Version of Record: https://www.sciencedirect.com/science/article/pii/S1746809420303475
Manuscript_f4361ec5af09216a4f2fa0fd86ea8abc

A Robust Deep Learning Approach for Automatic Classification of Seizures Against Non-seizures

Xinghua Yao[a], Xiaojin Li[a], Qiang Ye[b], Yan Huang[a], Qiang Cheng[a,*], Guo-Qiang Zhang[c,*]

[a]*Institute of Biomedical Informatics, University of Kentucky, Lexington, Kentucky, USA*
[b]*Department of Mathematics, University of Kentucky, Lexington, Kentucky, USA*
[c]*The University of Texas Health Science Center at Houston, Houston, Texas, USA*

Abstract

Identifying epileptic seizures through analysis of the electroencephalography (EEG) signal becomes a standard method for the diagnosis of epilepsy. Manual seizure identification on EEG by trained neurologists is time-consuming, labor-intensive and error-prone, and a reliable automatic seizure/non-seizure classification method is needed. One of the challenges in automatic seizure/non-seizure classification is that seizure morphologies exhibit considerable variabilities. In order to capture essential seizure patterns, this paper leverages an attention mechanism and a bidirectional long short-term memory (BiLSTM) to exploit both spatial and temporal discriminating features and overcome seizure variabilities. The attention mechanism is to capture spatial features according to the contributions of different brain regions to seizures. The BiLSTM is to extract discriminating temporal features in the forward and the backward directions. Cross-validation experiments and cross-patient experiments over the noisy data of CHB-MIT are performed to evaluate our proposed approach. The obtained average sensitivity of 87.30%, specificity of 88.30% and precision of 88.29% in cross-validation experiments are higher than using the current state-of-the-art methods, and the standard deviations of our approach are lower. The evaluation results of cross-patient experiments indicate that, our approach has good performance in comparisons with the current state-of-the-art methods and is more robust across patients.

Keywords:
attention mechanism, bidirectional LSTM, seizure/non-seizure classification, deep learning

1. Introduction

More than 50 million people in the world suffer from epilepsy [1]. Epilepsy is a central nervous system disorder, in which brain activity becomes abnormal, causing seizures or periods of unusual behaviors, sensations, and sometimes loss of awareness. An important technique to diagnose epilepsy is electroencephalography (EEG). An EEG signal records the electrical activities of the brain, and may reveal patterns of normal or abnormal brain electrical activities. In current clinical practices, EEG signals are collected from the brains by making use of either non-intrusive or implanted devices. The collected off-line EEG signals are then reviewed and analyzed by trained neurologists to identify characteristic patterns of the disease, such as pre-ictal spikes and seizures (A seizure is a sudden, uncontrolled electrical disturbance in the brain, which signifies epilepsy.), and to capture disease information, like seizure frequency, seizure type, etc. The obtained disease information is to provide supports for therapeutic decisions. This manual way of reviewing and analyzing is labor-intensive and error-prone, for it usually takes several hours for a well-trained expert to analyze one-day of recordings from one patient [2, 3, 4, 5, 6]. These limitations have motivated researchers to develop automated techniques to recognize seizure. In this paper, we focus on developing an automatic approach to classifying seizure signal segments and non-seizure segments from off-line EEG signals for assisting neurologists to make diagnosis.

One of critical challenges in the seizure/non-seizure classification is that seizure morphologies exhibit considerable inter-patient and intra-patient variabilities. Different machine learning methods and computational technologies have been applied to address this challenge. Seizure detection is often converted into a problem of seizure/non-seizure classification but more of a real-time flavor. Extensive studies have been conducted for constructing patient-specific detectors capable of detecting seizures [6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. In early studies, hand-crafted features are usually used as characteristics of seizure manifestations in EEG. More recent studies focus on applying deep learning models to seizure detection [4, 13, 16, 17, 18, 19]. Most of these studies adopt interesting technologies to help extracting seizure features. For example, signal processing techniques are used to filter the data; certain modules need to be pre-trained; multiple channels are utilized to extract spatial features, attention mechanisms are leveraged to learn channel-specific information, and temporal features are extracted by the sliding windows. However, to the best of our knowledge, either that channels are not differentiated, or that the attention mechanism is deployed after finishing local fea-

*Corresponding authors
Email addresses: Qiang.Cheng@uky.edu (Qiang Cheng), gqatcase@gmail.com (Guo-Qiang Zhang)

ture extraction. In the dynamic brain activities, spatial structure in a brain is relatively stable. It is better that the spatial feature extraction is conducted before temporal feature extraction. Such an order could reduce needless information or artifacts. About extracting temporal features, most studies only work in the forward direction. In fact, for seizure/non-seizure classification, the EEG signals can potentially provide some additional information in the backward direction [13].

Different brain regions are likely to have different contributions to the seizure. The characteristics of EEG data for epilepsy at different brain regions are different. The features of EEG signals at a time point are correlated with the past data and the future data. Besides, though EEG signals are in general dynamic and non-linear, during a sufficiently small time period, the signal may be considered to be stationary. For example, a signal with a duration of 1/256 seconds in the EEG data set CHB-MIT could be static. It only contains one data point. Based on the above three observations and inspired by an architecture in [20], we design a new approach by using bidirectional long short-term memory (BiLSTM) integrated with an attention mechanism. Firstly, we introduce an attention mechanism over EEG channels. Different weights are automatically assigned to signal channels at different brain regions according to how much they would affect the seizures. Secondly, the bidirectional long short-term memory technique is adopted to extract temporal features of EEG signals in both the forward and the backward directions. Thirdly, output sequences of the BiLSTM module are split into no-overlap patches in the time order. Each patch only contains a data point in the output sequences. All the patches are separately processed to extract features. With these three new ideas, we develop a novel approach for seizure/non-seizure classification in EEG signals. Cross-validation and cross-patient experiments are performed on the EEG data set CHB-MIT using the proposed approach. Signals on 17 common channels in the data set are selected and each signal is segmented according to a duration of 23 seconds. In the cross-validation experiments, we obtain the average sensitivity, specificity and precision of 87.3%, 88.3% and 88.29%, respectively, and the corresponding standard deviations of 0.0287, 0.0316 and 0.0265, respectively. For the cross-patient experiments, the average sensitivity, specificity and precision of 83.72%, 84.06% and 85.36% are respectively achieved, and the standard deviations being 0.1349, 0.1379 and 0.1020, respectively. These results exceed the current state-of-the-art performances on the noisy data of CHB-MIT in [17], [20] and [4]. The extensive experimental results show that the performance of the proposed new approach is promising and has high stability, with smaller variations compared to existing methods.

In brief, the main novelties of our paper include the following:

(1) An attention mechanism is utilized to capture spatial features of seizure. It distinguishes EEG signals from different brain regions and generates different attention weights for EEG data over different channels. The attention weights are explained by using EEG data segment examples.

(2) Bidirectional long short-term memory is combined with attention mechanism to extract temporal features. At each time step, the past spatially-weighted data and the future spatially-weighted data are analyzed.

(3) Experimental results on the noisy EEG data of CHB-MIT demonstrate that, the new approach can capture more robust seizure patterns than current state-of-the-art deep learning approaches, and overcome the inter-patient seizure variations better.

The rest of this paper is organized as follows. Section 2 describes related research work on automatic seizure/non-seizure classification. Section 3 presents our designed approach of BiLSTM with attention. In Section 4, evaluation of the proposed approach is performed in cross-validation and cross-patient experiments. Section 5 explains the attention mechanism and validates main modules in the proposed approach. Section 6 discusses the approach of BiLSTM with attention. Conclusions and future work are described in Section 7.

2. Related work

There is extensive research for seizure/non-seizure classification, which distinguishes seizure segments from non-seizure segments. Seizure detection, which is often of a real-time flavor, is often viewed as the seizure/non-seizure classification problem. The study of seizure detection can be divided into three categories. One category is using traditional machine learning methods [7, 8, 10, 11, 12, 21, 22, 23, 24, 25]. The second category is about signal processing methods and network techniques [6, 9, 15, 26, 27, 28, 29]. And the third category is using deep learning methods [4, 13, 16, 17, 18, 20, 30, 31, 32].

2.1. Work based on machine learning methods

With traditional machine learning methods, many previous works focus on developing patient-specific seizure detection methods [7, 8, 10, 11, 12, 24, 33].

Shoeb and Guttag proposed a patient-specific seizure detection method by using the support vector machine (SVM) [7]. The method leverages filters to extract spectral features over each channel, and then concatenate the feature vectors according to a fixed time length. Then, train the SVM model with the obtained feature vectors as the input. The method achieved a sensitivity of 96%, a median detection delay of 3 seconds and a median false detection rate of 2 per 24 hours. The sensitivity result is often used as a benchmark for patient-specific seizure detection on the data set CHB-MIT. The authors observed that the identity of channels could help differentiate between the seizure and the non-seizure activity.

Amin and Kamboh [8] designed an algorithm RUSBoost to process imbalanced seizure/non-seizure data, and used RUSBoost and the decision tree classifier to conduct patient-specific experiments with the CHB-MIT data set. The method was fast in training and achieved good performance with seizure detection accuracy of 97% and false detection rate of 0.08 per hour.

Table 1
Summary of existing EEG-based seizure detection methods.

Reference	EEG type	No. of subjects	No. of seizures	Patient-specific	Split data for training	Sens.	Spec.	Accu.	FAR
Acharya et al. [17]	scalp	10	n/a	no	70%	95%	90%	88.67%	n/a
Hussein et al. [20]	scalp	10	n/a	no	10-folds CV	100%	100%	100%	n/a
Shoeb and Guttag [7]	scalp	23	198	yes	Leave-one-out CV	96%	n/a	n/a	2 per 24 hours
Thodoroff et al. [4]	scalp	23	198	no	n/a	85%	n/a	n/a	0.8 per hour
Fergus et al. [21]	scalp	23	198	no	80%	88%	88%	n/a	n/a
Amin and Kamboh [8]	scalp	23	198	yes	50%	88%	n/a	n/a	0.0831 per hour
Yuan et al. [32]	scalp	23	198	no	5-folds CV	n/a	n/a	94.37%	n/a
Zandi et al. [6]	scalp	14	63	yes	Leave-one-out CV	90.5%	n/a	n/a	0.51 per hour
Saab and Gotman [27]	scalp	44	195	no	64%	76%	n/a	n/a	0.34 per hour
Kuhlmann et al. [28]	scalp	21	88	no	70%	81%	n/a	n/a	0.6 per hour
Wang et al. [29]	scalp	10	44	yes	5-folds CV	91.44%	99.34%	98.3%	n/a
Truong et al. [12]	intracranial	12	n/a	yes	Leave-one-out CV	89.4%	89.24%	n/a	n/a
Kharbouch et al. [24]	intracranial	10	67	yes	Leave-one-out CV	97%	n/a	n/a	0.6 per 24 hours

Sens. is an abbreviation for Sensitivity, Spec. for Specificity, Accu. for Accuracy, FAR for False Alarm Rate. These abbreviations are also used in Tables 3, 4, 5, 6 and 7.

Hunyadi et al. [10] presented seizure detection algorithm, which uses a nuclear norm regularization to convey spatial distribution information of ictal patterns. The algorithm extracted features from each channel, and then stacked them to analyze as one entity.

Truong et al. [12] proposed a automatic seizure detection method over intracranial electroencephalography (iEEG) data. First, supervised classifiers were used to select those channels that contribute the most to seizures. Features in the frequency and time domains were extracted, including spectral power and correlations between channel pairs. Then, Random Forest classifier was utilized for classification. This method has the state-of-the-art computational efficiency while maintaining the accuracy. In this method, selecting channels that contribute the most to seizures is to reduce the number of channels, thereby improving the computational efficiency.

The work in [7, 8, 10, 12] used data over multiple channels to extract spatial features. However, they did not apply different processing ways to the data with different channels.

Esbroeck et al. [11] proposed a multi-task learning framework to detect patient-specific seizure onset in the presence of intra-patient variability in seizure morphology. They considered distinguishing the windows of each seizure from non-seizure data as a separate task and treating the individual-seizure discrimination as another task. Compared to the standard SVM, testing results of the CHB-MIT data set indicated that their approach performed better in most cases.

Kiranyaz et al. [33] presented a systematic approach for patient-specific classification of long-term EEG. In the approach, EEG data were processed through band-pass filtering, feature extraction, epileptic seizures aggregation and morphologic filtering. Results of the data processing were input into collective network of binary classifiers to classify signal from each channel. Then, initial classification results over each channel were further learned and weighted by a dedicated classifier which makes final classification decision of each EEG frame. Over the CHB-MIT data set, [33] achieved an average sensitivity of 89.01% and an average specificity of 94.71%. High number of classifiers increased computational complexity of the approach.

In the patient-specific case, the data have no variations caused by different subjects. The performances of the patient-specific seizure/non-seizure classifiers are better than 90%. However, the patient-specific classifiers have a limitation of poor generalizability.

