Predict Epileptic Episodes from EEG Time Series with Deep Neural Networks

Liang Ma
Department of Biomedical Engineering
Columbia University
New York, NY, 10027
lm3397@columbia.edu

Mara Kaspers
Department of Biomedical Engineering
Columbia University
New York, NY, 10027
mkk2167@columbia.edu

***Abstract*—Identifying seizure discharges in electroencephalogram (EEG) is an early and crucial step of clinical diagnosis of epilepsy by neurologists. The manual review of an EEG for seizure detection is a laborious and error-prone process. Therefore automated seizure detection has been extensively studied. In recent years, deep learning techniques has been adopted in order to avoid manual feature extraction and selection. In this short project, we cursively compare the performance of two different types of deep neural networks on predicting epileptic episodes using EEG raw time series. We use a relatively small and publicly accessible dataset of surface EEG recorded from human subjects. Data is organized and sectioned into EEG epochs labeled into 5 categories, including seizure activity. We designed two tasks based on the data; one is classifying seizure activity against normal activity, and the other is classifying seizure against all other categories. A convolutional neural network (CNN) and a bidirectional long short term memory network (BiLSTM) are implemented to classify the inputs. The classification result for test data set showed that the LSTM, with accuracy above 0.97, greatly outperforms the CNN.**

I. Introduction

Epilepsy is a disease of the brain defined by recurrent unprovoked seizures and related syndromes[1]. The primary tool for seizure detection in a clinical setting is the electroencephalogram (EEG). Identifying seizure discharges in EEG is an early and crucial step of clinical diagnosis of epilepsy by neurologists. EEG continuously measures the electrical activity of the brain via electrodes placed on the scalp of the brain. It holds the advantage of being accessible, non-invasive, and able to record from large cortical areas. Manual inspection of long, continuous EEG for seizure detection is a time consuming and laborious process in both clinical and experimental settings. It can take many hours to meticulously examine days of EEG recordings for patients hospitalized to diagnose epilepsy. In an experimental setting, long-term EEG recordings (even up to several months) are often to be reviewed. Furthermore, the EEG readings made by different inspectors can be inconsistent as the criteria for abnormal EEG findings are experiential. Aside from manual annotation by experts with domain knowledge, classical methods include modeling signal characteristics such as time-frequency and power analyses. However these methods are also subject to high variations due to experimental setup, quality of recordings, and analysis pipelines.

Researchers have employed various machine learning techniques to automatically detect seizures[2]. Yet the extreme variability existing in EEG from different patients and sometimes the same patient causes significant difficulty in automatic detection[3]. Another source of difficulty is that EEG signals are highly non-stationary and nonlinear[4]. A generalized seizure detector requires extraction of discriminative features between seizure and non-seizure EEGs. Traditional methods include hand-engineered EEG feature extraction in the time domain, frequency domain, time-frequency domain, or combinations of these domains. Time domain features include amplitude and duration of waveform as well as their variation coefficients. Frequency domain features can be spectral characteristics from Fast Fourier Transform (FFT) or periodogram. Time-frequency decomposition usually comes from Short Time Fourier Transform (STFT) or wavelet transform. These features are extracted, statistically analyzed, ranked, and selected for classification. Combinations of multiple domain features are commonly used in the classification process. Classification methods include logistic regression, support vector machine (SVM), k-nearest neighbor, and more recently t-stochastic neighborhood embedding (t-SNE), as well as neural networks. In general, traditional approaches have two processes, feature extraction and classification. Both the identification of the appropriate features and the choice of a proper classifier can play important roles in optimizing algorithm performance. These processes depend heavily on domain expertise and consume a great deal of time and effort to select proper features and classifiers.

We have learned in class that deep learning approaches can automatically discover and learn discriminative features needed for the classification of inputs[5]. Recently, many studies have investigated deep learning for seizure detection. These studies have been based on different deep neural network structures, such as a fully connected neural network (FCNN)[6], convolutional neural network (CNN)[7], and recurrent neural network (RNN)[8]. We learned that CNNs and RNNs are especially good at processing biomedical signals and sequences. In order to take advantage of automatic feature discoveries we used as input only the raw EEG signals segmented into epochs. Using these time series, we trained, validated, tested and compared two models. We find that, for

the given dataset, a bidirectional long short-term memory network performs better and more consistent than a convolutional neural network does.

