Artificial Intelligence In Medicine 117 (2021) 102084

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Interpreting deep learning models for epileptic seizure detection on EEG signals

Valentin Gabeff [a,*], Tomas Teijeiro [a], Marina Zapater [a,b], Leila Cammoun [c], Sylvain Rheims [d,e], Philippe Ryvlin [c], David Atienza [a]

[a] Embedded Systems Laboratory (ESL), EPFL, Lausanne, Switzerland
[b] REDS Institute, University of Applied Sciences Western Switzerland (HEIG-VD/HES-SO), Yverdon-les-Bains, Switzerland
[c] Department of Clinical Neurosciences, Neurology Service, Centre Hospitalier Universitaire Vaudois (CHUV) and University of Lausanne, Lausanne, Switzerland
[d] Department of Functional Neurology and Epileptology, Hospices Civils de Lyon and University of Lyon, Lyon, France
[e] Lyon's Neurosciences Research Center (INSERM U1028/CNRS UMR 5292), Lyon, France

ARTICLE INFO

Keywords:
Epilepsy
EEG Seizure detection
Interpretable deep learning
Convolutional neural networks

ABSTRACT

While Deep Learning (DL) is often considered the state-of-the art for Artificial Intel-ligence-based medical decision support, it remains sparsely implemented in clinical practice and poorly trusted by clinicians due to insufficient interpretability of neural network models. We have approached this issue in the context of online detection of epileptic seizures by developing a DL model from EEG signals, and associating certain properties of the model behavior with the expert medical knowledge. This has conditioned the preparation of the input signals, the network architecture, and the post-processing of the output in line with the domain knowledge. Specifically, we focused the discussion on three main aspects: (1) how to aggregate the classification results on signal segments provided by the DL model into a larger time scale, at the seizure-level; (2) what are the relevant frequency patterns learned in the first convolutional layer of different models, and their relation with the delta, theta, alpha, beta and gamma frequency bands on which the visual interpretation of EEG is based; and (3) the identification of the signal waveforms with larger contribution towards the ictal class, according to the activation differences highlighted using the DeepLIFT method. Results show that the kernel size in the first layer determines the interpretability of the extracted features and the sensitivity of the trained models, even though the final performance is very similar after post-processing. Also, we found that amplitude is the main feature leading to an ictal prediction, suggesting that a larger patient population would be required to learn more complex frequency patterns. Still, our methodology was successfully able to generalize patient inter-variability for the majority of the studied population with a classification F1-score of 0.873 and detecting 90% of the seizures.

1. Introduction

Epilepsy is a neurological disease characterized by paroxysmal events, called seizures, arising from the abnormal activation of neuronal networks. This abnormal activation translates into changes in the pattern of electrical activity generated by the brain, which can be captured through electroencephalography (EEG). The disease affects 50 million people worldwide, among which 70% could live seizure-free with appropriate diagnosis and treatment according to the World Health Organization [1]. Conversely, 30% of patients with epilepsy continue to suffer unpredictably recurring seizures. For these patients, there is a crucial need for the development of devices to detect seizure events [2]. EEG is the gold standard method to detect all seizure types in hospitals [3,4], but no reliable wearable EEG is yet available to transfer this approach for very long-term monitoring at home. Yet, innovative wearables are being developed and might allow such monitoring in the near future, stressing the need for EEG-based online seizure detection and interpretable models.

With the bloom of Deep Learning (DL) in the biomedical field, several methods have been developed to detect and predict seizure events from EEG of epileptic patients primarily recorded during short in-hospital monitoring with standard scalp-EEG or intracerebral electrodes [4]. Though some methods reported excellent performances, most used offline analysis with significant pre-processing and transformation of the

* Corresponding author.
E-mail address: valentin.gabeff@epfl.ch (V. Gabeff).

https://doi.org/10.1016/j.artmed.2021.102084
Received 19 December 2020; Received in revised form 27 April 2021; Accepted 29 April 2021
Available online 1 May 2021

EEG signal not compatible with the aim of online, long-term, ambulatory low-power operations.

Overcoming the challenge of efficient characterization of seizure events on a large and heterogeneous population of patients is also a crucial step for the transfer to clinical applications. Indeed, current devices suffer from generalization difficulties to unseen patients and they often need to be fine-tuned to each patient as a result of important patient inter-variability in epileptic disorders [4]. Training DL networks often requires dividing the input EEG into short segments, typically between 0.5 and 30 seconds [4]. Classification metrics of those individual segments is a necessary step to characterize model performance, but if we aim to provide a model for seizure detection, one should also assess the performance on longer stretches of EEG signals carrying transitions between interictal and ictal phases.

Given that the "black-box" nature of DL is a common obstacle to the transfer of applications in clinics [5], we aim to explore not just the features learned by a DL model but also the input data properties leading to a classification decision with the view to improve the validity of the method and potentially the understanding of the related pathology. Delineating the key characteristics of the input data for an efficient classification can then justify the conception of the model architecture and the choice of processing methods.

This work is thus a step towards strengthening the relationship between the current knowledge of EEG signals in epileptic disorders and the development of transferable DL methods for characterization of seizure events. In the continuity of the work with the e-Glass as an EEG monitoring wearable device [6], our study focuses on electrodes placed over the temporal brain regions and explores performance of the model both at the segment and seizure levels, with an extensive discussion on how these two levels are related. As a result, we report a non-patient-specific online method using raw EEG to detect seizures, and investigate the features learned by the model, providing a visual feedback of the decisive patterns for seizure detection on the EEG signal.

2. Related work

A common follow-up practice for epilepsy requires that the patient documents each seizure in a paper or electronic diary to later provide an appropriate therapy [2,7]. Unfortunately, seizure events are often underreported, partly due to seizure-induced amnesia of seizure events [8]. This justifies the development of wireless recording devices coupled with an electronic diary to detect, record and eventually predict or forecast seizure events. Existing wearable devices often only detect a fraction of seizures, called generalized tonic-clonic seizures (GTCS), thanks to easy-to-observe GTCS-induced variation of various biosignals, including surface-electromyography (sEMG), electrodermal activity (EDA) and 3D-accelerometry [9,10,6]. However, these biosignals do not currently offer a reliable way to detect the majority of non-GTCS seizures [11].