In [21], Fergus et al. presented a method for seizure detection across subjects based on traditional machine learning techniques, and obtained 88% in Sensitivity and 88% in Specificity over the CHB-MIT data set by selecting features in multiple brain regions. The method mainly consists of four steps, which are data filtering, feature extraction, feature selection and training classifiers. In cross-validation experiments, EEG signals in CHB-MIT were segmented according to a segment length 60 seconds, one seizure segment was truncated for each seizure, non-seizure segments were extracted from non-seizure EEG records as many as seizure segments. The produced experiment data consist of 171 seizure segments and 171 non-seizure segments. On the average, each seizure segment contains 40s seizure data. Additionally, after segmenting EEG signals [21] used a bandpass filter and second order butterworth filters to extract the EEG data in the bandwidth 0.5Hz-30Hz.

2.2. Work based on signal processing and network techniques

Based on signal processing techniques, Zandi et al. proposed a wavelet-based algorithm for real-time detection of epileptic seizures using scalp EEG [6]. In this algorithm, the EEG from each channel was decomposed by wavelet packet transform, and a patient-specific measure was developed by using wavelet coefficients to separate the seizure and non-seizure states. Utilizing the measure, a combined seizure index was derived for each epoch of every EEG channel. Appropriate channel alarms were generated by inspecting the combined seizure index.

Acharya et al. [26] presented a method for the automatic detection of normal, pre-ictal, and ictal conditions from EEG signals. Four entropy features, including approximate entropy, sample entropy, and two phase entropies, were extracted. The extracted features were input into the classifier to do classification. Over the EEG data set provided by University of Bonn, seven classifiers were fed with extracted entropies to show the effectiveness of the features.

In [27], Saab and Gotman developed an online seizure alert system, which employed a wavelet decomposition module, feature extraction module and data segmentation module to compute a probability of that a seizure activity happens. The system was trained using 652 h of scalp EEG data. And it was tested on a separate EEG dataset including 360 h of scalp EEG data. A sensitivity of 76.0% and a false detection rate of 0.34 per hour were obtained.

Kuhlmann et al. [28] analyzed seizure detection features and their combinations using a probability-based seizure detection framework designed by [27]. The experiments for the analyses were performed on 525 h of scalp EEG data. Experimental results showed that, a detector based on a combination of features achieved a sensitivity of 81%, false positive rate of 0.60 per hour, and median detection delay of 16.9 s.

Wang et al. [29] presented a new approach based on partial directed coherence (PDC) analysis to detect seizure intervals of epilepsy patients. The PDC analysis was utilized as a mechanism to extract features, such as the direction and intensity of information flow related to EEG channels. Features of the outflow information were fed to a support vector machine classifier for discriminating interictal periods and ictal periods. The presented method was evaluated on a scalp EEG data set which included ten patients and 88 seizures. For each patient, 5-fold cross-validation was performed. Experimental results showed an average sensitivity of 91.44%, average specificity of 99.34%, and average accuracy of 98.30%.

Zhou et al. [15] proposed a seizure detection algorithm using lacunarity and Bayesian linear discriminant analysis (BLDA). In the algorithm, wavelet decomposition on EEGs was conducted with five scales, and the wavelet coefficients at scales 3, 4, and 5 were selected. Features including lacunarity and fluctuation index were extracted from the selected scales, and then they were fed to the BLDA for training and classification. Patient-specific experiments were performed on intracranial EEG data from the Epilepsy Center of the University Hospital of Freiburg. The obtained average sensitivity was 96.25%, with an average false detection rate of 0.13 per hour and a mean delay time of 13.8s. The obtained precision results for eleven patients were less than 50%.

By leveraging network technologies, Fan and Chou [9] utilized a complex network model to represent EEG signals, and integrated it with spectral graph theory to extract spatial-temporal synchronization patterns for detecting seizure onsets in real-time. The method was tested on 23 patients from the CHB-MIT data set. The resulting patient-specific sensitivity surpassed the benchmark methods.

2.3. Work based on deep learning methods

Recently, deep learning techniques have been developed rapidly and applied to solve the seizure/non-seizure classification problem.

Vidyaratne et al. [13] proposed a deep recurrent architecture by combining Cellular Neural Network and Bidirectional Recurrent Neural Network. The bidirectional recurrent neural network was deployed into each cell in the cellular neural network, and it was utilized to extract temporal features in the forward and the backward directions. Each cell interacts with its neighbor cells to extract local spatial-temporal features. The computed results in the cellular neural network were output into a multi-layered perceptron. In the perceptron, samples were classified based on a trained threshold. In order to satisfy the input requirements of cellular neural network, the authors proposed a mapping which organizes EEG signals into a 2D grid arrangement. Patient-specific experiments were conducted over the EEG data of five patients from the CHB-MIT data set. The obtained sensitivities were all 100% for the five patients. In their experiments, the raw EEG data were preprocessed using a bandpass filter between 3Hz and 30Hz in order to extract seizure activity data.

Golmohammadi et al. [16] explored seizure-detection performances of two neural networks over the data source of TUH EEG Corpus introduced in [34]. Their experiment results showed that the convolutional long short-term memory (LSTM) network is better than the convolutional GRU network. And also the impacts of initialization methods and regularization methods over the performance were experimented. The two models in [16] did not utilize attention mechanism.

Hussein et al. [20] designed a deep neural network for seizure/non-seizure classification by using LSTM as a main module. The approach extracts temporal features by using LSTM. Evaluation was performed on the EEG data set provided by University of Bonn. Testing results mostly reached 100%. In [17], Acharya et al. presented a 13-layers deep neural network for seizure/non-seizure classification by using convolutional neural network (CNN). Over the Bonn EEG data set, the obtained average sensitivity and specificity were 95% and 90%, respectively. For the experiments in [20] and [17], the two approaches extracted seizure features from the data on one channel to conduct classification. Each record in the Bonn EEG data set is the data from only one channel.

In [4], Thodoroff et al. designed a recurrent convolutional neural network to capture spectral, spatial and temporal patterns of seizures. The EEG signals were firstly transformed into images by using Polar Projection, cubic interpolation, and Fast Fourier transform. The image-based representation of EEG signals was to exploit the spatial locality in seizures. Created images were fed to the convolution neural network. The output vectors of the convolution neural network were organized to be sequences in chronological order. The sequences were then input into the bidirectional recurrent neural network to produce classified seizure/non-seizure results. Both patient-specific experiments and cross-patient experiments were performed. The patient-specific experiment results were similar to the results in [7]. And the cross-patient testing sensitivity was 85% on average. In the two kinds of experiments, the convolution neural network was pre-trained alone. And the transfer learning technology was utilized to overcome the problem of small amount of data in the patient-specific experiments. The proposed recurrent convolutional neural network in [4] is complicated.

Ansari et al. [30] aimed to automatically optimize feature selection for seizure detection. They utilized deep CNN to extract optimal features, and then fed the features to random forest to

do classification. In evaluation experiments, EEG recordings of 26 and 22 neonates were taken as training data and testing data, respectively. A false alarm rate of 0.9 per hour and a sensitivity of 77% were achieved. The proposed method needed no predefined features, and surpassed three classic feature-based approaches.

Yuan et al. [32] presented a unified multi-view deep learning framework to capture brain abnormalities associated with seizures based on multi-channel scalp EEG signals. In the framework, an autoencoder-based model was constructed to learn inter and intra correlations of EEG channels. The learned correlations were combined with features to detect seizures, which were extracted in supervised learning via spectrogram representation. In order to evaluate the proposed method, 5-folds cross-validation experiments were performed. The performances reach to accuracy of 94.37%, F1-score of 85.34%, and area under receiver operating characteristic curve (shortly, AUC-ROC) of 95.72%.

In [18], Yuan et al. proposed a model ChannelAtt, an end-to-end multi-view deep learning model with channel-aware attention mechanism, to detect seizures in EEG signals. The model employed a global-based attention, which can score contributions of channels dynamically and capture relationships among channels. In the evaluation of the model, EEG signal data from 9 patients in the EEG dataset CHB-MIT were taken as experimental data, and hold-out validations were conducted. Four metrics were used to measure the seizure-detection performance, which include area under precision-recall curve (shortly, AUC-PR), AUC-ROC, F1-score, and Accuracy. Experimental results show AUC-PR of 0.9651, AUC-ROC of 0.9847, F1-score of 97.85%, and Accuracy of 96.61%. The global-based attention mechanism in the model ChannelAtt captures correlations between individual EEG channel and the global state of brain. The correlations is described to be weights. The weights represent reflect contributions of channels to the brain activities. Features on channels are summed according to the weights. The weighted sum is treated as local features, and is concatenated with the global features to discriminate seizures and non-seizures. The model ChannelAtt is only evaluated on a part of EEG data in the dataset CHB-MIT.

In [31], Yuan et al. explored another way to fuse global features and local information in multi-channel biosignals. A deep fusional attention network, namely FusionAtt, was developed. In the model FusionAtt, two convolutional encoders and a fusional attention mechanism are designed. The two convolutional encoders are employed to extract global features and channel-specific features. The attention mechanism is to assign different attention energies to channels. Attention energies represent the importance to the target task. The model FusionAtt is evaluated using two clinical tasks, including multi-channel EEG seizure detection and multivariate PSG sleep stage classification. For the task of seizure detection, the proposed model is tested on the EEG dataset CHB-MIT, and the performances reach to AUC-ROC of 0.9556, AUC-PR of 0.9119, F1 score of 86.75%, and accuracy of 95.06%.

Zhang et al. [19] developed a new deep neural network model, which can learn seizure-specific representations from the raw EEG signals to classify seizures and non-seizures. The model extracts seizure features and patient information separately through adversarial training. The extracted seizure features are insensitive to the patient identity, and the extracted patient information is insensitive to the seizure state. At the mean time, an attention mechanism is designed to learn the importance of each EEG channel and calculate attention weights. The seizure features are combined with the attention weights, and they are computed to detect seizures. On a EEG dataset TUH corpus introduced in [34], the proposed method is evaluated in leave-one-out cross-validation experiments. A sensitivity of 97.4% and specificity of 88.1% are achieved.

Attention mechanisms were developed separately in [18, 19, 31] to differentiate channels in multi-channel EEG data. Channels were assigned different attention energies. The designed attention mechanisms helped improve the performances of models in [18, 19, 31]. There are differences among the attention mechanisms. [18] utilized local information and concatenated information to calculate attention energies. In [31], a gated function was leveraged to fuse the global and channel-specific information for the attention energy assignment. And [19] adopted a multiplication with local features to learn the attention weights.

Subsections 2.1, 2.2, and 2.3 review seizures/non-seizures classification research work according to the utilized EEG data analysis technologies. Table 1 summarizes the reported performances of existing EEG-based seizure detection methods in the recent years. The summarized methods include traditional machine learning methods, signal processing methods, and deep learning methods. They were evaluated in different validation methods and on different EEG data sets, and will be made comparisons with our proposed approach in Subsection 4.3.

3. Methods

3.1. Model design

EEG signal data is an important modality for the diagnosis of epilepsy. It is generally collected through placing electrodes on the scalp. Each electrode records brain activities in its located brain region. As different brain regions play different roles in the seizure procedure, the data collected at different brain regions record different characteristics of seizures. With the observations in [7], differences between seizure data and non-seizure data are related to channels. To exploit the differences of signals from different brain regions, we will use an attention mechanism to assign different weights to data from different channels.

Brain activities are continuous, and EEG signals could be regarded as continuous records of brain activities when ignoring the sampling effects. The brain activity at a time point is correlated with past signal data, and could also be analyzed from future signal data. To leverage correlations from both directions, we perform BiLSTM for analyzing EEG sequence data.

EEG signal is dynamic and non-linear. Due to the dynamic nature, certain statistical characteristics of EEG signals change over time. However, the EEG signal segments have similar statistical temporal and spectral features for a sufficiently small

time duration [20, 35]. A splitting operation is executed on the output sequences of bidirectional processing. Each output sequence is split into patches in the same order. Each patch only contains one data point. The patches are further extracted features through full connection operations separately and concurrently.

Based on the above three ideas and inspired by [20], we develop a new approach of BiLSTM with attention (shortly, attention BiLSTM) in order to classify seizure segments and non-seizure segments. Raw EEG signals are split into data segments according to a fixed time span. The split data segments are automatically weighted through an attention mechanism, i.e., for each segment, signal data from different channels are multiplied with different weights. In our attention mechanism, a fully connected module and a non-linear function are employed to obtain a matrix, in which each element is in an interval of [0, 1]. And then, means of elements in the obtained matrix are calculated along the second array dimension. The means are spatial features of EEG signal segments, and they are treated as weights on channels. A weight on a channel represents that, how reliable the characteristics of EEG signals in a brain region are with respect to signifying seizures. After adding weights, the data segments are fed to bidirectional LSTM module. The BiLSTM module extracts features in both forward and backward directions. For output sequences of BiLSTM, data at each time step are separately input into a full connection module. Then, the extracted features are averaged over all the time steps in order to achieve global features of a segment. Finally, the labels of data segments are calculated by a fully connected module with the Softmax function.