II. Materials and Methods

A. Data

Data is downloaded from kaggle.com, an open access, web-based data science community for data and code sharing. The data originally comes from University of California Irvine Machine Learning Repository[9]. The EEG data itself are multivariate time series with real and integer values.

It is organized into 5 different folders, each with 100 files, with each file representing a single subject/person. Each file is a recording of brain activity for 23.6 seconds. The corresponding time-series is sampled into 4097 data points. Each data point is the value of the EEG recording at a different point in time. So there are in total 500 individuals with each having 4097 data points for 23.5 seconds.

The original authors divided and shuffled every 4097 data points into 23 chunks, each chunk contains 178 data points for 1 second, and each data point is the value of the EEG recording at a different point in time. So now we have 23 x 500 = 11500 pieces of information(row), each information contains 178 data points for 1 second(column), the last column represents the label y 1,2,3,4,5.

The response variable is y in column 179, the Explanatory variables X1, X2, ..., X178.

y contains the category of the 178-dimensional input vector. Specifically y in 1, 2, 3, 4, 5:

1 - Recording of seizure activity

2 - They recorder the EEG from the area where the tumor was located

3 - Yes they identify where the region of the tumor was in the brain and recording the EEG activity from the healthy brain area

4 - eyes closed, means when they were recording the EEG signal the patient had their eyes closed

5 - eyes open, means when they were recording the EEG signal of the brain the patient had their eyes open

All subjects falling in classes 2, 3, 4, and 5 are subjects who did not have epileptic seizure. Only subjects in class 1 have epileptic seizure (Fig. 1). We noted that it is characterized by composite oscillatory patterns and high amplitude.

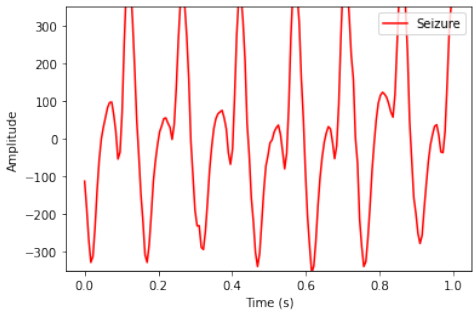


Fig. 1. Example recording of epileptic activity.

B. Tasks

We want to start with a simple task. The numbers of EEG epochs for each category are comparable with each other, so we first sought to classify epochs of epileptic seizure against recordings from healthy regions of the brain (Fig. 2). The two categories are balanced. There are 1488 epochs of seizure activity, and 1462 healthy epochs. In total there are 2950 epochs. The labels for healthy recordings were changed from 3s to 0s.

This subset of data is further divided into training set, validation set, and test set. 80 percent of the epochs are first randomly selected as the training set; of the remaining 20 percent, half of it was randomly designated as the validation set and the other half the test set. We also expanded the dimensionality and the final shapes are:

- training set: (2360, 178, 1)
- validation set: (295, 178, 1)
- test set: (295, 178, 1)

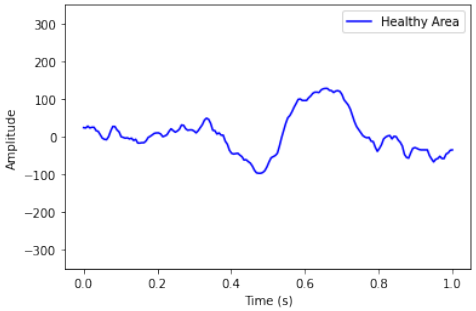


Fig. 2. Example recording from a healthy region.

We designed the next task to take advantage of the unused portion of the dataset, categories 2, 4 and 5 (Fig. 3, 4, 5). In addition to epileptic and healthy recordings, we include 1497 epochs of EEG from tumor regions, 1450 epochs of recording when the subject has eyes closed, and 1508 epochs or recording when the subject has eyes open. All categories other than seizure were relabeled as 0s.

This subset of data is further divided into training set, validation set, and test set. 80 percent of the epochs are first randomly selected as the training set; of the remaining 20 percent, half of it was randomly designated as the validation set and the other half the test set. We also expanded the dimensionality and the final shapes are:

- training set: (5924, 178, 1)
- validation set: (740, 178, 1)
- test set: (741, 178, 1)

C. Network structures

For each of the two tasks we sought to compare two models of different network type, a convolutional neural network and a bidirectional long short-term memory network. We have learned and seen examples that these types of networks are suitable for biomedical signals and sequences. We designed,

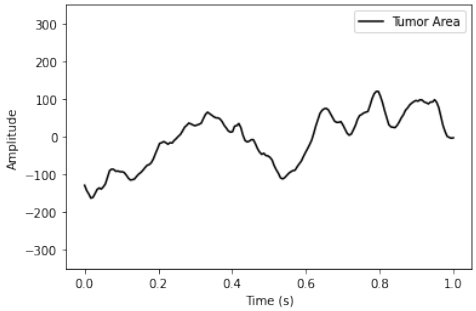


Fig. 3. Example recording from tumor area.

