibi, N. Ghassemi, M. Khodatars, M. Jafari, S. Hussain, R. Alizadehsani, P. Moridian, A. Khosravi, H. Hosseini-Nejad, M. Rouhani, et al. Epileptic seizure detection using deep learning techniques: A review. *arXiv preprint arXiv:2007.01276*, 2020.

[Sho09] A.H. Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

[SKG15] K. Samiee, P. Kovacs, and M. Gabbouj. Epileptic seizure classification of EEG time-series using rational discrete short-time Fourier transform. *IEEE Transactions on Biomedical Engineering*, 62(2):541–552, 2015.

[SKS+11] A.H. Shoeb, A. Kharbouch, J. Soegaard, S. Schachter, and J. Guttag. An algorithm for detecting seizure termination in scalp EEG. In *Proceedings of 2011 the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1443–1446. IEEE, 2011.

[SLuC15] C. Satirasethawong, A. Lek-uthai, and K. Chomtho. Amplitude-integrated EEG processing and its performance for automatic seizure detection. In *Proceedings of the 2015 IEEE International Conference on Signal and Image Processing Applications*, pages 551–556. IEEE, 2015.

[SSGMDEP19] R. San-Segundo, M. Gil-Martín, L. F. D'Haro-Enríquez, and J. M. Pardo. Classification of epileptic EEG recordings using signal transforms and convolutional neural networks. *Computers in Biology and Medicine*, 109:148–158, 2019.

[SSY+20] T. Sakai, T. Shoji, N. Yoshida, K. Fukumori, Y. Tanaka, and T. Tanaka. Scalpnet: Detection of spatiotemporal abnormal intervals in epileptic EEG using convolutional neural networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1244–1248. IEEE, 2020.

[TBB+11] D.J. Thurman, E. Beghi, C.E. Begley, A.T. Berg, J.R. Buchhalter, D. Ding, D.C. Hesdorffer, W.A. Hauser, L. Kazis, R. Kobau, et al. Standards for epidemiologic studies and surveillance of epilepsy. *Epilepsia*, 52(7):2–26, 2011.

[TDY+19] X. Tian, Z. Deng, W. Ying, K. Choi, D. Wu, B. Qin, J. Wang, H. Shen, and S. Wang. Deep multi-view feature learning for EEG-based epileptic seizure detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(10):1962–1972, 2019.

[TES+20] H. Takahashi, A. Emami, T. Shinozaki, N. Kunii, T. Matsuo, and K. Kawai. Convolutional neural network with autoencoder-assisted multiclass labelling for seizure detection based on scalp electroencephalography. *Computers in Biology and Medicine*, page 104016, 2020.

[TLLP20] F. Tang, Y. Liu, Y. Li, and Z. Peng. A unified multi-level spectral–temporal feature learning framework for patient-specific seizure onset detection in EEG signals. *Knowledge-Based Systems*, 205:106152, 2020.

[TTM+11] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G.B. Boylan. Performance assessment for EEG-based neonatal seizure detectors. *Clinical Neurophysiology*, 122(3):474–482, 2011.

[TYK16] N.S. Tawfik, S.M. Youssef, and M. Kholief. A hybrid automated detection of epileptic seizures in EEG records. *Computers & Electrical Engineering*, 53:177–190, 2016.

[VI17] L.S. Vidyaratne and K.M. Iftekharuddin. Real-time epileptic seizure detection using EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):2146–2156, 2017.

[YXJZ17] Y. Yuan, G. Xun, K. Jia, and A. Zhang. A multi-view deep learning method for epileptic seizure detection using short-time Fourier transform. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 213–222. ACM, 2017.

## Appendix A. Gradient of ScoreNet

This section describes full details of gradient calculation of the ScoreNet model. The gradients of the loss function with respect to the parameter are required to update the parameters. Consider the gradients computed at the epoch $i$. From (4) and (7), we can easily derive that

$$\frac{\partial \hat{y}_i}{\partial b4} = \hat{y}_i(1-\hat{y}_i), \quad \frac{\partial \hat{y}_i}{\partial a_4} = c_i \tilde{o}_l \hat{y}_i(1-\hat{y}_i), \quad \frac{\partial c_i}{\partial b_1} = c_i\left(1-c_i\right), \quad \frac{\partial c_i}{\partial a_1} = c_i\left(1-c_i\right)\mathbf{z}_i.$$