Recently, DL provided new opportunities to address the classification of EEG signals. Since it is a data-driven method, feature extraction is strongly simplified, at the cost of requiring much larger datasets. Some methods used Recurrent Neural Networks (RNN) to account for the temporal nature of EEG [12]. However, most methods have used Convolutional Neural Networks (CNN) [13–19]. Not only their architecture and mechanisms have been constantly improved due to their thorough utilization in the field of modern computer vision, but they often provide better results than simple RNN architectures for the classification of EEG time-series [20]. In 2015, de Aguiar *et al.* presented a weightless DL architecture to predict seizures online with an anticipation between 2 and 30 seconds before the onset [17]. The method is patient-dependent and achieves accuracies between 0.725 and 0.99 on the EPILEPSIAE dataset [21]. The authors did not perform any signal processing aside from the encoding and the architecture of the network was carefully designed to prepare for online processing on an embedded device. More recently, Yuan *et al.* [15] developed an auto-encoder followed by a shallow multi-view CNN architecture that fosters intra- and inter-channel dependencies from spectrograms. The proposed framework achieved a F1-score of 0.85 on the CHB-MIT dataset [22] for the classification of 3-second segments between the ictal and pre-ictal or interictal classes, in a patient-independent fashion. Despite a very good performance, the method lacks clinical applicability as it was not used to detect seizures on longer stretches of EEG recording.

Detection of seizure events from continuous EEG is commonly done by processing the short segments of test EEG signals in a time linear fashion and reporting an aggregated performance (*e.g.* seizure sensitivity). This helps to understand the behavior of the model during continuous monitoring and it is a necessary step for the development of monitoring devices. Most studies report performance on short segments but few additionally report sensitivity, precision and accuracy for complete seizures episodes. We refer to these two levels of performance as follows: "segment-level" depicts performance on the short segments of test EEG, while "seizure-level" refers to performance on the individual seizure events. In [13], the authors proposed a simple CNN architecture that predicts a seizure onset 5 to 35 minutes following the alarm if 8 out of 10 consecutive segments are classified as pre-ictal. On the CHB-MIT dataset, the patient-dependent method achieves a sensitivity of 81.2% with a false positive rate of 0.16 per hour. In [13,17], the authors aimed at predicting the seizure onset, while [15] focuses on detecting short ictal segments. Our work lies at the intersection between both approaches, where we aim to detect seizure onsets as early as possible (and possibly before) by training the network with interictal and ictal segments.

Since features in a DL model are automatically extracted during training without human intervention, decisions are rarely interpretable for clinicians. Some studies still advance to explain network dynamics when processing EEG signals [23–25]. In [23], the authors visualized the important frequencies and spatial locations for each class in a motor imagery task using causal inference. Specifically, they reported how increasing or suppressing the spectral amplitude of a frequency component causes changes in output neuron activation. Hartmann and colleagues extended the approach to all convolutional layers of the network [24]. They discovered that spectral amplitude is an important feature for the imagery task in the last layers of the model while modulating spectral phase affects early layers. The authors also showed that every layer extracts different frequencies by observing the median of the input samples that yielded the strongest activation for each filter. For the same task, Sturm *et al.* used Layer-wise Relevance Propagation (LRP) to visualize the most important features on the input signal by propagating the relevance of each neuron to each class from the output back to the input [25]. Their results showed that imagining right hand movements was associated with relevant features in the left hemisphere and inversely for the left hand, as expected. In this work, we take inspiration from these visualization methods to bring explanations on network dynamics in the context of epilepsy detection.

3. Materials and methods

3.1. Dataset

The dataset used in this study comes from the REPO$_2$MSE cohort, whose characteristics were previously reported in [26]. It contains multi-channel scalp-EEG recordings from 568 patients with epilepsy, and annotations of seizure onsets from an experienced epileptologist.A total of 1212 distinct seizures, each in one record file, are available with each record being cut to contain a single seizure with a median of 3.0 minutes of interictal recording before the annotated seizure onset. EEG recordings were either sampled at 256 Hz (89.6%), 512 Hz (10.2%) or 1024 Hz (0.2%). The median number of files containing a seizure was of 2 per patient, with a maximum of 9 and a minimum of 1.

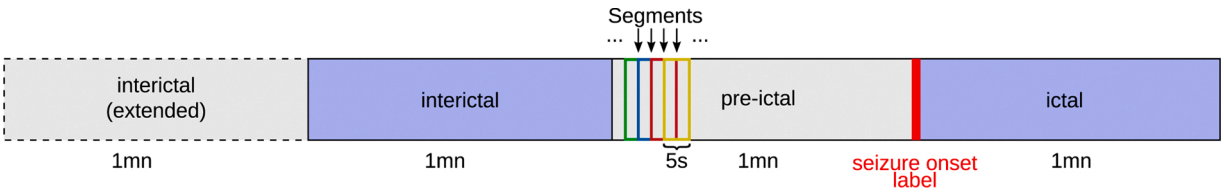

Fig. 1. Segmentation of EEG signal. Signals are cut into four parts relative to the seizure onset label. The ictal portion extends up to one minute after seizure onset. The pre-ictal portion is considered as the minute before seizure onset and is preceded by one minute of interictal signal. An additional minute of interictal is added in for post-processing with a difference filter, as explained in Section 3.5. Ictal and interictal portions (blue) are used for training and computation of metrics(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). Grey regions are only used for prediction but are not taken into account for the calculation of the evaluation metrics.

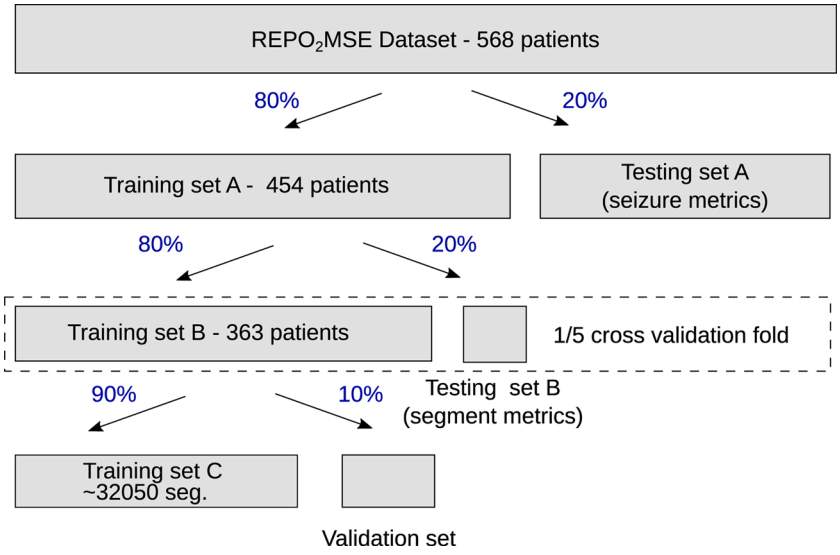

Fig. 2. Database splitting. Data from 568 patients are used. 20% are left out for testing at the seizure-level (test set A). 80% are used for both cross-validation training (B) to obtain metrics at the segment-level and again for training to obtain metrics at the seizure-level without cross-validation (A). 5-fold CV requires 80% of training data (B) as the training set. Because we use early stopping, we further select 90% of the training set (C) as the final training set and leave 10% as validation metrics during training. All steps except the last ensure data independence at the patient level. The number of 5-seconds segments used for training is not fixed at each fold as not all patients have the same amount of data. 57.6% of the full database is used for training at each cross-validation fold and 72% for final training.