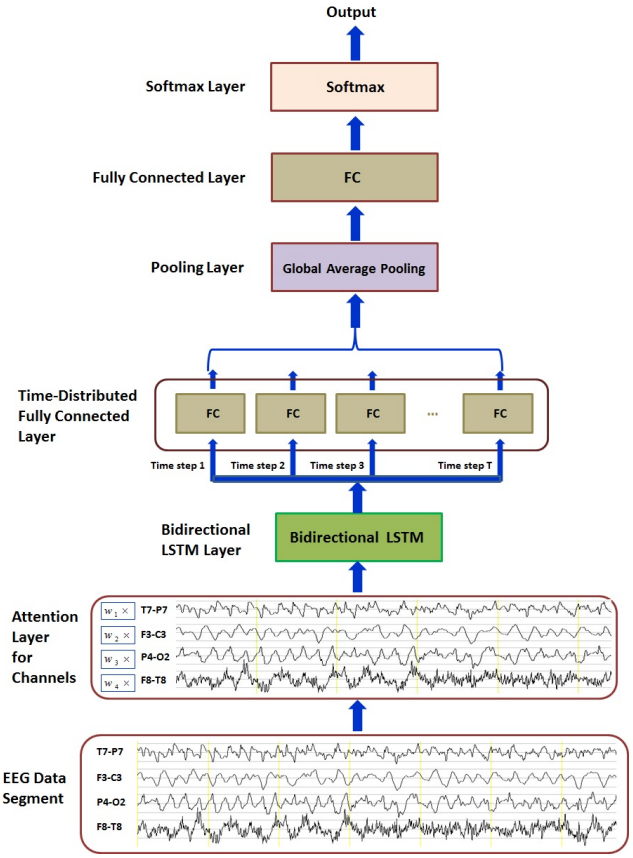


Fig. 1. Architecture of the proposed approach. T7-P7, F3-C3, P4-O2 and F8-T8 represent channels. W_1, W_2, W_3 and W_4 are weights on the four channels, respectively.

3.2. Model architecture and algorithm

Our model architecture consists of five modules, including attention layer, BiLSTM module, time-distributed fully-connected layer, pooling layer and fully-connected layer with Softmax. The designed architecture is presented in Fig. 1.

3.2.1. Attention layer

The attention layer, shown in Fig. 2, is to generate attention weights for each channel and then executes an element-wise multiplication. The original data are input into a fully connected module with a nonlinear activation function. The outputs of the fully connected module are averaged over all the time steps. Then, the obtained average values are copied to be shared at all time steps. In this way, an attention weight matrix is achieved. Finally, the attention matrix is element-wisely multiplied with the original inputs. The attention layer is computed using the following equations:

$$Y_1 = f_{re_1}(X_0) \tag{1}$$
$$Y_2 = \sigma(Y_1 * W_{al} + B_{al}) \tag{2}$$
$$Y_3 = f_{re_2}(Y_2) \tag{3}$$
$$Y_4 = f_{av}(Y_3) \tag{4}$$
$$Y_5 = f_{cy}(Y_4) \tag{5}$$
$$Y_{al} = X_0 \odot Y_5 \tag{6}$$

Here, X_0 denotes an input tensor of size (n_{sm}, n_{sp}, n_{ch}). Symbols n_{sm}, n_{sp}, n_{ch} represent the number of samples, the number of time steps, and the number of signal channels, respectively. Y_1 is a matrix of size (n_{ss}, n_{ch}), $n_{ss} = n_{sm} * n_{sp}$, W_{al} a weight matrix of size (n_{ch}, n_{ch}), a bias matrix B_{al} of size (n_{ss}, n_{ch}), and Y_2 with size (n_{ss}, n_{ch}). A symbol $\sigma(\cdot)$ represents a non-linear function, like $softmax(\cdot)$ and $sigmoid(\cdot)$. Y_3 is a matrix of size (n_{sm}, n_{sp}, n_{ch}), Y_4 of size (n_{sm}, n_{ch}), Y_5 of size (n_{sm}, n_{sp}, n_{ch}), and Y_{al} an output matrix of attention layer with shape (n_{sm}, n_{sp}, n_{ch}). Functions $f_{re_1}(\cdot)$ and $f_{re_2}(\cdot)$ are to reshape a matrix, $f_{av}(\cdot)$ is a function of computing averages along with the second axis of matrix, and $f_{cy}(\cdot)$ is an copying operation to share the averages over all the time steps. The symbol $\odot$ means an element-wise multiplication between matrices.

3.2.2. BiLSTM module

The BiLSTM module processes the input sequence separately according to the forward order and the backward order, and synthesize the forward outputs and the backward outputs [36, 37]. Its main procedure is presented in Fig. 3. In either forward order or backward order, the sequence is computed in the same way as LSTM, in which the computation can be described by using Eqs. (7)−(12) according to [38] and [39]. The synthesizing operations can be concatenation or summation.

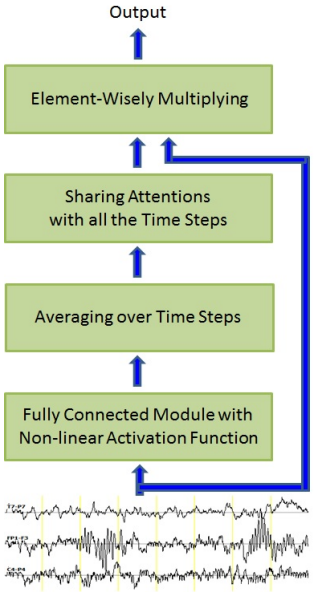


Fig. 2. Work flow of attention layer.

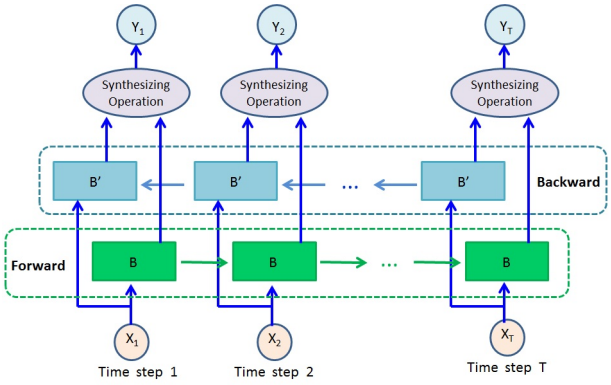


Fig. 3. Work flow of BiLSTM module.

$$\text{Block input} \quad \widetilde{C}_t = \varphi(X_t^{in} * W_{ce} + Y_{t-1}^{bo} * R_{ce} + B_{ce}) \tag{7}$$

$$\text{Input gate} \quad G_t^{ig} = \sigma(X_t^{in} * W_{ig} + Y_{t-1}^{bo} * R_{ig} + B_{ig}) \tag{8}$$

$$\text{Forget gate} \quad G_t^{fg} = \sigma(X_t^{in} * W_{fg} + Y_{t-1}^{bo} * R_{fg} + B_{fg}) \tag{9}$$

$$\text{Output gate} \quad G_t^{og} = \sigma(X_t^{in} * W_{og} + Y_{t-1}^{bo} * R_{og} + B_{og}) \tag{10}$$

$$\text{Cell} \quad C_t = C_{t-1} \odot G_t^{fg} + \widetilde{C}_t \odot G_t^{ig} \tag{11}$$

$$\text{Block output} \quad Y_t^{bo} = \psi(C_t) \odot G_t^{og} \tag{12}$$

Here, X_t^{in} is an input matrix of size (n_{sm}, n_{ch}) at the time step t, and Y_t^{bo} an output matrix of size (n_{sm}, n_{fe_1}) at the time step t, where n_{fe_1} is a dimensionality of extracted feature space. Matrices G_t^{ig}, G_t^{fg}, G_t^{og}, $\widetilde{C}_t$, and C_t represent input gate state, forget gate state, output gate state, a block input, and cell state at the time step t, respectively. Input weights matrices W_{ce}, W_{ig}, W_{fg} and W_{og} are with shape (n_{ch}, n_{fe_1}). Recurrent weights matrices R_{ce}, R_{ig}, R_{fg}, and R_{og} are of size (n_{fe_1}, n_{fe_1}). Bias matrices B_{ce}, B_{ig}, B_{fg}, and B_{og} are of size (n_{sm}, n_{fe_1}). $\varphi(\cdot)$, $\sigma(\cdot)$, and $\psi(\cdot)$ are non-linear activation functions. The symbol $\odot$ means element-wise multiplication.

For the output Y_{al} in the attention layer, it is split into n_{sp} components according to time steps, i.e., $X_1, X_2, \cdots, X_{n_{sp}}$, with each one being a matrix of size (n_{sm}, n_{ch}). These components form a sequence of $X_1 X_2 \cdots X_{n_{sp}}$ in a chronological orders. For the sequence $X_1 X_2 \cdots X_{n_{sp}}$, the variable X_{in_t} in Eq. (7) has different values in the forward and the backward order. Its value at the time step t in the forward order is X_t, and the value in the backward order is $X_{n_{sp}-t+1}$. Based on Eqs. (7)−(12), a forward output sequence $\mathscr{Y}_{fd}$ is obtained in the forward order, and a backward output sequence $\mathscr{Y}_{bd}$ for the backward order. We use $\mathscr{Y}_{fd}(t)$ to denote the t-th item in the sequence $\mathscr{Y}_{fd}$, i.e., the forward output at the time step t, and $\mathscr{Y}_{bd}(t)$ for the backward output at the time step t. The two output sequences $\mathscr{Y}_{fd}$ and $\mathscr{Y}_{bd}$ are then synthesized as follows:

$$\mathscr{Y}_{blm}(t) = \Phi(\mathscr{Y}_{fd}(t), \mathscr{Y}_{bd}(n_{sp} - t + 1)) \tag{13}$$

Here, $t = 1, \cdots, n_{sp}$. $\Phi(\cdot)$ means an operation, which has two options, i.e., concatenation and summation. $\mathscr{Y}_{blm}$ represents the synthesized sequence of the forward output sequence and the backward output sequence, and $\mathscr{Y}_{blm}(t)$ of size (n_{sm}, n_{fe_2}) means the t-th item in the sequence $\mathscr{Y}_{blm}$, i.e., the output of BiLSTM module at the time step t. n_{fe_2} is a dimensionality of output space of BiLSTM module.

3.2.3. *Time-distributed fully-connected layer*

The time-distributed fully-connected layer is to further extract features at each time step. It executes fully-connected operations separately and simultaneously for inputs at each time step. And the fully-connected operations use linear functions as activation functions. Time-distributed layer could help improve executing efficiency when processing signal data with high sampling frequency. At each time step, the computation procedure is described as follows:

$$\mathscr{Y}_{dl}(t) = \mathscr{Y}_{blm}(t) * W_{dl} + B_{dl}. \tag{14}$$

Here, $t = 1, 2, \cdots, n_{sp}$. Matrix $\mathscr{Y}_{dl}(t)$ of size (n_{sm}, n_{fe_3}), is the output at the time step t in time-distributed fully-connected layer, where n_{fe_3} is a dimensionality of extracted feature space in the time-distributed layer. W_{dl} denotes a weight matrix of size (n_{fe_2}, n_{fe_3}), B_{dl} a bias matrix of size (n_{sm}, n_{fe_3}). All the time-step components $\{\mathscr{Y}_{dl}(t),\ t = 1, \cdots, n_{sp}\}$ compose a matrix Y_{dl} of size $(n_{sm}, n_{sp}, n_{fe_3})$ as the output of the time-distributed fully-connected layer.

3.2.4. *Pooling layer*

The pooling layer in our architecture executes the average pooling operation in order to extract global features of each sample. The operation takes the output matrix Y_{dl} of size $(n_{sm}, n_{sp}, n_{fe_3})$ from the time-distributed fully-connected layer as inputs, computes a mean value of the time-step data for each sample in the matrix Y_{dl}, and outputs a matrix Y_{ap} of size (n_{sm}, n_{fe_3}).

3.2.5. *Fully connected layer and Softmax layer*

Fully connected layer executes a fully connected operation to extract further features and to reduce the last dimension of input

matrix into number of classes. It does not use activation function. A matrix multiplication and a matrix addition are executed in the fully connected layer according to Eq. (15). Computed results from the fully connected layer are passed into the Softmax layer to calculate probabilities that each sample belongs to a class. For the Softmax layer, its computation is described as Eq. (16).

$$Y_{fcl} = Y_{ap} * W_{fcl} + B_{fcl} \tag{15}$$

$$Y_{sl} = softmax(Y_{fcl}) \tag{16}$$

Here, W_{fcl} and B_{fcl} denotes weights matrix of size (n_{fe_3}, n_c) and bias matrix of size (n_{sm}, n_c), respectively. n_c is the number of classes. Y_{fcl} is an output matrix of size (n_{sm}, n_c) in the fully-connected layer. Function $softmax(\cdot)$ calculates probabilities about each sample belonging to each class. Y_{sl} is an output of the Softmax layer.

The pseudo-codes of the proposed seizure/non-seizure classification approach of BiLSTM with attention are shown in Algorithm 1.