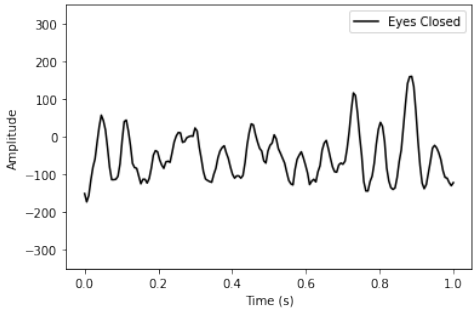


Fig. 4. Example recording when the patient has eyes closed.

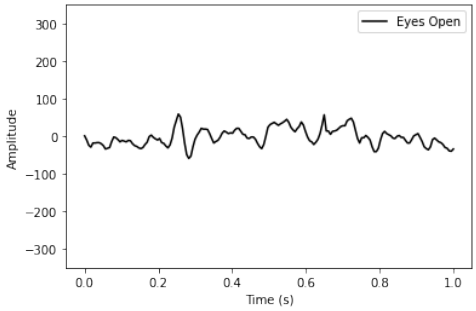


Fig. 5. Example recording when the patient has eyes open.

implemented and assessed the models using keras from tensorflow. Experiments were carried out in an online python notebook in Google Colab. A copy of data is stored and accessed with Google Drive.

Our CNN has 9 hidden layers. The input layer has shape (None, 178, 1) which corresponds to one epoch of EEG. After the input layer, the first hidden layer a one-dimensional convolution layer with 32 filters and kernel size 6. The output is batch normalized and fed through a 2-element max pooling layer. A second convolution layer with 64 filters and kernel size 3 follows. The output of the second convolution is also batch normalized and max pooled. The 7th layer is a flattening layer. Flattened output is fed into a densely connected layer with 32 hidden units, followed by another dense layer with 16 hidden units; both layers uses rectified linear unit (ReLU) as activation function. The output layer has size 2 and uses soft max as activation. In total there are 95,474 parameters, 95,282 of which are trainable.

The input layer of our BiLSTM model is the same as that of

Model: "CNN"

Layer (type)	Output Shape	Param #
inputs_cnn (InputLayer)	[(None, 178, 1)]	0
conv1d (Conv1D)	(None, 173, 32)	224
batch_normalization (BatchNo	(None, 173, 32)	128
max_pooling1d (MaxPooling1D)	(None, 87, 32)	0
conv1d_1 (Conv1D)	(None, 85, 64)	6208
batch_normalization_1 (Batch	(None, 85, 64)	256
max_pooling1d_1 (MaxPooling1	(None, 43, 64)	0
flatten (Flatten)	(None, 2752)	0
dense (Dense)	(None, 32)	88096
dense_1 (Dense)	(None, 16)	528
dense_2 (Dense)	(None, 2)	34

Total params: 95,474
Trainable params: 95,282
Non-trainable params: 192

Fig. 6. Summary of Convolutional Neural Network.

the CNN, with (None, 178, 1) in shape. The input is first fed through a densely connected layer with 32 hidden units and ReLU as activation. The key layer is the bidirectional recurrent layer, and we specified the recurrent units as long short-term memory units; there are 128 LSTM units. The output goes through a dropout layer with probability 0.3. The output is then batch normalized, and fed through a dense layer with 64 hidden units. This is followed by another dropout layer with 0.3 probability and another layer of batch normalization. The output layer has size 2 and uses soft max as activation. In total there are 182,786 parameters, 182,146 of which are trainable.

Model: "BiLSTM"

Layer (type)	Output Shape	Param #
inputs_lstm (InputLayer)	[(None, 178, 1)]	0
dense (Dense)	(None, 178, 32)	64
bidirectional (Bidirectional	(None, 256)	164864
dropout (Dropout)	(None, 256)	0
batch_normalization (BatchNo	(None, 256)	1024
dense_1 (Dense)	(None, 64)	16448
dropout_2 (Dropout)	(None, 64)	0
batch_normalization_1 (Batch	(None, 64)	256
dense_3 (Dense)	(None, 2)	130

Total params: 182,786
Trainable params: 182,146
Non-trainable params: 640

Fig. 7. Summary of Bidirectional Long Short Term Memory Network.