3.2. Preparing the input

Pre-processing steps were minimized for an online method. Raw EEG signals were down-sampled to 256 Hz and cut in overlapping segments of 5 seconds. Although a time-frequency representation of the signal is commonly employed [15,13], we believe that a careful design of the model architecture taking into account how experts visually interpret raw EEG signals in time and space may perform as well as more explicit representations. This study focuses on four EEG channels “F7-T7”, “F8-T8”, “T7-P7” and “T8-P8” to prepare future long-term recordings using wearable EEG systems like the e-Glass [6].

The segmentation scheme of each EEG signal is represented on Fig. 1. For each available seizure, we considered one minute of interictal recording and one minute of ictal recording. The duration between ictal and interictal classes was equal to balance the number of training samples. The interictal segment started two minutes and ended one minute prior to seizure onset. The immediate pre-ictal minute before seizure onset was not used for training to guarantee a clear separation between interictal and ictal segments, following the recommendations of the clinical experts. The ictal segment started at seizure onset. Some seizures might last less than one minute, leading the ictal segment to include some immediate post-ictal recording. Because this post-ictal segment is often difficult to distinguish from the ictal phase itself, and might also be informative for seizure detection, we did not attempt to separate true ictal from immediate post-ictal activities during the one minute of ictal recording. The interictal segments were considered as “negative class” segments, while the ictal segments were considered as “positive class” segments. Each segment is then subdivided in windows of 5 seconds with 50% overlap for data augmentation and centered with respect to the median of the window. This allows to balance training data with each file including 23 negative samples and 23 positive samples. The windowing strategy gives a resolution of 2.5 seconds when predicting seizures at the seizure-level.

The method to split the data between training, validation and testing sets is illustrated in Fig. 2. 80% of the data are used for training and testing at the segment-level and 20% for testing at the seizure-level. They are referred to as training and testing sets A, respectively. We perform 5-fold cross-validation (CV) splitting on the patient list with 10% of validation data at each fold on the training set to obtain metrics at the segment-level and fine-tune hyper-parameters used for detection of the seizure onset. They are referred to as training and testing sets B. 10% of training set B is used as validation to monitor the training performance of training set C and implements early-stopping. When training on set A without cross-validation, 10% of the samples are used for validation (not represented).

3.3. Network architecture

We developed a CNN, as this type of architecture has shown promising results in epilepsy detection and classification [13,14,17]. The CNN input are 2D gray-scale images with dimension 4 channels x 1280 time samples (5 seconds x 256 Hz). The model architecture is illustrated in Fig. 3. It is composed of three blocks of convolutional layers followed by two fully connected (FC) layers with a single output for binary classification (*i.e.*, ictal vs. interictal). Each block of convolutional layers consists of two units of a convolutional layer, followed by a Batch Normalization (BN) operation and Rectified Linear Unit (ReLU) activation. BN is used to re-center the data and to ensure a non-linear ReLU activation, as this has been proven to speed-up training and improve model performance [27]. A max pooling operation follows each block to

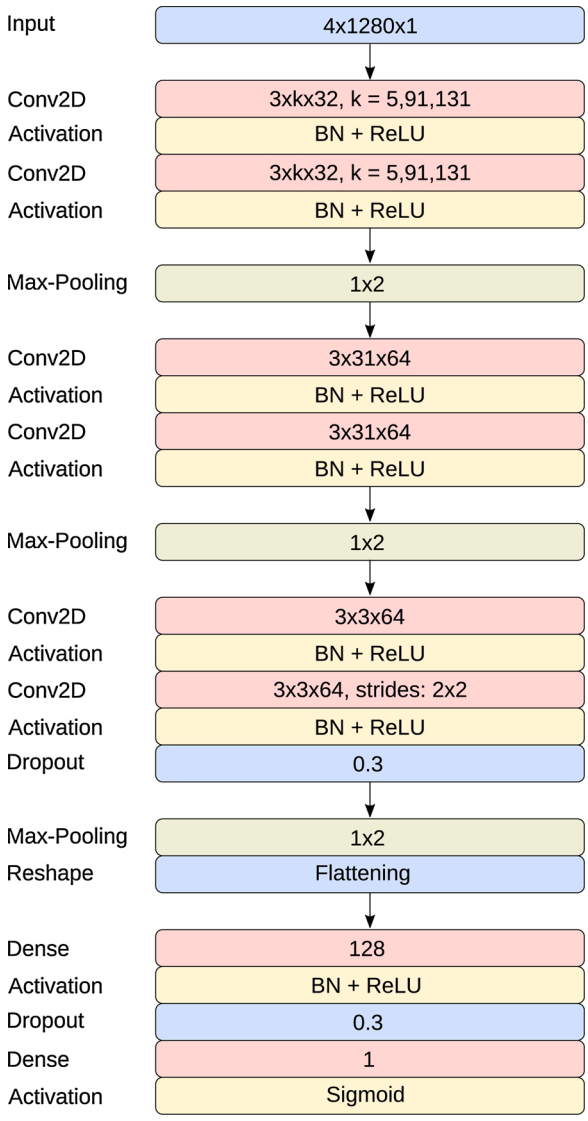

Fig. 3. Neural network architecture. The network architecture is composed of 3 convolutional blocks and two fully connected layers. Each block is followed by a pooling operation. We compare 3 versions of this network with different values of k in the first layer kernels.

reduce dimensionality and to improve temporal invariance of the input. Convolutional layers in the first block contains 32 $3\times k$ kernels, 64 3×31 kernels in the second block, and 64 3×3 kernels in the last block with a stride of 1×1 in all blocks. To explore the features learned by the model in the first layer, we tested different values of k and compared models with kernel size of 3×5, 3×91, and 3×131. In what follows we refer to each one of these models according to the first layer kernel size. A kernel of length 131 along the time dimension can capture frequencies as low as 2 Hz, while a kernel of length 5 can only explicitly extract high-gamma frequencies in the first two layers. In the latter case, the following 3×31 block can also extract lower frequencies. The temporal sizes of the kernels were chosen arbitrarily and fine-tuned experimentally, but they were kept distinct enough to extract different ranges of frequencies.

A 30% dropout (DO) operation is used before each FC layer and 0.05 weighted kernel and bias regularization are used to avoid overfitting. Models are trained for a maximum of 120 epochs with early-stopping if no improvement of the validation loss is made after 15 epochs. Stochastic Gradient Descend (SGD) optimizer is used with a learning-rate of 0.005. The training was performed on a computing server with 2 AMD EPYC 7551 32-Core Processors and 500 GB of RAM, equipped with an Nvidia Tesla T4 GPU. The models were implemented with Keras using the Tensorflow v1.4 backend. Training on set A takes 99 minutes for 50 epochs, or a mean time of 118 seconds per epoch.