Algorithm 1. Seizure/Non-seizure Classification over EEG Data using the Attention BiLSTM Approach

Input: X_0, the matrix of EEG data segments
Output: Y_{pred}, the matrix of classification results

1: Initialize matrices $W_{al}, B_{al}, W_{ce}, W_{ig}, W_{fg}, W_{og}, R_{ce}, R_{ig}, R_{fg}, R_{og}, B_{ce}, B_{ig}, B_{fg}, B_{og}, W_{dl}, B_{dl}, W_{fcl}, B_{fcl}$
2: Compute the output matrix Y_{al} using the input X_0 and Eqs. (1)−(6)
3: Split Y_{al} into n_{sp} components $\{X_1, X_2, \cdots, X_{n_{sp}}\}$ according to time steps, and compose a sequence $X_1 X_2 \cdots X_{n_{sp}}$ in chronological order
4: Compute a forward output sequence $\mathscr{Y}_{fd}$ for the sequence $X_1 X_2 \cdots X_{n_{sp}}$ based on Eqs. (7)−(12)
5: Compute a backward output sequence $\mathscr{Y}_{bd}$ for the inverse sequence $X_{n_{sp}} \cdots X_2 X_1$ based on Eqs. (7)−(12)
6: Synthesize sequences $\mathscr{Y}_{fd}$ and $\mathscr{Y}_{bd}$ by using Eq. (13), and achieve a sequence $\mathscr{Y}_{blm}$
7: Compute a sequence $\mathscr{Y}_{dl}$ by using Eq. (14), and then compose a matrix Y_{dl} according to time steps
8: Compute matrix Y_{ap} by averaging values over time steps for each sample in Y_{dl}
9: Compute matrix Y_{sl} according to Eqs. (15) and (16)
10: Compute the column position of the maximal element in each row of Y_{sl}, and achieve classification results Y_{pred}
11: Return Y_{pred}

4. Evaluation

In this section, we evaluate the approach of BiLSTM with attention by performing cross-validation experiments and cross-patient experiments over the noisy scalp EEG data set of CHB-MIT. Our evaluation mainly adopts three standard metrics, including the sensitivity, the specificity and the precision. The cross-validation experiment is that, data from all the patients are randomly split into three mutually disjoint sets, i.e., training set, validation set and testing set. The training set and validation set are used to train a model, and the testing set is to assess the ability of the trained model. To reduce variability, ten rounds of cross-validation are performed for each seizure/non-seizure classification approach in our experiments. Then, average values and standard deviations over results in the ten rounds are calculated. In a cross-patient experiment, one patient is selected as a testing subject, and all the other patients as training and validation subjects. Data from the training and validation subjects are to train a model, and data from the testing subject are to test the trained model. In our cross-patient experiments, 23 patients in CHB-MIT are separately selected as test subject to assess the performance of our proposed approach, and then the overall performance over the 23 patients is analyzed.

4.1. Data

4.1.1. CHB-MIT data set

The data set CHB-MIT contains 686 EEG recordings from 23 patients of different ages ranging from 1.5 years to 22 years. The recordings include 198 seizures. The used sampling frequency is 256 Hz. Each recording contains a set of EEG signals with different channels. Most recordings are one hour long, and some are for two or four hours. The EEG recordings are grouped into 24 cases and stored in EDF data files. Each EDF file corresponds to an EEG recording. In each case, the signal data were recorded from a single patient. Case Chb21 was obtained 1.5 years after Case Chb01 from the same patient. Each data file contains data over 23 or more channels. There exist data files in which the data over some channels were missing. And some data files, for example, Chb12_27.edf, Chb12_28.edf and Chb12_29.edf, have different channel montages from other seizure files. In our experiments, we did not use the data in the above three EDF files.

4.1.2. Data segmentation

In order to extract effective seizure features, 17 common channels were selected, i.e., for each patient, the data of 17 common channels were used for seizure/non-seizure features extraction. The 17 common channels were P4-O2, FP2-F4, P7-O1, C4-P4, F7-T7, C3-P3, FP1-F7, F8-T8, FZ-CZ, CZ-PZ, F3-C3, T7-P7, P8-O2, FP1-F3, F4-C4, FP2-F8, and P3-O1, respectively. Each data record was split into data segments with the length of 23 seconds from the beginning to the end without overlapping. According to annotation files which mark the starting time and the ending time of each seizure, it could be determined whether a data segment contains a seizure or not. In our experiments, if a segment contained a seizure, it was considered as a seizure segment; otherwise, it was a non-seizure segment. In the seizure segments, the lengths of seizure data varied from 1s to 23s, with the average length being 16.9s. Among all seizure segments, the portion of the seizure signal less than 7s was 14.7%, the part containing more than 10s accounted for 76.1%, and the part containing more than 17s accounted for 59.8%.

As a result of the splitting, 665 seizure segments were obtained. The 665 seizure data segments were taken as a part of our experiment data. For evaluation over a balanced data, 665 non-seizure segments in each experiment were randomly selected from all the non-seizure segments without using random seed.

4.2. Cross-validation seizure/non-seizure classification

The deep learning approach in [20] uses LSTM as a main module (shortly, LSTM approach) to detect seizures. The LSTM approach is evaluated through cross-validation experiments over the EEG data set from University of Bonn [40], showing the state-of-the-art performance. We will compare our approach with the LSTM approach. And also our approach will be compared with a convolutional neural network approach (for short, CNN approach) in [17]. Since the data in Bonn EEG data set is strictly processed, and does not contain any artifacts, and is small in size, we choose to use the noisy CHB-MIT data set for the cross-validation experiments.

The LSTM approach [20] and the CNN approach [17] do not provide all the source codes. Thus, we implemented the two approaches according to their descriptions. The implemented LSTM approach and CNN approach were tested. For the binary classification of A-E in [20], three kinds of cross-validation experiments were performed to evaluate our implemented LSTM approach, including hold-out cross-validation, 10-fold cross-validation, and leave-one-out cross-validation. In the hold-out cross-validation for the A-E classification, a ratio between training data and testing data was set to be the corresponding ratio in [20], i.e., 33.33% for training data, and 66.67% for testing data. Our experimental results were sensitivity of 100.00%, specificity of 100.00%, and accuracy of 100.00%. In the 10-fold cross-validation for the A-E classification, the obtained sensitivity was 99.00%, specificity of 100.00%, and accuracy of 99.50%. In the leave-one-out cross-validation for the A-E classification, the obtained sensitivity, specificity, and accuracy were all 100.00%. For the binary classification of ABCD-E in [20], we conducted hold-out cross-validation experiments using the implemented LSTM approach according to the training-set-size/testing-set-size ratio in [20]. With respect to the three metrics, i.e., sensitivity, specificity, and accuracy, our achieved results were all 100.00%. For the three-class classification problem of A-C-E in [20], a hold-out cross validation was performed with the same data-splitting ratio as in [20]. The received results using our implemented LSTM approach were as follows: average sensitivity of 98.65%, average specificity of 99.34%, and average accuracy of 99.11%. The above experimental results were reaching to or near the reported performances in [20]. By using the same data set (i.e., Bonn EEG data set) and evaluation method as in [17], 10-fold cross-validation was performed for our implemented CNN approach. In the cross-validation experiments, our obtained results were as follows: sensitivity of 99.00%, specificity of 98.50%, accuracy of 98.67%, and precision of 97.06% for normal segments; sensitivity of 98.00%, specificity of 99.00%, accuracy of 98.67%, and precision of 98.00% for preictal segments; sensitivity of 97.00%, specificity of 99.50%, accuracy of 98.67%, and precision of 98.98% for seizure segments; average sensitivity of 98.00%, average specificity of 99.00%, average accuracy of 98.67%, and average precision of 98.01% for the above three classes of segments. The results in our implemented CNN approach were a little better than the reported results in [17]. Then based on the two implementations, we experimented with the CHB-MIT data set to compare them with our proposed approach of attention BiLSTM.

In each cross-validation experiment, all the seizure segments were utilized as a part of experiment data, and non-seizure segments with the same quantity were randomly selected. The training set, validation set and testing set were obtained by randomly splitting the experiment data set according to the ratio 70:15:15. We tuned and determined parameters to achieve the best performance for the three approaches, including the LSTM approach, the CNN approach, and our attention BiLSTM approach. And for each approach, ten cross-validation experiments were carried out based on the correspondingly well-tuned parameters.

For cross-validation experiments using the LSTM approach, our parameters were set as follows: The number of hidden states was 120 in the LSTM layer, that in the time-distributed computing layer was 60, the optimizer was RMSprop, the learning rate was 0.0007, the batch size was 30, and the number of epochs was 30. For the CNN approach in [17], it contains five convolutional layers, five max pooling layers, and three fully connected layers, and its parameters setting in our cross-validation experiments was as follows: The number of hidden states in the first two convolutional layers was 100, that in each of the second two convolutional layers was 200, that in the fifth convolutional layer was 260, that in the first fully connected layer was 100, that in the second fully connected layer was 50, the parameter alpha was 0.01 in the LeakyReLU activation function, the optimizer was Adam, the learning rate was 0.001, the batch size was 30, and the number of epochs was 50. For the proposed approach of BiLSTM with attention, we took a window size of 5888, $tanh(\cdot)$ as cell output activation function, and $sigmoid(\cdot)$ as recurrent activation function. That is, for the two functions $\varphi(\cdot)$ and $\psi(\cdot)$ in Eqs.(7)-(12) their values were $tanh(\cdot)$, and $sigmoid(\cdot)$ as the value of $\sigma(\cdot)$. When tuning hyper-parameters for the proposed approach, we mainly focused on two hyper-parameters, i.e., the number of epoch and the learning rate. According to the training loss curve and the validation loss curve in Fig. 4, we set the number of epoch as 35. For the learning rate, five different values were experimented, and their validation accuracy curves were presented in Fig. 5. The validation accuracy curve corresponding to the learning rate of 0.0013 show better accuracy trend. Our well-tuned parameters in the cross-validation experiments of the proposed approach were as follows: The number of hidden states in the bidirectional LSTM layer was 140, that in the time-distributed layer was 70, the merging mode in the bidirectional LSTM was concatenation, the optimizer was RMSprop, the learning rate was 0.0013, the batch size was 30, and the number of epochs was 35. And the total number of trainable parameters is 197,078. Based on the above parameter settings and the data segmentation on the data set CHB-MIT in Subsection 4.1.2, for each batch of EEG segments, the size of input matrix is (30, 5888, 17) in our model, the attention layer outputs a matrix of size (30, 5888, 17), the LSTM module in the BiLSTM module outputs a matrix of size (30, 5888, 140), the BiLSTM modules output matrix of size (30, 5888, 280), the time-distributed fully-connected layers output matrix of size (30, 5888, 70), the

pooling layers output of size (30, 70), and the output matrix of size (30, 2) for the last fully-connected layer.

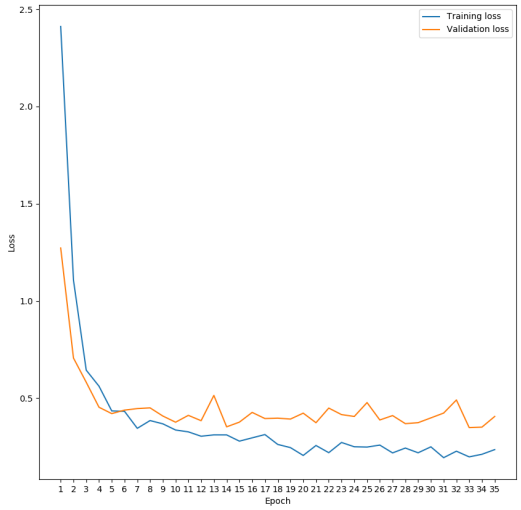


Fig. 4. Training loss curve and validation loss curve for the proposed approach.

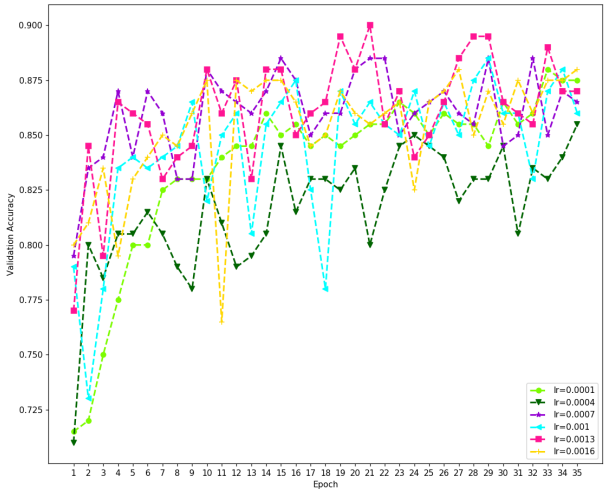


Fig. 5. Validation accuracy trends in five learning rates for the proposed approach.

The cross-validation results using the LSTM approach, including Sensitivity, Specificity, F1 score, Precision, Accuracy, AUC-ROC, the average and the standard deviation, are shown in Table 2. And the results by using the CNN approach and our approach of attention BiLSTM are presented in Tables 3 and 4, respectively.