Similarly, from (5), the partial derivatives of $s_i$ with respect to $b_2$ and $a_2$ are

$$\frac{\partial s_i}{\partial b_2} = 1 - s_i^2, \quad \frac{\partial s_i}{\partial a_2} = \left(1 - s_i^2\right)\mathbf{z}_i.$$

In addition, according to (6), the gradients of $\tilde{o}_l$ with respect to $b_3$ and $a_3$ are

$$\frac{\partial \tilde{o}_l}{\partial b_3} = \tilde{o}_l\left(1-\tilde{o}_l\right), \quad \frac{\partial \tilde{o}_l}{\partial a_3} = \frac{1}{N_l}\tilde{o}_l\left(1-\tilde{o}_l\right)\sum_{j\in G_l} s_j.$$

To compute the gradient of $\mathcal{L}$, we first split the loss $\mathcal{L}$ into a summation of local losses: $\mathcal{L}(y,\hat{y}) = \sum_{l=1}^{m} \mathcal{L}_l(y,\hat{y})$ where $\mathcal{L}_l$ is a local loss. For instance, $\mathcal{L}_l(y,\hat{y}) = (1/N)\sum_{i\in G_l}\left[y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right]$ when the binary cross-entropy is employed, and $G_l$ contains indices of the group $l$. The gradients of each local loss are then calculated separately. Let $\alpha_l$ and $\beta_l$ be the first and last indices of candidates in the group $l$. According to (7), the derivatives of the local loss with respect to $b_4$ and $a_4$ are as follows:

$$\frac{\partial \mathcal{L}_l}{\partial b_4} = \sum_{i\in G_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial b_4} = \sum_{i\in G_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i}\hat{y}_i\left(1-\hat{y}_i\right) = \begin{bmatrix}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_{\alpha_l}} & \cdots & \frac{\partial \mathcal{L}_l}{\partial \hat{y}_{\beta_l}}\end{bmatrix}\begin{bmatrix}\hat{y}_{\alpha_l}(1-\hat{y}_{\alpha_l}) \\ \vdots \\ \hat{y}_{\beta_l}(1-\hat{y}_{\beta_l})\end{bmatrix}.$$

$$\begin{aligned}\frac{\partial \mathcal{L}_l}{\partial a_4} &= \sum_{i\in G_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial a_4} = \sum_{i\in G_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i}c_i\tilde{o}_l\hat{y}_i\left(1-\hat{y}_i\right) \\ &= \tilde{o}_l\begin{bmatrix}c_{\alpha_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_{\alpha_l}} & \cdots & c_{\beta_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_{\beta_l}}\end{bmatrix}\begin{bmatrix}\hat{y}_{\alpha_l}(1-\hat{y}_{\alpha_l}) \\ \vdots \\ \hat{y}_{\beta_l}(1-\hat{y}_{\beta_l})\end{bmatrix}.\end{aligned}$$

Next, similar to backpropagation in neural networks, the gradient of the local loss over $b_3$ is

$$\begin{aligned}\frac{\partial \mathcal{L}_l}{\partial b_3} &= \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \tilde{o}_l} \frac{\partial \tilde{o}_l}{\partial b_3} = \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} a_4 c_i \hat{y}_i (1 - \hat{y}_i) \tilde{o}_l \left(1 - \tilde{o}_l\right) \\ &= a_4 (1 - \tilde{o}_l) \tilde{o}_l \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} c_i \hat{y}_i (1 - \hat{y}_i) = a_4 (1 - \tilde{o}_l) \frac{\partial \mathcal{L}_l}{\partial a_4}.\end{aligned}$$

We can substitute $\frac{\partial \mathcal{L}_l}{\partial a_4}$ for $\tilde{o}_l \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} c_i \hat{y}_i (1 - \hat{y}_i)$ so that these terms are not repeatedly computed. Consequently, the gradients of the local loss with respect to the other ScoreNet parameters are demonstrated below.