3.4. Metrics

We report accuracy, sensitivity, precision, and F1-score both at the segment and seizure-levels. Metrics at the segment-level are reported by concatenating results for each CV-fold while Area Under the receiver operating characteristic Curve (AUC) is averaged across CV folds (testing sets B), as advised in [28]. Results at the seizure-level are reported after training on 80% of patients (training set A) and testing on the remaining 20% (testing set A).

As there is no signal without seizures in the dataset, metrics at the seizure-level are computed according to the negative and positive parts of the signal. A seizure is correctly detected if the onset is detected in the ictal part of the signal and counted as a false positive if detected in the interictal part.

3.5. Post-processing

The performance of the seizure detection method needs to be assessed at the time scale of seizure events and not only on short individual segments. Since the detection of seizure onset does not require that all sub-segments are classified correctly, we expect the performance at the seizure-level to be higher than at the segment-level. Two aggregation methods are employed to detect seizure events from the predictions of successive segments. The first follows a Bayesian approach. The evidence for each class is computed by taking the product of continuous predictions over a sliding window of size W. A window is considered positive if the log-odds of the ictal evidence over the interictal evidence is superior to a threshold th_b.

The second method applies a difference filter to successive predictions to detect the transition between the pre-ictal and ictal segments. Previous output probability for a segment at time $t = -M$ is subtracted from the output probability at time $t = 0$, where M is the length of the difference filter. A window was considered positive if it reached a threshold th_d. To account for values of M up to 23 samples, we also considered the 1 minute of signal before the interictal part so that the false positive rate is not artificially reduced in the interictal part.

The Bayesian method focuses on detecting segments of high ictal evidence, while the difference method addresses the detection of the transition between immediate pre-ictal and the seizure onset. The hyperparameters of each method were optimized in a grid-search fashion on the concatenated CV folds of set B and were then used to generate metrics at the seizure-level on the test set A. To compare the differences of dynamics between models, we also computed their performance at the seizure-level for each aggregation method with a fixed set of hyperparameters.

3.6. Visualization

We applied two visualization methods to understand the decision dynamics of the network on the EEG signal. First, we explored what are the inputs maximizing the first layer kernels using gradient ascent. Inputs are initialized with random samples in the range [-10 μV;10 μV]. As the resulting signals were sinusoids, we computed their power spectrum with the Welch's method and reported the main frequency components for each channel [29]. This method does not inform if a given frequency component contributes to the ictal or interictal class. Accordingly, maximized inputs are fed back to the network and probability at the output is kept. From this probability, we may then infer that a frequency component (or a combination of them) is associated to the ictal class if it has an output probability close to 1, and to the negative class otherwise. For the 3×5 model only, we additionally perform the same analysis but

Table 1
Metrics at the segment-level. Accuracy, sensitivity, precision, and F1-score are reported by concatenating CV outputs (test sets B). Network output is converted to binary classification according to two decision thresholds chosen arbitrarily to explore performance at low- and high-sensitivity. Overall, 3×5 model exhibits the highest sensitivity and 3×131 the best precision. Averaging AUC across folds shows that all three models perform equally.

Model	3×91		3×131		3×5	
Threshold	0.150	0.850	0.150	0.850	0.150	0.850
Accuracy	0.705	0.686	**0.711**	0.644	0.657	0.690
Sensitivity	0.930	0.405	0.914	0.307	**0.962**	0.416
Precision	0.481	0.966	0.508	**0.981**	0.353	0.964
F1-score	0.759	0.563	**0.760**	0.463	0.737	0.573
AUC	0.866 ± 0.02		0.867 ± 0.022		0.866 ± 0.026	

maximizing filters of the last convolutional block instead of the first to see if a model can progressively build a representation of lower frequencies after several pooling operations.

Second, we visualized the learned features back on the EEG signal using SHAP values (SHapley Additive exPlanations) [30]. This method compares the output difference between a baseline EEG signal and a given input and propagates this difference back to the input signal, similarly as in [31]. A positive SHAP value indicates an input data point that led to a positive difference output and therefore a contribution to the positive class.

The first method is based on the model weights at the first layer and constructs preferred activation patterns. The second method is activation-based and has potential for clinical applications. It requires a forward pass of an input sample to highlight ictal features on the signal.

4. Experimental results

In this section, we first discuss the detection performance of the three models both at the segment and seizure-levels. We then explore the results of the maximized inputs and of the DeepLIFT visualization method.

4.1. Model performance at the segment level

Table 1 shows the classification performance metrics at the segment-level for the 3×91, 3×131 and 3×5 models. We applied two different decision thresholds: 0.15 and 0.85 for high and low sensitivity respectively. Fig. 4 displays the prediction polarisation at the output neuron. The 3×131 model yields the best F1-score of 0.76 at the low threshold and is the model with output probabilities that are the most shifted towards the negative class. On the contrary, the 3×5 model has output probabilities shifted towards the ictal class and the best F1-score of 0.577 at a high decision threshold. The 3×5 model keeps relatively high sensitivity at a high threshold because the distribution is more shifted towards the positive class than for the other models and the opposite is observed for the 3×131 model. The 3×91 model is the most balanced with F1-scores of 0.759 and 0.566 for decision thresholds of 0.15 and 0.85, respectively. All models have an AUC of 0.87.

4.2. Model performance at the seizure level

4.2.1. Evidence aggregation

Results of grid search selection for the hyperparameters W and th_b in the Bayesian aggregation approach are displayed on Fig. 5 and the selected hyper-parameter values in Table 2. Optimum W^* are 5, 7, 5 and th_b^* are 1.5, 2.5, 1.5 for models 3×5, 3×31 and 3×131, respectively. The optimal windows correspond to 12.5 to 15 seconds of signal and optimal log-odds ratio to an aggregated probability approximately between 0.8 and 0.9, meaning that high evidence is required to predict a positive

Table 2
Optimized hyper-parameters for the Bayesian approach.

Parameters	W^*	th_b^*
3×5	5	1.5
3×91	7	2.5
3×131	5	1.5

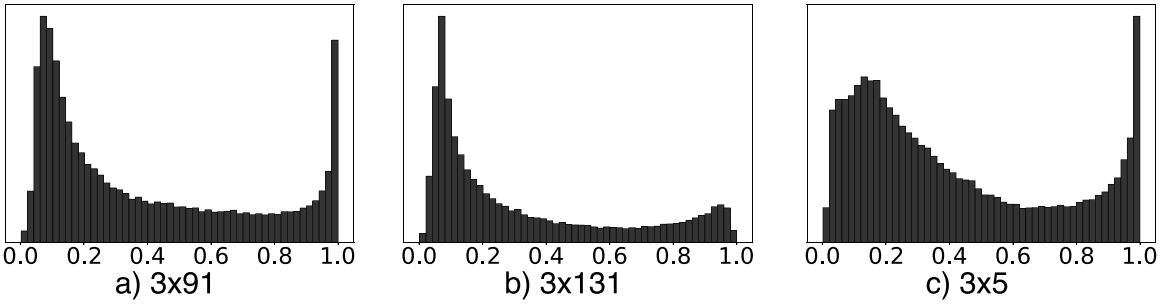

Fig. 4. Distribution of output probabilities. Output probabilities of each CNN are plotted as a histogram. 3×5 model has probabilities shifted to the right, suggesting a higher sensitivity. On the contrary 3×131 model has few outputs close to one indicating a potentially lower sensitivity.