For the LSTM approach, the achieved average sensitivity, average specificity, average precision and average AUC-ROC are respectively 84.00%, 84.30%, 84.62%, and 0.9151. By using the approach of attention BiLSTM, the obtained average sensitivity of 87.30%, average specificity of 88.30%, average precision of 88.29% and average AUC-ROC of 0.9470 are better

Table 2
Cross-validation results using the LSTM approach.

Iter.	Sens.	Spec.	F1 Sco.	Prec.	Accu.	AUC-ROC
1	0.8800	0.8700	0.8756	0.8713	0.8750	0.9492
2	0.8000	0.9300	0.8556	0.9195	0.8650	0.9158
3	0.8400	0.8600	0.8485	0.8571	0.8500	0.9054
4	0.7900	0.9000	0.8360	0.8876	0.8450	0.9341
5	0.8400	0.8600	0.8485	0.8571	0.8500	0.8982
6	0.8200	0.8800	0.8454	0.8723	0.8500	0.9091
7	0.8700	0.7700	0.8286	0.7909	0.8200	0.9277
8	0.8000	0.8100	0.8040	0.8081	0.8050	0.9074
9	0.8600	0.8300	0.8473	0.8350	0.8450	0.9189
10	0.9000	0.7200	0.8257	0.7627	0.8100	0.8852
Ave.	0.8400	0.8430	0.8415	0.8462	0.8415	0.9151
Std.	0.0355	0.0593	0.0183	0.0450	0.0217	0.0175

Iter. is an abbreviation for Iteration, F1 Sco. for F1 Score, Prec. for Precision, Ave. for Average, and Std. for Standard Deviation. These abbreviations are also used in Tables 3, 4, 5, 6, and 7.

Table 3
Cross-validation results using the CNN approach.

Iter.	Sens.	Spec.	F1 Sco.	Prec.	Accu.	AUC-ROC
1	0.8900	0.7600	0.8357	0.7876	0.8250	0.9153
2	0.8300	0.9500	0.8830	0.9432	0.8900	0.9623
3	0.8700	0.7100	0.8056	0.7500	0.7900	0.8783
4	0.8300	0.8000	0.8177	0.8058	0.8150	0.8764
5	0.8200	0.8100	0.8159	0.8119	0.8150	0.8982
6	0.9000	0.7600	0.8411	0.7895	0.8300	0.9207
7	0.8200	0.8500	0.8325	0.8454	0.8350	0.8938
8	0.7700	0.9100	0.8280	0.8953	0.8400	0.9102
9	0.8100	0.9000	0.8482	0.8901	0.8550	0.9153
10	0.8900	0.6600	0.7982	0.7236	0.7750	0.8895
Ave.	0.8430	0.8110	0.8306	0.8242	0.8270	0.9060
Std.	0.0403	0.0877	0.0229	0.0653	0.0306	0.0239

than the LSTM approach. For the F1 score and accuracy, the approach of attention BiLSTM also exceeds the LSTM approach. And the standard deviations of by the attention BiLSTM approach are mostly less than the LSTM approach. It can be seen that the proposed approach of attention BiLSTM not only classifies seizures more accurately than the LSTM approach, but is also more stable.

For the CNN approach, the obtained average sensitivity, average specificity, average precision and average AUC-ROC are 84.30%, 81.10%, 82.42% and 0.9060, respectively. Our model outperforms the CNN approach in sensitivity, specificity and precision. For the average accuracy and the average F1 score, our approach also has higher values than the CNN approach. And the standard deviations in our method are smaller than the CNN approach. These experimental results show that, the proposed approach of attention BiLSTM has better performance in the seizure/non-seizure classification than the CNN approach.

Ten receiver operating characteristic (ROC) curves are plotted in Fig. 6 for the ten iterations in Table 2, Fig. 7 for the ten iterations in Table 3, and Fig. 8 for the ten iterations in Table 4. The ROC curves in Fig. 8 are closer to the corresponding left-hand borders than the ROC curves in Fig. 6 and in Fig. 7, and also reach to the top borders faster. Our proposed approach discriminates seizures and non-seizures better than the LSTM approach and the CNN approach.

Table 4
Cross-validation results using the proposed approach.

Iter.	Sens.	Spec.	F1 Sco.	Prec.	Accu.	AUC-ROC
1	0.8700	0.8900	0.8788	0.8878	0.8800	0.9319
2	0.9300	0.8500	0.8942	0.8611	0.8900	0.9487
3	0.8800	0.9200	0.8980	0.9167	0.9000	0.9479
4	0.8500	0.9000	0.8718	0.8947	0.8750	0.9332
5	0.9100	0.8200	0.8708	0.8349	0.8650	0.9604
6	0.8500	0.9200	0.8808	0.9140	0.8850	0.9564
7	0.8800	0.8900	0.8844	0.8889	0.8850	0.9554
8	0.8300	0.8500	0.8384	0.8469	0.8400	0.9215
9	0.8800	0.9100	0.8934	0.9072	0.8950	0.9646
10	0.8500	0.8800	0.8629	0.8763	0.8650	0.9497
Ave.	0.8730	0.8830	0.8774	0.8829	0.8780	0.9470
Std.	0.0287	0.0316	0.0168	0.0265	0.0168	0.0131

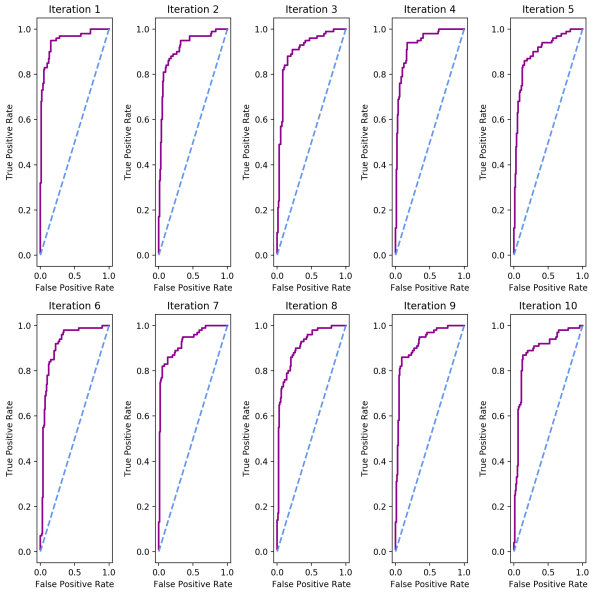


Fig. 6. Ten receiver operating characteristic curves for the ten iterations in Table 2 (the LSTM approach).

4.3. Cross-patient seizure/non-seizure classification

For cross-patient seizure/non-seizure classification, each experiment takes data of one patient as testing data, and other patients' data as training data and validation data according to the ratio 85:15. Because the two cases Chb01 and Chb21 are records from the same patient. The two cases were utilized together either as testing data or training-validation data. In each experiment, all the seizure data segments from each patient were utilized, and non-seizure data segments were randomly selected with the same number of seizure segments. So, the data was balanced in each experiment.

For each patient, we used her/his EEG data as testing data and data of other patients as training-validation data, and obtained the sensitivity, specificity, F1 score, precision, and accuracy. Separately using the LSTM approach, the CNN approach, and the proposed approach, cross-patient experiments were performed. The cross-patient results are listed in Table 5, Table 6, and Table 7, respectively. The performance of the proposed approach reached to an average sensitivity of 83.72%, average

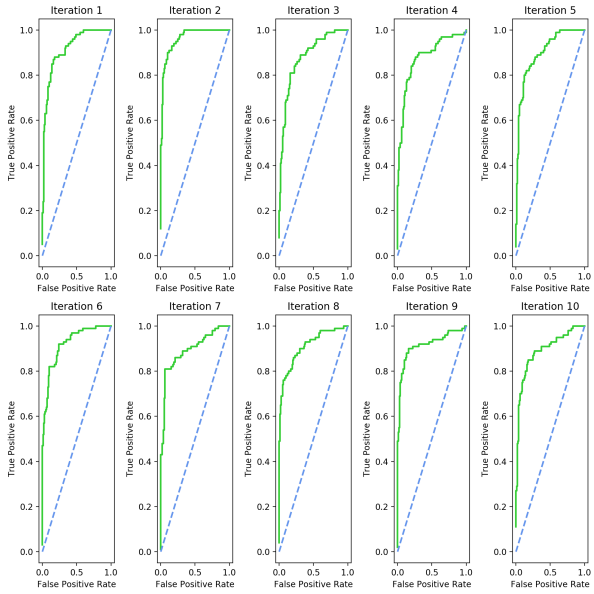


Fig. 7. Ten receiver operating characteristic curves for the ten iterations in Table 3 (the CNN approach).

specificity of 84.06%, average F1 score of 83.63%, average precision of 85.36%, and average accuracy of 83.89%. The overall performance of the proposed approach was better than the LSTM approach and the CNN approach. Compared to the LSTM approach, the average sensitivity and the average specificity of the proposed approach improved with 4.72% and 3.74%, respectively. In comparison with the CNN approach, although the performance of the attention BiLSTM was worse over the case Chb09 , it was better over the most of other 22 cases. The average sensitivity and the average specificity of the attention BiLSTM enhanced 1.79% and 6.63% separately. Also, the standard deviations of sensitivity, specificity, F1 score, precision, and accuracy of the proposed approach were 0.1349, 0.1379, 0.0888, 0.1020, and 0.0833, respectively, which showed that the attention BiLSTM approach was more stable than the two approaches. Fig. 9 shows the sensitivities, the specificities and the precisions in the form of bar chart.

In [4], Thodoroff et al. utilize a recurrent convolutional neural network (recurrent CNN) and obtain an average sensitivity 85% in cross-patient experiments over the CHB-MIT data set. According to Fig. 7(a) and Fig. 7(c) in [4], for six cases Chb06, Chb12, Chb13, Chb14, Chb15 and Chb16, the obtained sensitivity results are not good, only around 20% for Chb06 and Chb14. For other seventeen cases the sensitivity results are mostly 100%. The two cases Chb01 and Chb21 are tested separately for recurrent CNN. Our method achieved better sensitivities in the above cases, all exceeding 50%, although the sensitivity of the remaining cases were less than 100%. Fig. 10 presents the sensitivity comparisons between the method of recurrent CNN and our approach of BiLSTM with attention for the above six cases. And Fig. 11 shows sensitivities of 21 common-tested cases. The 21 cases do not contain Chb01, Chb21 and Chb24. Over the common-tested cases, our standard

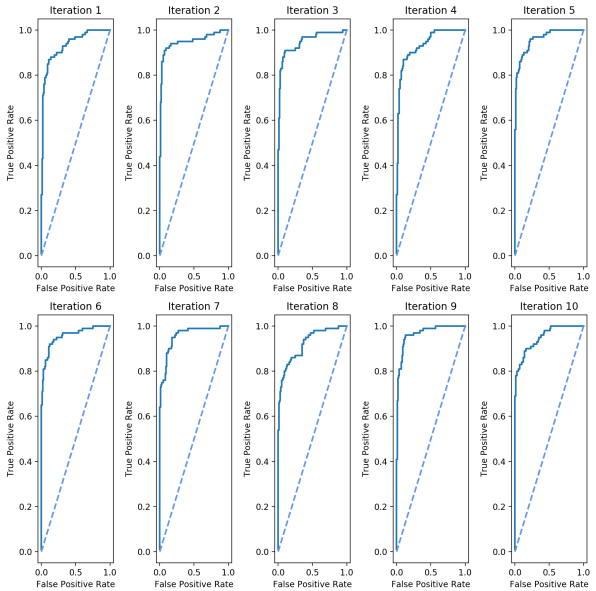


Fig. 8. Ten receiver operating characteristic curves for the ten iterations in Table 4 (the proposed approach).

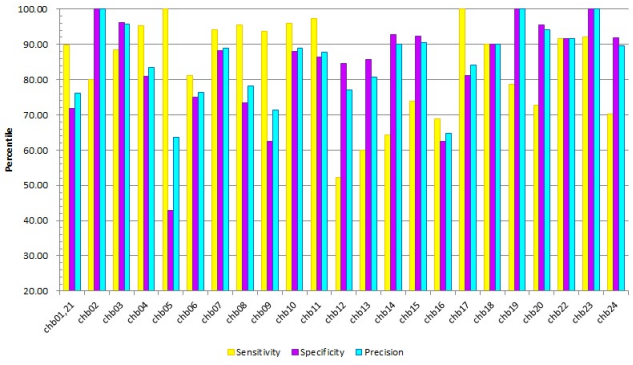


Fig. 9. (Color online) Bar chart illustations of cross-patient sensitivity, specificity and precision over 24 cases for the attention BiLSTM.

deviations for sensitivity and specificity are 0.1374 and 0.1407, respectively. The results indicated that our sensitivity results are more concentrative, and in this sense, the proposed approach of attention BiLSTM is more stable.