For both models we use Adam algorithm for stochastic optimization. We choose sparse categorical crossentropy as objective function to be minimized. Our metric for assessing model performance is prediction accuracy on the test dataset. During training, we also iteratively saved the model parameters with the highest validation accuracy, and compared with the final model.

III. Results and Conclusion

A. Experiment results

We trained a CNN and a bidirectional LSTM model on the two different categorical combinations of our EEG dataset; epileptic episodes against non-epileptic episodes; and epileptic episodes against all 4 remaining categories. In both experiments, the LSTM model resulted in higher accuracy and lower loss compared to the CNN model (Table 1).

	Epileptic VS Healthy		Epileptic VS All	
	CNN*	**BiLSTM**	**CNN***	**BiLSTM**
Epochs	500	100	100	100
Training Loss	0.6931	0.0207	0.6931	0.0265
Training Accuracy	0.4580	0.9947	0.8194	0.9912
Validation Loss	0.6931	0.1192	0.6931	0.1001
Validation Accuracy	0.4949	0.9729	0.8230	0.9770
Test Accuracy	0.5492	0.9831	0.8313	0.9798
Best Model Test Acc	0.9220	0.9797	0.8421	0.9838

Results varied widely between training sessions, the results represented here are example values of one session

TABLE I
SUMMARY OF PERFORMANCE METRICS CNN AND BILSTM

When trained on the epileptic (n=1,488) versus healthy (n=1,462) EEG data, the CNN model generated inconsistent accuracy metrics between training sessions. The final validation accuracy varied roughly from 0.1-0.9. Here we report a final test accuracy of 0.55, while the test accuracy from a saved best model equals to 0.922. Additionally, the training and validation accuracy curves, over the course of 500 training epochs, behaved erratically with high peaks and low valleys (Fig. 8A). Further, the validation and training loss values are equal at 0.6931 and the curves are superimposed and do not exhibit expected asymptomatic behavior over training time (Fig. 8B).

The LSTM model trained on the same two categorical variables of healthy versus epileptic EEG data showed more consistent results. The validation accuracy assessed on the model was 0.973 and the training accuracy remained higher than the validation accuracy (Fig. 9A). We report a test accuracy of 0.983, while the test accuracy from a saved best model equals to 0.980. Additionally, the training and validation loss after 100 epochs were 0.0068 and 0.0739, respectively, suggesting appropriate fitting of the data (Fig. 9B).

The CNN model trained on the full dataset categorized as epileptic (n=1,488) versus all other categories (n=5,917), behaved similarly compared to the previous experiment on the smaller subset of data (Fig. 10). Both the training and validation accuracy varied widely from 0.1-0.9 between training sessions and showed erratic behavior within training sessions. Again, the training and validation loss were equal at 0.693. Here we report a final test accuracy of 0.831, while the test accuracy from a saved best model equals to 0.842.

The bidirectional LSTM model trained on the full dataset showed similar validation loss and accuracy compared to when trained on the smaller subset in the first test (Fig. 11). Again, the performance metrics were significantly better compared to the CNN model. The validation accuracy and loss were 0.977 and 0.108, respectively, and the training loss and accuracy were better compared to the validation counterparts, indicating appropriate fitting of the model. We report a final test accuracy of 0.980, while the test accuracy from a saved best model equals to 0.984.

Comparing the results we report the more sohpisticated model (BiLSTM) with larger amount of data (epilepsy vs all other) yielded the best performance.

B. Conclusion and discussion

We have demonstrated that a bidirectional LSTM model is more accurate at predicting epileptic episodes from EEG data compared to a CNN model of the same depth. By comparing the models' performance metrics when trained on two different categorizations of the same dataset, we showed that the BiLSTM model produces higher validation accuracy and lower validation loss compared to the CNN model, regardless of categorization. Additionally, we find that the CNN model produces inconsistent and erratic accuracy metrics, whereas the BiLSTM model can consistently predict epileptic activity with 0.97 accuracy.