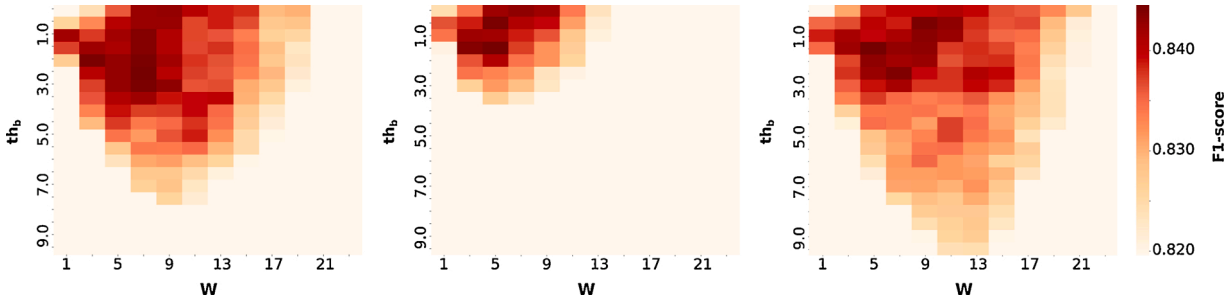

Fig. 5. Hyper-parameters space - Bayesian approach. a. 3×91, b. 3×131, c. 3×5. F1 scores are smoothly distributed in the hyper-parameter space of every model, showing a low risk of overfitting. The 3×5 model has optimal hyper-parameters shifted towards high log-odds as a consequence of a high sensitivity at the segment-level. The opposite is observed for the 3×131 model.

Table 3
Metrics at the seizure-level - Bayesian approach.

Model	3×91	3×131	3×5
Sensitivity	0.834	**0.895**	0.886
Precision	0.825	0.795	**0.808**
Accuracy	0.83	0.845	**0.847**
F1-score	0.83	0.852	**0.853**

Table 4
Hyper-parameters optimization - Difference filter.

Parameters	M^*	th_b^*
3×5	17	0.45
3×91	15	0.5
3×131	21	0.45

output.

Accuracy, sensitivity, precision, and F1-score are reported in Table 3. The 3×131 model has the highest sensitivity with 89.5% of the test seizures detected, while the 3×91 has the lowest one with a value of 83.4%. The 3×5 model has the best F1-score of 0.853 and the highest precision, with less than 9.82% of false positives.

4.2.2. Difference filter aggregation

Optimal M^* and th_b^* are reported in Table 4 and the F1-score grids in Fig. 6 for the difference filter aggregation method. Optimal difference windows are 17, 15, 21 and optimal detection thresholds are 0.45, 0.5, 0.45 for the 3×5, 3×91 and 3×131 models, respectively. The optimal time scale to compute the difference of prediction between two samples is then between 37.5 and 52.5 seconds, and requires a prediction difference close to 50%.

Table 5 shows the metrics at the seizure-level for the difference filter method. The 3×131 model performs best with an F1-score of 0.873 and 90% of the seizures detected, while the 3×91 model has the lowest false positive rate in the interictal part with a precision of 0.891. This can be explained by a higher th_b^* and shorter M^*.

4.2.3. Model differences

When comparing Tables 1, 3 and 5, we observe that performance at the seizure-level is higher than at the segment-level for both aggregation methods, as expected. At the segment-level, the F1-scores never exceeded 0.76, while it reaches 0.873 for the 3×131 model with the difference filter method. Performance differences between models at the segment-level are not conserved at the seizure-level. Indeed, the 3×131 model was yielding the highest precision and the 3×5 model the highest sensitivity, while it is the opposite at the seizure-level for the Bayesian approach. This effect is less pronounced for the difference filter.

Results of model comparison with a fixed set of hyper-parameters are shown in Tables 7 and 6 . They reveal the previous dynamic observed at the segment-level where 3×5 model is yielding the highest sensitivity and 3×131 model the highest precision. Hyper-parameter optimization tends to make methods converge to a common behavior by compensating differences at the seizure-level.

4.2.4. Aggregation methods comparison

Fig. 7 compares the outputs of both aggregation methods on 229 test signals, represented as rows. The left column shows the continuous probability outputs of the model and the middle and right columns represent the post-processed binary outputs for the Bayesian approach and the difference filter, respectively. The Bayesian approach often leads to higher sensitivity and longer stretches of ictal detection. False positives in the interictal part are reduced for the difference method, as this method can detect drifts in output probabilities stronger than th_b^* for signals with probabilities constantly above 0.5. This explains why some signals are considered as fully ictal in the Bayesian approach while not

Table 5
Metrics at the seizure-level - Difference filter.

Model	3×91	3×131	3×5
Sensitivity	0.817	0.904	**0.908**
Precision	**0.891**	0.834	0.76
Accuracy	0.854	**0.869**	0.834
F1-score	0.848	**0.873**	0.846

Table 6
Metrics at the seizure-level with $W = 7$ - Bayesian approach 3×91 model performs the best for low threshold and 3×131 is the best classifier at a high threshold. This last result is rather surprising as the observations are opposed to the difference filter method. Antithetic dynamics of 3×131 and 3×5 models is still preserved.

Model	3×91		3×131		3×5	
Threshold	0.5	3.0	0.5	3.0	0.5	3.0
Sensitivity	0.904	0.817	0.895	0.865	0.974	0.93
Precision	0.803	0.847	0.782	0.852	0.664	0.729
Accuracy	0.854	0.832	0.838	0.858	0.819	0.83
F1 score	0.861	0.829	0.847	0.859	0.843	0.845

Table 7
Metrics at the seizure-level with $M = 19$ - Difference filter. 3×131 model performs the best at a low threshold and 3×5 at a high threshold. This highlights again the opposite performance of both models.