Table 1 summarizes existing EEG-based seizure detection approaches, which were evaluated on different EEG data and with different validation methods. The five approaches in Shoeb and Guttag [7], Thodoroff et al. [4], Fergus et al. [21], Amin and Kamboh [8], and Yuan et al. [32], were evaluated on the scalp EEG data set CHB-MIT. [7] and [8] adopted patient-specific validation method. [4], [21] and [32] utilized non-patient-specific validation method. The developed model in [4] is the recurrent CNN. In the above paragraph, we demonstrate that the recurrent CNN is less stable than our proposed approach. For the model in [21], its achieved performance is 88% sensitivity and 88% specificity. Our approach reaches to a sensitivity of 87.3% and specificity of 88.3%, which are comparable to the presented model in [21]. Yuan et al. [32] obtained an accuracy of 94.37% which is greater than our accuracy results. Acharya et al. [17] developed the CNN approach, and Hussein et al. [20] presented the LSTM approach. For the two approaches, Subsection 4.2 and Subsection 4.3 demonstrate that our proposed approach outperforms them.

Table 5
Cross-patient experiment results using the LSTM approach.

Case	Sens.	Spec.	F1 Sco.	Prec.	Accu.
Chb01,21	0.7692	0.8718	0.8108	0.8571	0.8205
Chb02	0.9000	0.9000	0.9000	0.9000	0.9000
Chb03	0.8462	0.9231	0.8800	0.9167	0.8846
Chb04	1.0000	0.8095	0.9130	0.8400	0.9048
Chb05	0.9643	0.1071	0.6750	0.5192	0.5357
Chb06	0.6875	0.9375	0.7857	0.9167	0.8125
Chb07	1.0000	0.8235	0.9189	0.8500	0.9118
Chb08	0.8444	0.7556	0.8085	0.7755	0.8000
Chb09	0.8750	0.7500	0.8235	0.7778	0.8125
Chb10	0.9200	1.0000	0.9583	1.0000	0.9600
Chb11	0.9730	0.8378	0.9114	0.8571	0.9054
Chb12	0.0845	0.9155	0.1446	0.5000	0.5000
Chb13	0.3143	0.7143	0.3929	0.5238	0.5143
Chb14	0.6429	0.2143	0.5294	0.4500	0.4286
Chb15	0.7476	0.7184	0.7368	0.7264	0.7330
Chb16	0.6250	0.6875	0.6452	0.6667	0.6562
Chb17	1.0000	0.8750	0.9412	0.8889	0.9375
Chb18	0.9000	0.9500	0.9231	0.9474	0.9250
Chb19	0.8571	0.9286	0.8889	0.9231	0.8929
Chb20	0.6818	1.0000	0.8108	1.0000	0.8409
Chb22	0.9167	0.9167	0.9167	0.9167	0.9167
Chb23	1.0000	1.0000	1.0000	1.0000	1.0000
Chb24	0.6216	0.8378	0.6970	0.7931	0.7297
Ave.	0.7900	0.8032	0.7831	0.8064	0.7966
Std.	0.2227	0.2185	0.1972	0.1638	0.1596

5. Model analysis

5.1. Interpretations of attention mechanism

Our attention mechanism is designed for distinguishing signals from different brain regions and produces different weights for the signals. In the attention layer, a kernel matrix and a bias matrix are needed, and they are trained together with other modules in our model. Based on the two matrices, the weights of channels, which correspond to different brain regions, are calculated according to the input data. In fact, different epilepsy patients have different seizure patterns and EEG signal is dynamic. For one patient, experienced seizures may have different types and may come from different brain regions. Therefore, it is reasonable to calculate adaptively channel weights in our attention mechanism. Fig. 12 and Fig. 13 show attention weight distributions on 17 channels in two data segments from two patients (i.e., Chb11 and Chb03), which are computed by the attention mechanism in the same trained model. These two figures show that our attention mechanism can adaptively calculate the channel weights of signal data from different patients.

Table 6
Cross-patient experiment results using the CNN approach.

Case	Sens.	Spec.	F1 Sco.	Prec.	Accu.
Chb01,21	0.8205	0.8462	0.8312	0.8421	0.8333
Chb02	0.9000	0.9000	0.9000	0.9000	0.9000
Chb03	0.8846	0.9615	0.9200	0.9583	0.9231
Chb04	0.9524	0.9048	0.9302	0.9091	0.9286
Chb05	0.7143	0.8929	0.7843	0.8696	0.8036
Chb06	0.7500	0.8125	0.7742	0.8000	0.7812
Chb07	0.8235	1.0000	0.9032	1.0000	0.9118
Chb08	0.7778	0.9333	0.8434	0.9211	0.8556
Chb09	0.8750	0.9375	0.9032	0.9333	0.9062
Chb10	0.8800	0.9200	0.8980	0.9167	0.9000
Chb11	0.9730	0.8919	0.9351	0.9000	0.9324
Chb12	0.5211	0.4225	0.4966	0.4744	0.4718
Chb13	0.6571	0.4286	0.5897	0.5349	0.5429
Chb14	0.8571	0.5000	0.7273	0.6316	0.6786
Chb15	0.9903	0.1456	0.6962	0.5368	0.5680
Chb16	0.8125	0.5625	0.7222	0.6500	0.6875
Chb17	0.8125	0.7500	0.7879	0.7647	0.7812
Chb18	0.9000	0.6000	0.7826	0.6923	0.7500
Chb19	0.8571	0.9286	0.8889	0.9231	0.8929
Chb20	0.6818	0.8636	0.7500	0.8333	0.7727
Chb22	0.8333	0.9167	0.8696	0.9091	0.8750
Chb23	0.8400	0.8800	0.8571	0.8750	0.8600
Chb24	0.7297	0.8108	0.7606	0.7941	0.7703
Ave.	0.8193	0.7743	0.8066	0.8074	0.7968
Std.	0.1059	0.2168	0.1088	0.1462	0.1276

Table 7
Cross-patient experiment results using the proposed approach.

Case	Sens.	Spec.	F1 Sco.	Prec.	Accu.
Chb01,21	0.8974	0.7179	0.8235	0.7609	0.8077
Chb02	0.8000	1.0000	0.8889	1.0000	0.9000
Chb03	0.8846	0.9615	0.9200	0.9583	0.9231
Chb04	0.9524	0.8095	0.8889	0.8333	0.8810
Chb05	1.0000	0.4286	0.7778	0.6364	0.7143
Chb06	0.8125	0.7500	0.7879	0.7647	0.7813
Chb07	0.9412	0.8824	0.9143	0.8889	0.9118
Chb08	0.9556	0.7333	0.8600	0.7818	0.8444
Chb09	0.9375	0.6250	0.8108	0.7143	0.7813
Chb10	0.9600	0.8800	0.9231	0.8889	0.9200
Chb11	0.9730	0.8649	0.9231	0.8780	0.9189
Chb12	0.5211	0.8451	0.6218	0.7708	0.6831
Chb13	0.6000	0.8571	0.6885	0.8077	0.7286
Chb14	0.6429	0.9286	0.7500	0.9000	0.7857
Chb15	0.7379	0.9223	0.8128	0.9048	0.8301
Chb16	0.6875	0.6250	0.6667	0.6471	0.6563
Chb17	1.0000	0.8125	0.9143	0.8421	0.9063
Chb18	0.9000	0.9000	0.9000	0.9000	0.9000
Chb19	0.7857	1.0000	0.8800	1.0000	0.8929
Chb20	0.7273	0.9545	0.8205	0.9412	0.8409
Chb22	0.9167	0.9167	0.9167	0.9167	0.9167
Chb23	0.9200	1.0000	0.9583	1.0000	0.9600
Chb24	0.7027	0.9189	0.7879	0.8966	0.8108
Ave.	0.8372	0.8406	0.8363	0.8536	0.8389
Std.	0.1349	0.1379	0.0888	0.1020	0.0833

In some areas of the brain, EEG signals during seizures show many differences with signals at non-seizures. The differences, such as frequency and magnitude, could be used to indentify seizure and non-seizure. The attention mechanism captures signal characteristics and assigns large weight values to the channels, which could distinguish seizure and non-seizure segments. In our experiments, it was observed that relatively large weights were assigned to channels with great differences between seizure signals and non-seizure signals. An example of attention weights of 17 channels for a seizure segment is shown in Fig. 12; the channels of F8-T8, P3-O1 and FP2-F8 have the large weights compared to other channels. In Fig. 14(a) and Fig. 14(b), the actual signals over the above three channels change (i.e., six purple panels) much in the rate of change of magnitude. For the actual signals over channels P4-O2 and P8-O2 (i.e., four green panels), the differences of magnitude change rate between Fig. 14(a) and Fig. 14(b) are relatively small. A rate of change of amplitude is computed on 128 successive data points in a signal on a channel, and it is a difference between a maximal data point and a minimal data point divided by a duration. In an average rate of change of amplitude, differences between seizure segment and non-seizure segment on five channels in Fig. 14 are as follows: $396.84\mu V/s$ for Channel F8-T8, $314.01\mu V/s$ for Channel P3-O1, $461.12\mu V/s$ for Channel FP2-F8, $193.60\mu V/s$ for Channel P4-O2, and $182.01\mu V/s$ for Channel P8-O2. As shown in Fig. 12, the assigned weights over Channel P4-O2 and Channel P8-O2 are small.

The actual signals of the channel P4-O2 (i.e., two green panels) in Fig. 15(a) and Fig. 15(b) manifest small differences in magnitudes. A difference of rate of change of amplitude between seizure segment and non-seizure segment on Channel P4-O2 in Fig. 15 is $380.82\mu V/s$. The attention mechanism produces small weight for the channel P4-O2 so that the corresponding signal data is not treated as critical evidences to classify seizure/non-seizure. The signals over channels T7-P7, FP2-F8 and P3-O1 (i.e., six purple panels) change a lot from the non-seizure Fig. 15(a) to the seizure Fig. 15(b). Such changes could differentiate seizure/non-seizure segments. With respect to the rate of change of amplitude, differences between seizure segment and non-seizure segment on three channels in Fig. 15 are as follows: $1075.52\mu V/s$ for Channel T7-P7, $748.56\mu V/s$ for Channel FP2-F8, and $427.38\mu V/s$ for Channel P3-O1. So, the three channels are assigned large attention weights, as shown in Fig. 13.

5.2. Validations of BiLSTM and attention mechanism

The approach of attention BiLSTM is developed in the inspiration of the LSTM approach in [20]. In the development, the performances of bidirectional LSTM and attention mechanism are separately explored. By using parameters with the best performances in the tuning procedures, ten rounds of cross-validation experiments are performed separately for testing the two modules. When testing the module of bidirectional LSTM, the parameters are set as follows: The learning rate is 0.001,

Table 8
Cross-validation results for modules in the attention BiLSTM approach.

Module	Sensitivity	Specificity	F1 Score	Precision	Accuracy
Bidirectional LSTM	0.8630±0.06	0.8280±0.05	0.8477±0.01	0.8373±0.03	0.8455±0.01
Attention LSTM	0.8340±0.05	0.8870±0.04	0.8564±0.02	0.8828±0.03	0.8605±0.02
Attention BiLSTM	0.8730±0.0287	0.8830±0.0316	0.8774±0.0168	0.8829±0.0265	0.8780±0.0168

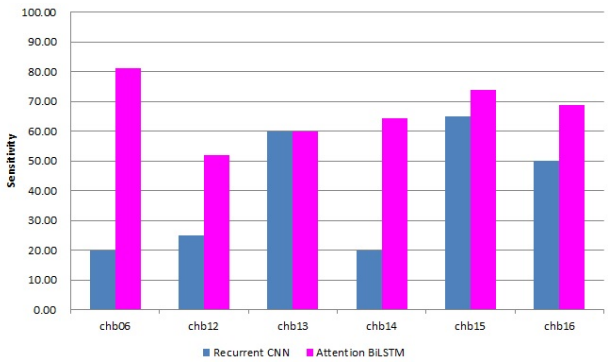


Fig. 10. (Color online) Comparison of cross-patient sensitivity over 6 cases between attention BiLSTM and recurrent CNN.

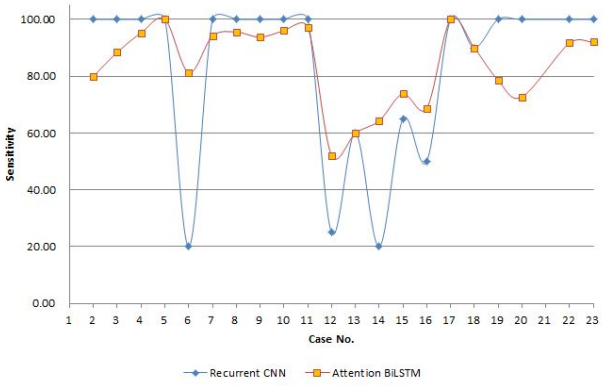


Fig. 11. (Color online) Comparison of cross-patient sensitivity over 21 common cases between attention BiLSTM and recurrent CNN.