$$\begin{aligned}\frac{\partial \mathcal{L}_l}{\partial a_3} &= \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \tilde{o}_l} \frac{\partial \tilde{o}_l}{\partial a_3} \\ &= \sum_{i \in G_l} \left[ \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} a_4 c_i \hat{y}_i (1 - \hat{y}_i) \tilde{o}_l \left(1 - \tilde{o}_l\right) \frac{1}{N_l} \sum_{j \in G_l} s_j \right] \\ &= \frac{1}{N_l} \left[ a_4 (1 - \tilde{o}_l) \tilde{o}_l \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} c_i \hat{y}_i (1 - \hat{y}_i) \right] \left( \sum_{j \in G_l} s_j \right) \\ &= \frac{1}{N_l} \frac{\partial \mathcal{L}_l}{\partial b_3} \sum_{j \in G_l} s_j. \end{aligned} \tag{A.1}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_l}{\partial b_2} &= \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \tilde{o}_l} \frac{\partial \tilde{o}_l}{\partial \sum s_j / N_l} \frac{\partial \sum s_j / N_l}{\partial b_2} \\ &= \sum_{i \in G_l} \left[ \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} a_4 c_i \hat{y}_i (1 - \hat{y}_i) \tilde{o}_l \left(1 - \tilde{o}_l\right) \frac{a_3}{N_l} \sum_{j \in G_l} (1 - s_j)^2 \right] \\ &= \frac{a_3}{N_l} \left[ a_4 (1 - \tilde{o}_l) \tilde{o}_l \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} c_i (\hat{y}_i - y_i) \right] \left( \sum_{j \in G_l} (1 - s_j^2) \right) \\ &= \frac{a_3}{N_l} \frac{\partial \mathcal{L}_l}{\partial b_3} \sum_{j \in G_l} (1 - s_j^2). \end{aligned} \tag{A.2}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_l}{\partial a_2} &= \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \tilde{o}_l} \frac{\partial \tilde{o}_l}{\partial \sum s_j / N_l} \frac{\partial \sum s_j / N_l}{\partial a_2} \\ &= \sum_{i \in G_l} \left[ \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} a_4 c_i \hat{y}_i (1 - \hat{y}_i) \tilde{o}_l \left(1 - \tilde{o}_l\right) \frac{a_3}{N_l} \sum_{j \in G_l} \left[ (1 - s_j^2) \mathbf{z}_j \right] \right] \\ &= \frac{a_3}{N_l} \left( a_4 (1 - \tilde{o}_l) \tilde{o}_l \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} c_i (\hat{y}_i - y_i) \right) \left( \sum_{j \in G_l} (1 - s_j^2) \mathbf{z_j} \right)\end{aligned}$$

$$= \frac{a_3}{N_l}\frac{\partial \mathcal{L}_l}{\partial b_3}\begin{bmatrix}\mathbf{z}_{\alpha_l} & \cdots & \mathbf{z}_{\beta_l}\end{bmatrix}\begin{bmatrix}\left(1-s_{\alpha_l}^2\right)\\ \vdots \\ \left(1-s_{\beta_l}^2\right)\end{bmatrix}. \tag{A.3}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_l}{\partial b_1} &= \sum_{i\in G_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial c_i}\frac{\partial c_i}{\partial b_1} \\
&= \sum_{i\in G_l}\left[\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i}a_4\tilde{o}_l\hat{y}_i(1-\hat{y}_i)c_i\left(1-c_i\right)\right] \\
&= a_4\tilde{o}_l\begin{bmatrix}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_{\alpha_l}} & \cdots & \frac{\partial \mathcal{L}_l}{\partial \hat{y}_{\beta_l}}\end{bmatrix}\begin{bmatrix}\hat{y}_{\alpha_l}(1-\hat{y}_{\alpha_l})c_{\alpha_l}(1-c_{\alpha_l})\\ \vdots \\ \hat{y}_{\beta_l}(1-\hat{y}_{\beta_l})c_{\beta_l}(1-c_{\beta_l})\end{bmatrix}.
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_l}{\partial a_1} &= \sum_{i\in G_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial c_i}\frac{\partial c_i}{\partial a_1} = \sum_{i\in G_l}\left[\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i}a_4\tilde{o}_i\hat{y}_i(1-\hat{y}_i)\cdot c_i\left(1-c_i\right)\mathbf{z}_i\right] \\
&= a_4\tilde{o}_l\begin{bmatrix}\mathbf{z}_{\alpha_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_{\alpha_l}} & \cdots & \mathbf{z}_{\beta_l}\frac{\partial \mathcal{L}_l}{\partial \hat{y}_{\beta_l}}\end{bmatrix}\begin{bmatrix}\hat{y}_{\alpha_l}(1-\hat{y}_{\alpha_l})c_{\alpha_l}(1-c_{\alpha_l})\\ \vdots \\ \hat{y}_{\beta_l}(1-\hat{y}_{\beta_l})c_{\beta_l}(1-c_{\beta_l})\end{bmatrix}.
\end{aligned} \tag{A.4}$$

These calculations are similar to backpropagation in neural network, and we can see that some terms are repeatedly used for calculation. For instance, $\frac{\partial \mathcal{L}_l}{\partial b_3}$ is used in (A.1), (A.2), and (A.3). Hence, we can compute these repeated terms once and cache them to reduce time for the calculations. Finally, the gradients of $\mathcal{L}$ are obtained by combining the gradients of $\mathcal{L}_l$, *i.e.*, $\nabla\mathcal{L} = \sum_{l=1}^{m}\nabla\mathcal{L}_l$.