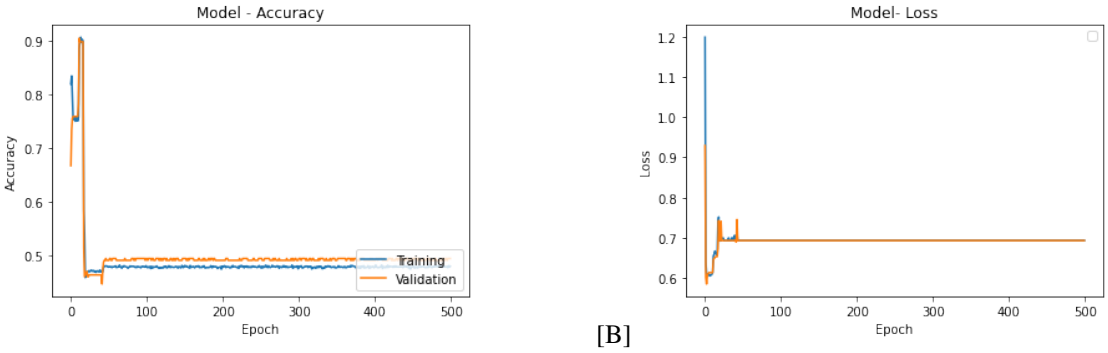
We attribute the inconsistent performance by CNN to our deliberate data division scheme. When assigning data into training, validation, or test sets, we did not maintain proportional balance between the categories; in one execution the proportion of seizure activity in the training set might be very low while the proportion in the validation set or the test set might be high, and the opposite can occur in a different execution. This fact further strengthens our claim of BiLSTM supremacy since it is invariant to the proportional imbalance.

Although the lack of exquisite relugarization in the CNN model or its relatively smaller size might play a role in the difference, we believe that the ability of the BiLSTM architecture to learn bidirectional time dependencies, which are characteristic of EEG data and even more pronounced in the strong oscillatory pattern in seizure activities, allowed for its superior performance over the CNN model.

It is important to note that a different type of network or architecture might work better given different input modalities. A recent publication showed that CNN is the most suitable structure for automated seizure detection when applied to 2D images of raw EEG waveform[10]. We emphasize that our claim is based on a specific task on a particular set of data. Nevertheless, our exploration and experimentation have been a fruitful learning experience. Further directions of research may include identifying mechanisms through which BiLSTM outperforms CNN, finding the best input modality for each network type, or integrating distinct network types into one model suitable for a variety of data.

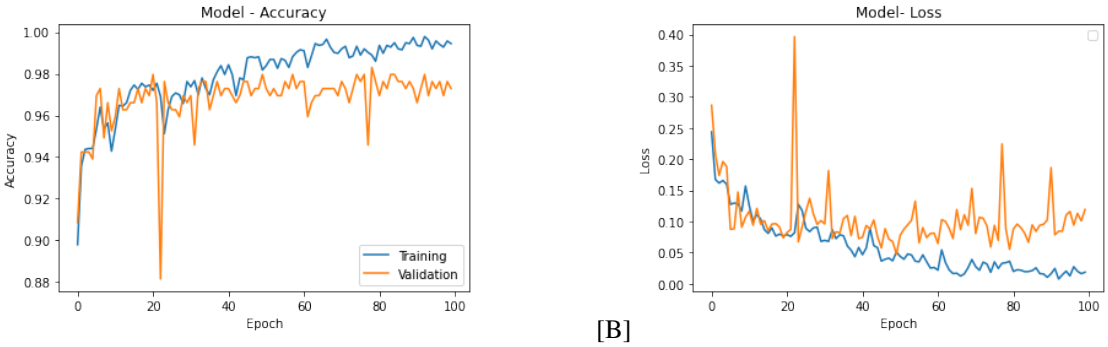
ACKNOWLEDGMENT

We would like to thank Prof. Sajda and the wonderful TAs for all the valuable guidance and assistance throughout the course, especially in such an unusual and uncertain semester.