Model	3×91		3×131		3×5	
Threshold	0.4	0.6	0.4	0.6	0.4	0.6
Sensitivity	0.904	0.729	0.908	0.764	0.93	0.795
Precision	0.764	0.943	0.782	0.948	0.721	0.913
Accuracy	0.834	0.836	0.845	0.856	0.825	0.854
F1 score	0.845	0.817	0.854	0.841	0.842	0.845

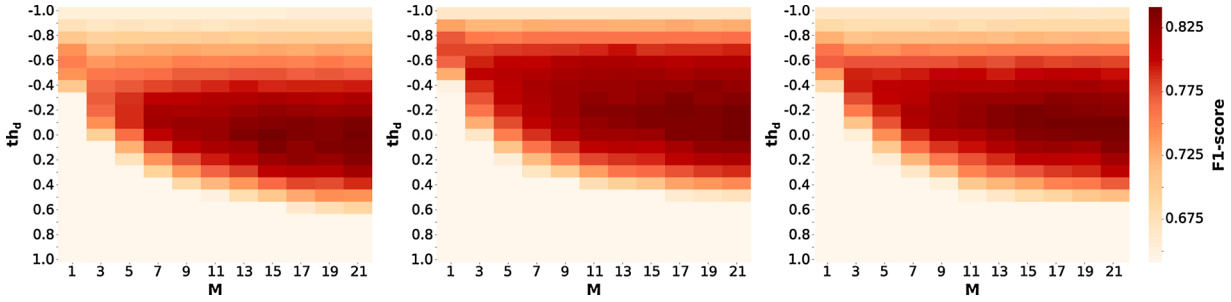

Fig. 6. Hyper-parameters space - Difference filter. a. 3×91, b. 3×131, c. 3×5. Hyper-parameters space is smooth reducing the risk of overfitting. 3×131 model space is shifted towards lower threshold values. All models seem to have an optimal size of filter beyond 23. This factor is limited by the short time of recording before seizure onset label.

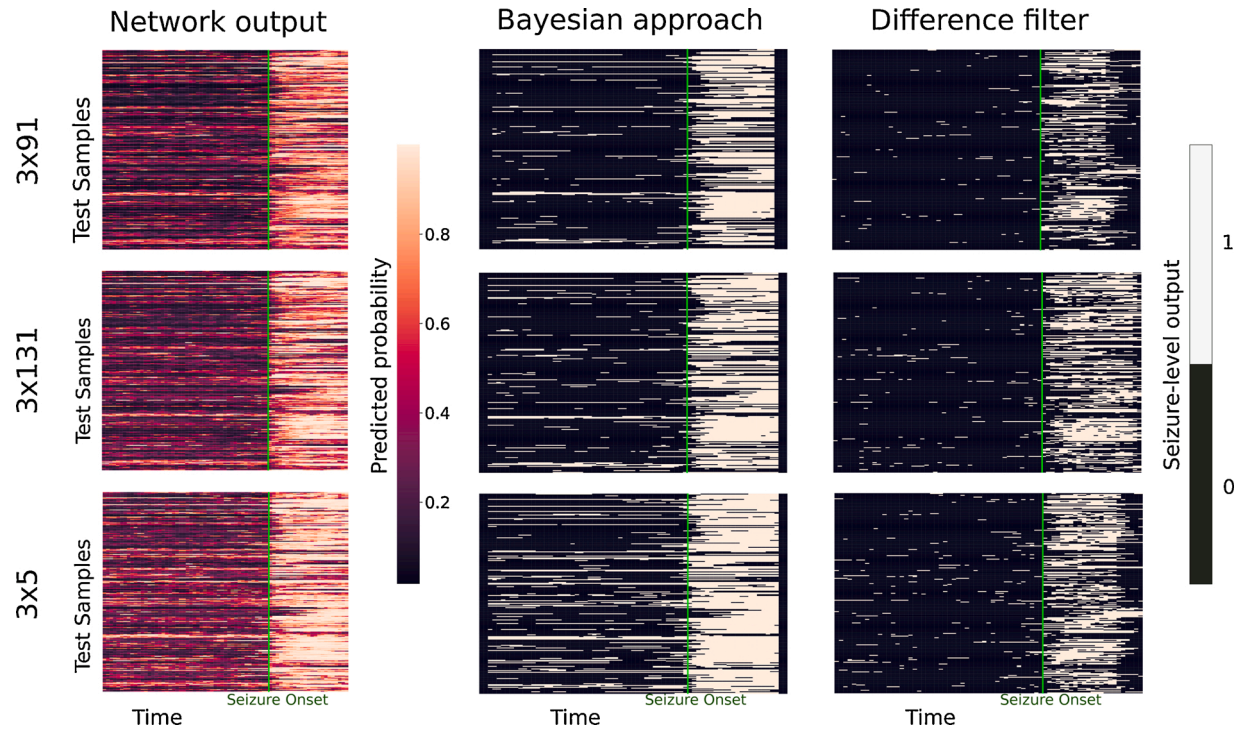

Fig. 7. Comparison of post-processing methods. Network outputs (left column) are presented as heatmaps with values ranging from 0 to 1 for the negative and positive class respectively. The raw outputs are then transformed by the Bayesian approach (middle column) or the difference filter method (right column). The Bayesian approach closely resembles the network output, with high sensitivity but also a higher false positive rate. On the contrary, the difference filter is more focused on the seizure onset and has a lower false positive rate.

being detected as such by the difference method.

For each model, falsely classified segments mostly correspond to signals being classified either as fully ictal or fully interictal with no clear transition between immediate pre-itcal segments and seizure onsets. These signals are consistent for each model showing that part of the seizure signals could never be properly detected with any of the proposed models.

4.3. Network interpretability

4.3.1. Maximized inputs

Fig. 8 shows the result of the maximized input visualization using gradient ascent. The output of 9 out of the 32 available filters eliciting strongest activation are represented. Distinct sinusoidal patterns are observed for 3×91 and 3×131 models after the third maximized inputs. This suggests that first layer kernels extract specific frequency components in the input EEG and supports the choice for large kernel sizes. High-frequency components are always present in top maximized inputs indicating that they are the most decisive features in the first layers. Synchronicity with a small phase shift between channels of maximized inputs can also be observed in most examples, suggesting that spatial correlation is an important feature learned by the model.

Tables 8 and 9 display the main frequency components of maximized inputs along with their respective output probabilities when feeding them to the network. We did not report the main frequencies for the 3×5 model as the power spectrum was broad and no singular frequencies could be identified. For 3×131 and 3×91 models, the majority of maximized inputs contain frequencies in the alpha, beta, and low-gamma bands. However, maximized inputs eliciting the strongest activation response contain high-gamma frequencies between 70 and 100 Hz. Those filters also lead to strong activation of the ictal class. Results highlight that most maximized inputs contribute to the ictal class. However, frequency components around 8 Hz are associated with activation of the interictal class.