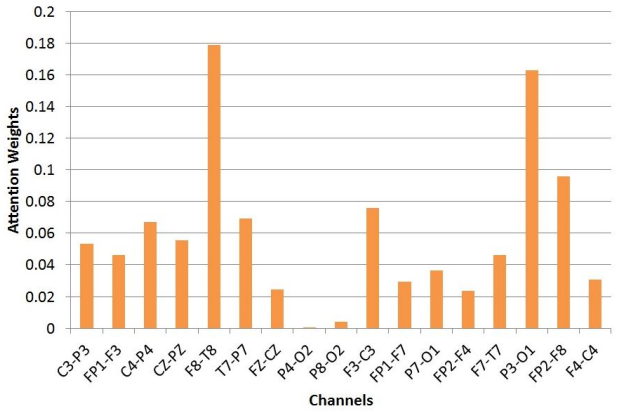


Fig. 12. Attention weights on channels for a seizure segment in Chb11.

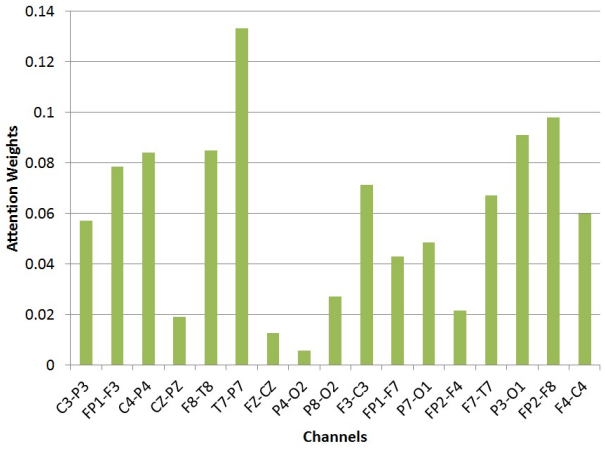
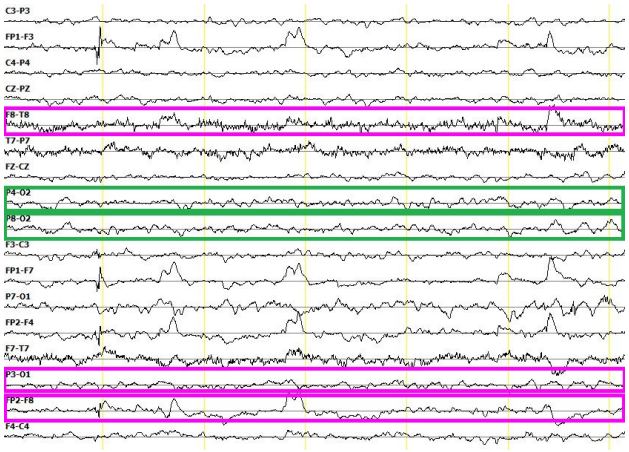


Fig. 13. Attention weights on channels for a seizure segment in Chb03.

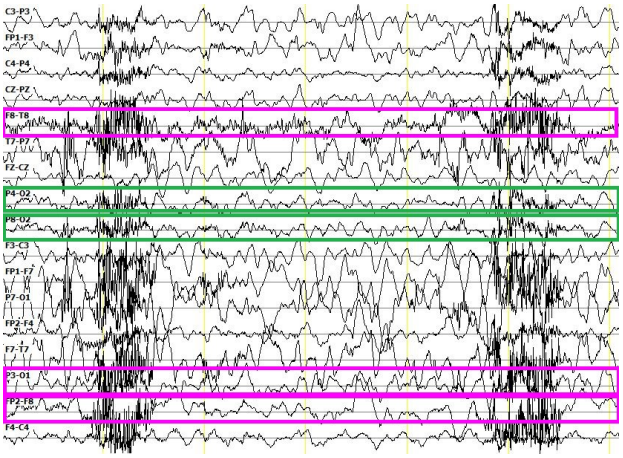
the number of hidden states in the bidirectional LSTM is 100, that in time-distributed layer is 50, the optimizer is RMSprop, batch size is 30, and the number of epochs is 30. For the testing of attention mechanism, the parameters are: The learning rate is 0.001, the number of hidden states in the module LSTM is 100, that in time-distributed layer is 50, the optimizer is RMSprop, batch size is 30, and the number of epochs is 25. The obtained cross-validation results are shown in Table 8. The results indicate that in comparisons with the LSTM approach results in Table 2, the bidirectional LSTM obtains better sensitivity but worse specificity, and the attention LSTM achieves greater specificity but a little smaller sensitivity. Only using the bidirectional LSTM module or the attention LSTM module does not absolutely improve the performance. After combining the two modules in the approach of attention BiLSTM, both the sensitivity and the specificity are improved with 3.3% and 4%, respectively. Thus, both bidirectional LSTM and attention mechanism play important roles in the approach of attention BiLSTM for seizure/non-seizure classification. They are not redundant in the attention BiLSTM. The bidirectional LSTM is deployed to learn temporal information in the EEG data, and the attention mechanism is to extract spatial information in the multi-channel signal data.

6. Discussion

In this paper, we design a novel approach of BiLSTM with attention for seizure/non-seizure classification in off-line EEG data. Cross-patient and cross-validation experiments across pa-

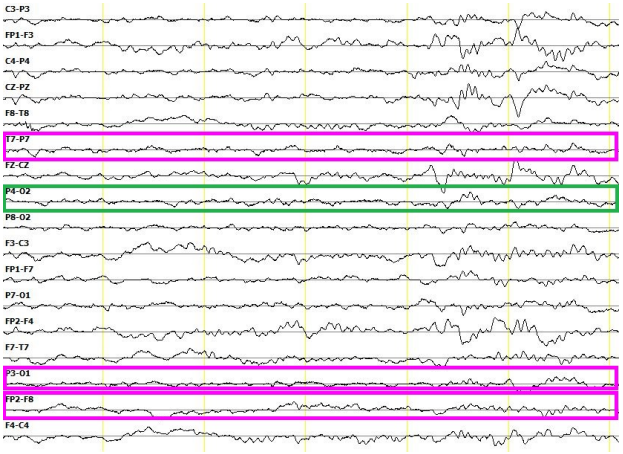


(a) Signals in a non-seizure segment from Chb11.

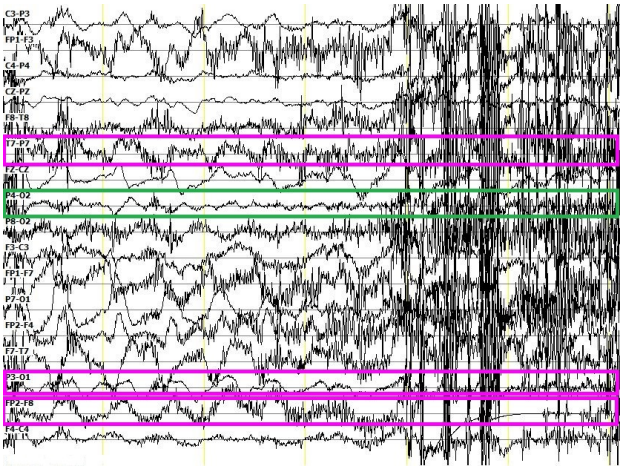


(b) Signals in a seizure segment from Chb11.

Fig. 14. (Color online) Visualization of signals on channels in a non-seizure segment and a seizure segment from Chb11. Purple panels represent channels with large signal changes, and green panels for channels with small signal changes.



(a) Signals in a non-seizure segment from Chb03.



(b) Signals in a seizure segment from Chb03.

Fig. 15. (Color online) Visualization of signals on channels in a non-seizure segment and a seizure segment from Chb03. The purple panels and green panels have the same meanings as in Fig. 14.

tients are separately applied to evaluations on the pediatric data set of CHB-MIT. When doing segmentation, a time length of 23 seconds is selected by referring to the segment length in Bonn EEG data set [40], and each data record in each case is split from the beginning to the end without overlapping. As a result, 665 seizure segments are obtained, and the lengths of seizure data vary from 1s to 23s in seizure segments. The length diversity of seizure data is aligned with a real-world situation. In each experiment, the 665 seizure segments were taken as a part of experimental data, and 665 non-seizure segments were randomly selected from the extracted non-seizure segments. Its randomness and sparsity reduce temporal correlations among non-seizure data segments, and avoid resulting in overly optimistic specificity results [7]. The above segmenting-data-record way and the selecting strategy of non-seizure segments make the evaluation of our approach be more reliable.

In the cross-validation experiments, the sensitivity, specificity and precision of our approach were better than the LSTM approach in [20] and the CNN approach in [17]. The improvements in the sensitivity, specificity, and precision over those two state-of-the-art approaches were 3.3%, 4%, 3.67% and 3%, 7.2%, 5.87%, respectively, and the standard deviations were less than the two approaches in comparison. As Table 8 shows, the better performances of our approach are attributed to the attention mechanism and the feature extraction in both forward and backward directions.

Among cross-patient experiment results in Table 7, there exist gaps. Over the six patients, including Chb05, Chb09, Chb12, Chb13, Chb14, and Chb16, either sensitivity or specificity were less than 70%. For the seven patients, i.e., Chb03, Chb07, Chb10, Chb11, Chb18, Chb22 and Chb23, all testing results were over 85%. The possible reason is that, for a child, the brain, meninges, skull, and head size change overtime [41]. Compared to the method of recurrent CNN proposed in [4], the performances of our method BiLSTM with attention were more stable. In [4], the convolution neural network module in recurrent CNN is pre-trained before training the whole model. Our attention BiLSTM approach does not need pre-training, and it directly processes raw data and extracts features. The REVEAL algorithm proposed in [42] achieved an average sensitivity of 61%. [5] used the automatic seizure detection system EpiScan

on the CHB-MIT data set and obtained an average sensitivity of 67%. The average sensitivity of our approach is much better than REVEAL and EpiScan.

It was explored that whether increasing the number of seizure segments by allowing overlaps among segments could help improve the performance of seizures/non-seizures classification or not. The EEG signals in the dataset CHB-MIT were segmented according to a segment length of 23s and overlapping length of 5s. In such a segmentation way, 848 seizure segments were obtained, which are more than the 665 seizure segments without overlaps. In order to make evaluation on a balanced data set, 848 non-seizure segments were randomly selected. Over the produced 1696 segments, cross-validation experiments were performed by using our attention BiLSTM approach. After tuning hyper-parameters well, ten cross-validations were conducted. The obtained results were an average sensitivity of 86.95%, average specificity of 88.91%, average F1 score of 87.80%, average precision of 88.81%, average accuracy of 87.93%, and average AUC-ROC of 0.9414. The improvements are few. Maybe, there exist two possible reasons. One is that the EEG data segments for the training are still not many enough to achieve improvements. The other is that overlaps among segments could cause over-fitting problem such that the training accuracy is excellent and the testing results are bad.

The application scenario of our approach is to automatically select all the seizure segments from the off-line EEG data records for neurologists analyses and to remove non-seizure segments from neurologists work. The automatic way aims to help neurologists reduce workloads and increase their productivities. Because of the off-line EEG data segments, extracting features in the forward direction and the backward direction and performing analyses are feasible in practices. In the application, those classified as seizure segments are sent to neurologists to make analyses, and those classified as non-seizure segments are out of neurologists analyses. The automatic classification of true negative segments (i.e., true non-seizure segments) reduces neurologists workloads. False negative segments (i.e., false non-seizure segments) are removed from neurologists analysis list, but this is not what neurologists expect. So, selecting as many seizure segments as possible and as accurately as possible is the target in the application. Metrics, such as sensitivity, specificity, and precision, are used to measure the performance of automatic seizure/non-seizure classification method in the application. For the metric of temporal false alarm rate, it means the number of samples that are falsely classified as being positive in a time unit. The time unit may be either one hour or one day. Our application focuses on that how many seizure segments are successfully selected and that how many non-seizure segments are correctly classified in total. The false seizure segments in the application do not increase neurologists workloads essentially. For example, a model in the application is with a false alarm rate of 10 per hour, a specificity of 80.00%, and a sensitivity of 80.00%. The example of model produces 240 false seizure segments one day, and can correctly select 80.00% non-seizure segments and 80.00% seizure segments. Although the number of produced false seizure segments is large, the work of reviewing the 80.00% non-seizure segments is not needed. The reduced workloads are much more. So in the application, the temporal false alarm rate is not taken as a metric to evaluate the seizures/non-seizures classification models.

Instead of directly training weights on channels, we utilize an attention mechanism to generate weights. In the directly training way, the obtained weights on channels are the same for all the patients. In fact, the seizure patterns of different patients are different, and different types of seizures have different patterns, and it is possible that one patient may have different types of seizures. Therefore, for data segments from different patients, the weights on channels, which describe the relative strength that signals signify seizures, need be different. In our attention mechanism, a kernel matrix and a bias matrix are obtained by training, and then the two trained matrices are performed transformations by combining with data segments. The outputs of transformations are attention weights for the data segments. The attention mechanism produces different weights for data segments from different patients. When evaluating our model on the data set CHB-MIT, the function $softmax(\cdot)$ is adopted in the attention mechanism, and the sum of weights on channels is 1 for each data segment. A channel weight represents relative strength about that characteristics of signals on the corresponding channel signifies seizures. A channel weight close to 0 indicates that corresponding signal characteristics are relatively weak to signify seizures. It does not imply that the channel has no contribution to seizure.