## Appendix B. Database summary

This section presents a record summary of the CHB-MIT Scalp EEG database as shown in Table B.3.

## Appendix C. Initial point selection

This section explains the details of selecting the initial points of ScoreNet parameters in this study. We selected three initial points due to the prior classification results based on the training data. From (A.3) and (A.4), as $z_i$ is

Table B.3: Record summary of the CHB-MIT Scalp EEG database.

| **Cases** | **# Records** | **EEG duration (sec)** | **# Seizures** | **Seizure duration (sec)** |
|---|---|---|---|---|
| chb01 | 42 | 145,988 | 7 | 449 |
| chb02 | 36 | 126,959 | 3 | 175 |
| chb03 | 38 | 136,806 | 7 | 409 |
| chb04 | 42 | 561,834 | 4 | 382 |
| chb05 | 39 | 140,410 | 5 | 563 |
| chb06 | 18 | 240,246 | 10 | 163 |
| chb07 | 19 | 241,388 | 3 | 328 |
| chb08 | 20 | 72,023 | 5 | 924 |
| chb09 | 19 | 244,338 | 4 | 280 |
| chb10 | 25 | 180,084 | 7 | 454 |
| chb11 | 35 | 123,257 | 3 | 809 |
| chb12 | 24 | 85,300 | 40 | 1,515 |
| chb13 | 33 | 118,800 | 12 | 547 |
| chb14 | 26 | 93,600 | 8 | 177 |
| chb15 | 40 | 144,036 | 20 | 2,012 |
| chb16 | 19 | 68,400 | 10 | 94 |
| chb17 | 21 | 75,624 | 3 | 296 |
| chb18 | 36 | 128,285 | 6 | 323 |
| chb19 | 30 | 107,746 | 3 | 239 |
| chb20 | 29 | 99,366 | 8 | 302 |
| chb21 | 33 | 118,189 | 4 | 203 |
| chb22 | 31 | 111,611 | 3 | 207 |
| chb23 | 9 | 95,610 | 7 | 431 |
| chb24 | 22 | 76,640 | 16 | 527 |
| **sum** | **686** | **3,536,540** | **198** | **11,809** |

an output of a classification problem of highly imbalanced data, $z$ is extremely sparse; $\mathbf{z}_i$ is usually a zero vector, and the gradients become vanished. Thus, the parameters $a_1$ and $a_2$ are slightly updated, and hardly converge to a good local minimum. In this case, we mainly initialized $a_1$ and $a_2$ first based on three situations, and the initialization of the other parameters was empirically obtained to avoid a convergence to poor optimal points. Initial points of $a_1$ and $a_2$ were based on the parameters of the counting based method, *i.e.*, the initial points were multiples of $\mathbf{1}$. When there are many false negatives from the prior classification step, magnitudes of $a_1$ and $a_2$ should be high to boost a seizure

candidate $c_i$ and we easily assign $s_i = 1$. On the other hand, the magnitudes of $a_1$ and $a_2$ should be small to suppress the effect of false positives when the number of isolated false positive is dominant. In addition, $a_1$ and $a_2$ should have the intermediate magnitudes relatively compared to these two cases if the numbers of false positives and negatives are slightly different. Finally, the initial values according to the circumstances are eventually listed in Table C.4.

Table C.4: List of initial points. Vector of ones is represented by $\mathbf{1}$.

| **Cases** | **Parameters** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ |
| High false positive | $\mathbf{1}$ | $\mathbf{1}$ | 3 | 3 | $-1$ | $-2$ | 0 | $-1$ |
| High false negative | $9 \cdot \mathbf{1}$ | $9 \cdot \mathbf{1}$ | 4 | 4 | $-4$ | $-8$ | 1 | $-1$ |
| Balance | $6 \cdot \mathbf{1}$ | $6 \cdot \mathbf{1}$ | 3 | 3 | $-3$ | $-9$ | 0 | $-1$ |