To study the importance for the model to extract high-gamma frequency bands in the maximized inputs, we raised the input range amplitude from 10 μV to 100 μV. Fig. 9 and Tables 10 and 11 show a much higher proportion of high frequency components in the maximized input in the 3×131 and 3×91 models. Hence, high frequency components seem responsible for extracting the high amplitude information in the signal. Maximized inputs containing 8 Hz frequency components still elicit the lowest output probabilities although now being classified as ictal. A large amplitude is then a distinctive feature of the ictal class in all three models, and overrides other features that were previously indicators of the interictal class.

Fig. 10 and Table 12 show the most important frequency components in the preferred inputs of the last convolutional block for the 3×5 model. Although a kernel of size 5 in the first layer can only extract frequencies as low as 51 Hz, the model builds a representation of the lower frequency components in the last convolutional layers of the network. Pooling operations are most likely contributing to this effect. We should keep in mind that the 3×5 model has a 3×31 layer in the second block, and therefore it is able to explicitly extract frequencies as low as 9 Hz. However, Table 12 shows that it can still learn frequency features as low as 4 Hz.

4.4. Inference visualization

We highlight input features characteristic of the ictal class by overlaying the matrix of SHAP values on the EEG signal, as shown on Fig. 11. As expected the ictal portion contains a higher proportion of positive SHAP values than the interictal part. On the example presented and consistently for every model, decisive features are generally spikes of high amplitudes in the EEG signal at a relatively consistent time frequency. Fig. 11 also shows some correlation of the SHAP values between channels.

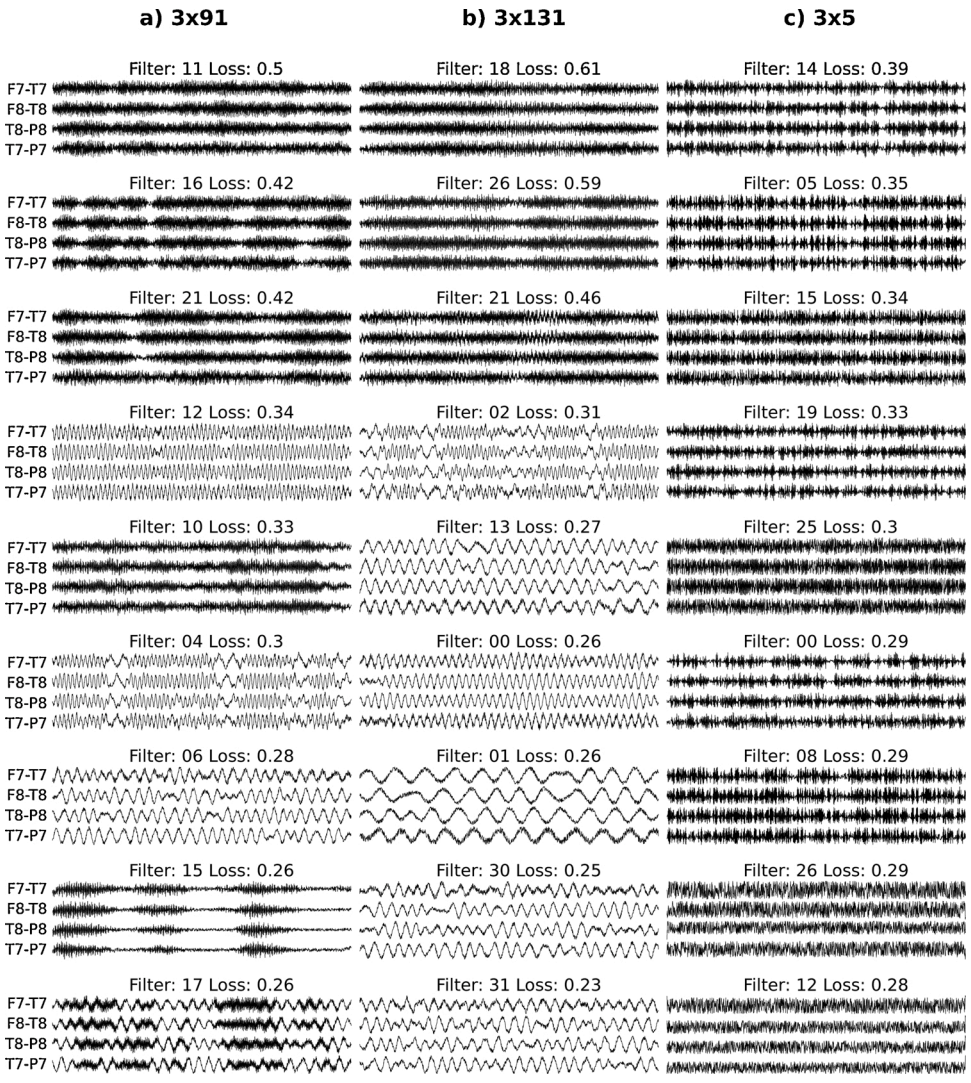

Fig. 8. Most significant maximized inputs for first layer kernels with low initial amplitude. Maximized inputs are sorted according to their contribution to a custom loss function after the last gradient ascent step. 3×5 model shows only high frequency detecting kernels on the first layer. Indeed, the kernel sizes can only detect 50Hz frequencies and above with a sampling frequency of 256Hz. 3×131 and 3×91 model both have lower frequency detecting filters. High frequencies detecting filters elicit the strongest response after the first layer in all models.

Table 8
Main frequency components of maximized inputs with low initial amplitude - 3×91. The top three maximized inputs contain high frequency components leading to strong activation of the ictal class. Only the 8Hz component is associated with activation of the interictal class.

Filter idx	F7-T7	F8-T8	T8-P8	T7-P7	pred.	loss
11	[97]	[97]	[97]	[97]	0.999	0.496
16	[97]	[97]	[97]	[98]	0.999	0.425
21	[97]	[97]	[97]	[97]	0.999	0.418
12	[14]	[14]	[14]	[14]	0.999	0.337
10	[72, 97]	[72, 97]	[72, 97]	[72, 97]	0.997	0.332
4	[14, 4]	[14, 4]	[14, 4]	[14, 4]	0.994	0.300
6	[5]	[5]	[6]	[6]	0.994	0.280
15	[72, 97]	[72, 97]	[72, 97]	[72, 97]	0.673	0.264
17	[97, 5]	[97, 5]	[97, 5]	[5, 97]	0.999	0.264
[...]	[...]	[...]	[...]	[...]	[...]	[...]
3	[8]	[8]	[8]	[8]	0.092	0.246

Table 9
Main frequency components of maximized inputs with low initial amplitude - 3×131. The first three maximized inputs contain high frequency components. As for 3×91 model, 8Hz components are associated with activation of the interictal class.