When designing attention mechanism, we tried different ways: one way is adding different attention weights over time steps, and another way is adding different attention weights over time steps and over channels. Our experimental results using the two ways were not good. One possible reason is that the role of each brain region in the whole brain state is generally stable in a short duration such as 23s. Finally, we choose to apply attention mechanism to channels and share the attention weights among time steps. Actually, different channels have different contributions to a seizure, and the contributions turn out to be correlated to the locations of brain regions, rather than the time. In addition, we applied our method to single channel data. The results with single channel data were not good. They were in agreement with the observation in [7]; that is, for some channels, the data morphology in seizure state is similar to that in non-seizure state.

7. Conclusions

This paper focuses on the problem of automatic seizure/non-seizure classification. Inspired by the architecture in [20], we analyze both spatial and temporal characteristics of seizures, and propose a novel deep learning-based approach by using the model of BiLSTM integrated with attention. The integration of an attention mechanism is to capture spatial features better, and the employment of the BiLSTM model is to extract more temporal features. The proposed approach is evaluated on the noisy EEG data set of CHB-MIT. The evaluation is across multiple patients and uses data from multiple brain regions. In the cross-validation experiments, we obtain sensitivity of 87.3%,

specificity of 88.3% and precision of 88.29%, which are better than the LSTM approach in [20] and the CNN approach in [17]. In the cross-patient experiments, the testing results are 83.72%-sensitivity, 84.06%-specificity and 85.35%-precision on average. Comparing to the model recurrent CNN in [4], our model BiLSTM with attention is more stable.

In the approach of BiLSTM with attention, the pooling layer adopts a globally-averaging way to extract holistic features of data segments. The problem whether such a way is the best or not for the seizure/non-seizure classification will be explored in the future. And also we want to investigate whether the length of data segments has effects on the sensitivity, the specificity and the precision.

References

[1] I. Megiddo, A. Colson, D. Chisholm, T. Dua, A. Nandi, R. Laxminarayan. Health and economic benefits of public financing of epilepsy treatment in India: An agent-based simulation model. Epilepsia, vol. 57, no. 3, pp. 464-474, 2016. https://doi.org/10.1111/epi.13294.

[2] J. Gotman, J. R. Ives, P. Gloor. Automatic recognition of inter-ictal epileptic activity in prolonged EEG recordings. Electroencephalography and Clinical Neurophysiology, vol. 46, no. 5, pp. 510-520, 1979. https://doi.org/10.1016/0013-4694(79)90004-X.

[3] J. Gotman. Automatic recognition of epileptic seizures in the EEG. Electroencephalography and Clinical Neurophysiology, vol. 54, no. 5, pp. 530-540, 1982. https://doi.org/10.1016/0013-4694(82)90038-4.

[4] P. Thodoroff, J. Pineau, A. Lim. Learning robust features using deep learning for automatic seizure detection. Proceedings of the 1st Machine Learning for Healthcare Conference; Los Angeles, CA, USA; 2016. Journal of Machine Learning Research, vol. 56, pp. 178-190, 2016.

[5] F. Fürbass, P. Ossenblok, M. Hartmann, H. Perko, A. M. Skupch, G. Lindinger, L. Elezi, E. Pataraia, A. J. Colon, C. Baumgartner, T. Kluge. Prospective multi-center study of an automatic online seizure detection system for epilepsy monitoring units. Clinical Neurophysiology, vol. 126, no. 6, pp. 1124-1131, 2015. https://doi.org/10.1016/j.clinph.2014.09.023.

[6] A. S. Zandi, M. Javidan, G. A. Dumont, R. Tafreshi. Automated real-time epileptic seizure detection in scalp EEG recordings using an algorithm based on wavelet packet transform. IEEE Transactions on Biomedical Engineering, vol. 57, no. 7, pp.1639-1651, 2010. https://doi.org/10.1109/TBME.2010.2046417.

[7] A. Shoeb, J. Guttag. Application of machine learning to epileptic seizure detection. Proceedings of the 27th International Conference on Machine Learning; pp. 975-982; Haifa, Israel; 2010.

[8] S. Amin, A. M. Kamboh. A robust approach towards epileptic seizure detection. Proceedings of IEEEE 26th International Workshop on Machine Learning for Signal Processing; pp. 1-6; Salerno, Italy; 2016.

[9] M. Fan, C. Chou. Detecting abnormal pattern of epileptic seizures via temporal synchronization of EEG signals. IEEE Transactions on Biomedical Engineering, vol. 66, no. 3, pp. 601-608, 2019. https://doi.org/10.1109/TBME.2018.2850959.

[10] B. Hunyadi, M. Signoretto, W. V. Paesschen, J. A. K. Suykens, S. V. Huffel, M. D. Vos. Incorporating structural information from the multichannel EEG improves patient-specific seizure detection. Clinical Neurophysiology, vol. 123, no. 12, pp. 2352-2361, 2012. https://doi.org/10.1016/j.clinph.2012.05.018.

[11] A. V. Esbroeck, L. Smith, Z. Syed, S. Singh, Z. Karam. Multitask seizure detection: Addressing intra-patient variation in seizure morphologies. Machine Learning, vol. 102, no. 3, pp. 309-321, 2016. https://doi.org/10.1007/s10994-015-5519-7.

[12] N. D. Truong, L. Kuhlmann, M. R. Bonyadi, J. Yang. Supervised learning in automatic channel selection for epileptic seizure detection. Expert Systems with Applications, vol. 86, pp. 199-207, 2017. https://doi.org/10.1016/j.eswa.2017.05.055.

[13] L. Vidyaratne, A. Glandon, M. Alam, K. M. Iftekharuddin. Deep recurrent neural network for seizure detection. Proceedings of 2016 International Joint Conference on Neural Networks; pp. 1202-1207; Vancouver, BC, Canada; 2016.

[14] H. Qu, J. Gotman. Improvement in seizure detection performance by automatic adaptation to the EEG of each patient. Electroencephalography and Clinical Neurophysiology, vol. 86, no. 2, pp. 79-87, 1993. https://doi.org/10.1016/0013-4694(93)90079-B.

[15] W. Zhou, Y. Liu, Q. Yuan, X. Li. Epileptic seizure detection using lacunarity and Bayesian linear discriminant analysis in intracranial EEG. IEEE Transactions on Biomedical Engineering, vol. 60, no. 12, pp. 3375-3381, 2013. https://doi.org/ 10.1109/TBME.2013.2254486.

[16] M. Golmohammadi, S. Ziyabari, V. Shah, E. V. Weltin, C. Campbell, L. Obeid, J. Picone. Gated recurrent networks for seizure detection. Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium; pp. 1-5; Philadelphia, Pennsylvania, USA; 2017.

[17] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. Computers in Biology and Medicine, vol. 100, no. 1, pp. 270-278, 2018. https://doi.org/10.1016/j.compbiomed.2017.09.017.

[18] Y. Yuan, G. Xun, F. Ma, Q. Suo, H. Xue, K. Jia, A. Zhang. A novel channel-aware attention framework for multi-channel EEG seizure detection via multi-view deep learning. Proceedings of the 2018 IEEE EMBS International Conference on Biomedical & Health Informatics; pp. 206-209; Las Vegas, USA; 2018.

[19] X. Zhang, L. Yao, M. Dong, Z. Liu, Y. Zhang, Y. Li. Adversarial representation learning for robust patient-independent epileptic seizure detection. arXiv prepint, arXiv:1909.10868, 2019.

[20] R. Hussein, H. Palangi, R. Ward, Z. J. Wang. Epileptic seizure detection: A deep learning approach. arXiv prepint, arXiv:1803.09848v1, 2018.

[21] P. Fergus, A. Hussain, D. Hignett, D. A. Jumeily, K. A. Aziz, H. Hamdan. A machine learning system for automated whole-brain seizure detection. Applied Computing and Informatics, vol. 12, no. 1, pp. 70-89, 2016. https://doi.org/10.1016/j.aci.2015.01.001.

[22] N. Nicalaou, J. Georgiou. Detection of epileptic electroencephalogram based on permutation entropy and support vector machines. Expert Systems with Applications, vol. 39, no. 1, pp. 202-209, 2012. https://doi.org/10.1016/j.eswa.2011.07.008.

[23] S. Nasehi, H. Pourghassem. Patient-specific epileptic seizure onset detection algorithm based on spectral features and IPSONN classifier. Proceedings of 2013 International Conference on Communication Systems and Network Technologies; pp. 186-190; Gwalior, India; 2013.

[24] A. Kharbouch, A. Shoeb, J. Guttag, S. S. Cash. An algorithm for seizure onset detection using intracranial EEG. Epilepsy & Behavior, vol. 22, no. 1, pp. S29-S35, 2011. https://doi.org/10.1016/j.yebeh.2011.08.031.

[25] Y. Zheng, J. Zhu, Y. Qi, X. Zheng, J. Zhang. An automatic patient-specific seizure onset detection method using intracranial electrocephalography. Neuromodulation, vol. 18, no. 2, pp. 79-84, 2015. https://doi.org/10.1111/ner.12214.

[26] U. R. Acharya, F. Molinari, S. V. Sree, S. Chattopadhyay, K.-H. Ng, J. S.Suri. Automated diagnosis of epileptic EEG using entropies. Biomedical Signal Processing and Control, vol. 7, no. 4, pp. 401-408, 2012. https://doi.org/10.1016/j.bspc.2011.07.007.

[27] M. E. Saab, J. Gotman. A system to detect the onset of epileptic seizures in scalp EEG. Clinical Neurophsiology, vol. 116, no. 2, pp. 427-442, 2005. https://doi.org/10.1016/j.clinph.2004.08.004.

[28] L. Kuhlmann, A. N. Burkitt, M. J. Cook, K. Fuller, D. B. Grayden, I. Seiderer, I. M. Y. Mareels. Seizure detection using seizure probability estimation: Comparison of features used to detect seizures. Annals of Biomedical Engineering, vol. 37, no. 10, pp. 2129-2145, 2009. https://doi.org/10.1007/s10439-009-9755-5.

[29] G. Wang, Z. Sun, R. Tao, K. Li, G. Bao, X. Yan. Epileptic seizure detection based on partial directed coherence analysis. IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 3, pp. 873-879, 2016. https://doi.org/10.1109/JBHI.2015.2424074.

[30] A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. Vos, S. V. Huffel. Neonatal seizure detection using deep convolutional neural networks. International Journal of Neural Systems, vol. 28, no. 0, 1850011, 2018. https://doi.org/10.1142/S0129065718500119.

[31] Y. Yuan, K. Jia. FusionAtt: Deep fusional attention networks for multi-channel biomedical signals. Sensors, vol. 19, no. 11, pp. 2429, 2019. https://doi.org/10.3390/s19112429.

[32] Y. Yuan, G. Xun, K. Jia, A. Zhang. A multi-view deep learning framework for EEG seizure detection. IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 1, pp. 83-94, 2019. http-

s://doi.org/10.1109/JBHI.2018.2871678.

[33] S. Kiranyaz, T. Ince, M. Zabihi, D. Ince. Automated patient-specific classification of long-term electroencephalography. Journal of Biomedical Informatics, vol. 49, pp. 16-31, 2014. https://doi.org/10.1016/j.jbi.2014.02.005.

[34] I. Obeid, J. Picone. The temple university hospital eeg data corpus. Frontiers in Neuroscience, vol. 10, pp. 196, 2016. https://doi.org/10.3389/fnins.2016.00196.

[35] H. Hassanpour, M. Shahiri. Adaptive segmentation using wavelet transform. Proceedings of 2007 International Conference on Electrical Engineering; pp.1-5; Lahore, Pakistan; 2007.

[36] M. Schuster, K. K. Paliwal. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, 1997. https://doi.org/10.1109/78.650093.

[37] A. Graves, J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, vol. 18, no. 5-6, pp. 602-610, 2005. https://doi.org/10.1016/j.neunet.2005.06.042.

[38] F. A. Gers, J. Schmidhuber, F. Cummins. Learning to forget: Continual prediction with LSTM. Neural Computation, vol. 12, no. 10, pp. 2451-2471, 2000. https://doi.org/10.1162/089976600300015015.

[39] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber. LSTM: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222-2232, 2017. https://doi.org/10.1109/TNNLS.2016.2582924.

[40] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, C. E. Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Physical Review E, vol. 64, 061907, 2001. https://doi.org/10.1103/PhysRevE.64.061907.

[41] P. J. Holt. Introduction to pediatric EEG. Atlanta: Emory University School of Medicine; c2017. Available from: https://www.pediatrics.emory.edu/divisions/neurology/education/pedeeg.html

[42] S. B. Wilson, M. L. Scheuer, R. G. Emerson, A. J. Gabor. Seizure detection: Evaluation of the Reveal algorithm. Clinical Neurophysiology, vol. 115, no. 10, pp. 2280-2291, 2004. https://doi.org/10.1016/j.clinph.2004.05.018.