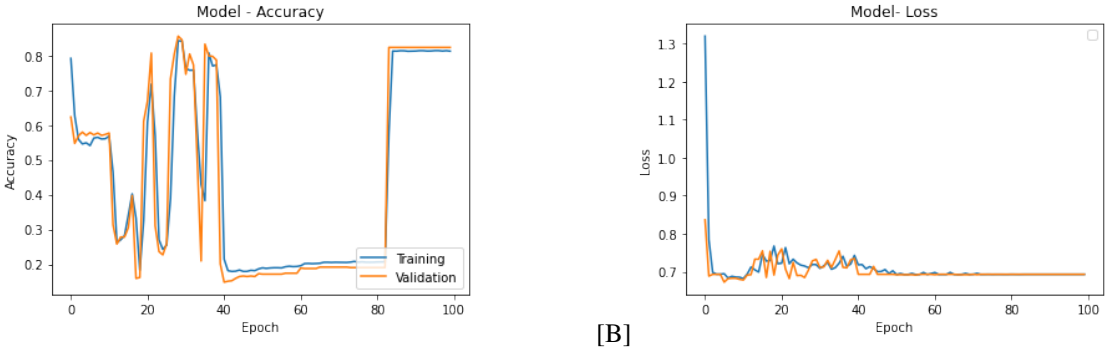
[A] [B]

Fig. 8. CNN Performance on Seizure v.s. Healthy



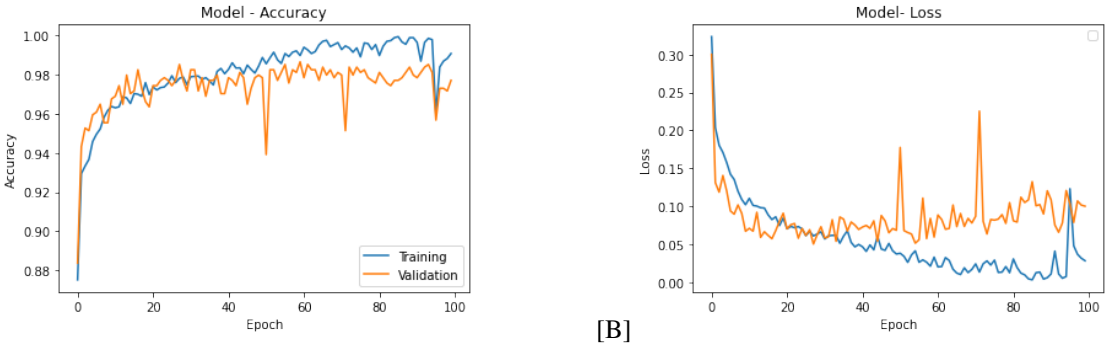
[A] [B]

Fig. 9. LSTM Performance on Seizure v.s. Healthy



[A] [B]

Fig. 10. CNN Performance on Seizure v.s. All Other Categories



[A] [B]

Fig. 11. LSTM Performance on Seizure v.s. All Other Categories

REFERENCES

[1] Fisher, R. S. et al. ILAE official report: a practical clinical definition of epilepsy. Epilepsia 55, 475–482, https://doi.org/10.1111/epi.12550 (2014).

[2] Mohseni, H. R., Maghsoudi, A. Shamsollahi, M. B. Seizure detection in EEG signals: a comparison of different approaches. In: Proceedings of the IEEE Engineering in Medicine and Biology Society Suppl, 6724–6727, https://doi.org/10.1109/IEMBS.2006.260931 (2006).

[3] McShane, T. A clinical guide to epileptic syndromes and their treatment. Arch. Dis. Child. 89(6), 591 (2004).

[4] Palus, M. Nonlinearity in normal human EEG: cycles, temporal asymmetry, nonstationarity and randomness, not chaos. Biol. Cybern. 75, 389–396 (1996).

[5] LeCun, Y., Bengio, Y. Hinton, G. Deep learning. Nature 521, 436–444, https://doi.org/10.1038/nature14539 (2015).

[6] Jang, H. J. Cho, K. O. Dual deep neural network-based classifiers to detect experimental seizures. Korean J Physiol Pharmacol 23, 131–139, https://doi.org/10.4196/kjpp.2019.23.2.131 (2019).

[7] Zhou, M. et al. Epileptic Seizure Detection Based on EEG Signals and CNN. Front. Neuroinform. 12, 95, https://doi.org/10.3389/fninf.2018.00095 (2018).

[8] Hussein, R., Palangi, H., Ward, R. K. Wang, Z. J. Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals. Clin. Neurophysiol. 130, 25–37, https://doi.org/10.1016/j.clinph.2018.10.010 (2019).

[9] Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, Elger CE (2001) Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Phys. Rev. E, 64, 061907

[10] Cho, KO., Jang, HJ. Comparison of different input modalities and network structures for deep learning-based seizure detection. Sci Rep 10, 122 (2020). https://doi.org/10.1038/s41598-019-56958-y