Filter idx	F7-T7	F8-T8	T8-P8	T7-P7	pred.	loss
18	[98, 72]	[98, 72]	[98, 72]	[98, 72]	0.999	0.608
26	[72, 98]	[72, 98]	[72, 98]	[72]	0.981	0.593
21	[98]	[98, 15]	[98]	[99]	0.999	0.460
2	[14, 5]	[15, 5]	[15, 5]	[15, 5]	0.998	0.314
13	[5]	[5]	[5]	[5]	0.669	0.265
0	[8]	[8]	[8]	[8]	0.050	0.258
1	[2]	[2]	[2]	[2]	0.580	0.258
30	[5]	[5]	[5]	[5]	0.993	0.246
31	[5]	[5]	[5]	[5]	0.998	0.233

5. Discussion

The methodology presented was elaborated to meet specific requirements. First, we aimed for an online seizure characterization method for potential wearable device applications. Second, the method needed to handle the high patient inter-variability in terms of EEG signals and seizure expression. Finally, since clinical applications commonly require interpretable models, we progressed towards this goal and visualized some features learned by the model to compare them with traditional reading of raw EEG in the context of seizure detection.

5.1. Evaluation of the methodology

Short segments of raw EEG from temporal electrodes were directly fed to the network to classify ictal and interictal signals. The output probabilities were aggregated using Bayesian reasoning or a difference

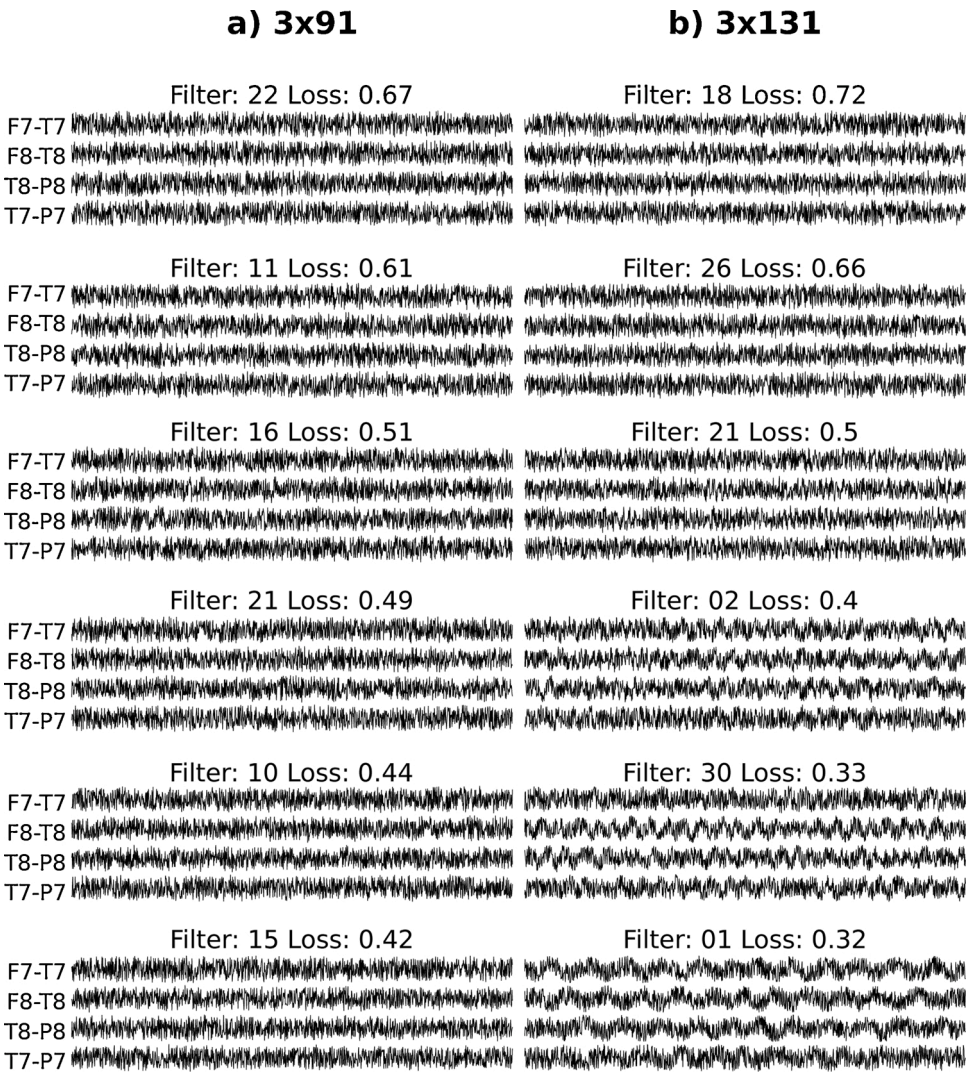

Fig. 9. Most significant maximized inputs for first layer kernels with high initial amplitude. Maximized inputs are now initialized with random samples with a magnitude up to 100μV. High amplitude initialization leads to amplification of high frequencies and stronger activation.

Table 10
Main frequency components of maximized inputs with high initial amplitude - 3×91. Most maximized inputs carry high-gamma frequency components. All associated probabilities are close to 1 and components around 8Hz lead to the lowest prediction probability.

Filter idx	F7-T7	F8-T8	T8-P8	T7-P7	pred.	loss
22	[96]	[96]	[97]	[97]	0.999	0.669
11	[97]	[97]	[97]	[97]	0.999	0.607
16	[97]	[96]	[97]	[97]	0.999	0.514
21	[97]	[97]	[97]	[97]	0.999	0.494
10	[72, 97]	[97, 73]	[97, 72]	[72, 97]	0.999	0.445
15	[97]	[97, 72]	[96, 72]	[72, 98]	0.999	0.423
12	[14]	[14]	[14]	[14]	0.999	0.401
4	[15]	[15, 4]	[14, 4]	[]	0.999	0.375
6	[5]	[4]	[5]	[6]	0.999	0.356
[...]	[...]	[...]	[...]	[...]	[...]	[...]
3	[8]	[8]	[8]	[8]	0.904	0.325

Table 11
Main frequency components of maximized inputs with high initial amplitude - 3×131. All the maximized inputs looking noisier, this model does not have a stronger proportion of high-gamma frequency components. As for the 3×91 model, most maximized inputs lead to strong activation of the ictal class. 8Hz components are associated with the lowest output probability.

Filter idx	F7-T7	F8-T8	T8-P8	T7-P7	pred.	loss
18	[98]	[98, 72]	[98, 71]	[98]	0.999	0.717
26	[72]	[72, 98]	[71, 98]	[71]	0.999	0.661
21	[99, 16]	[99, 15]	[99, 15]	[99, 16]	0.999	0.505
2	[5]	[4, 15]	[4, 15]	[]	0.999	0.396
30	[5]	[5]	[5]	[5]	0.999	0.332
1	[2]	[3]	[3]	[2]	0.994	0.317
13	[4]	[4]	[5]	[5]	0.999	0.316
31	[6]	[6]	[5, 7]	[5]	0.999	0.314
0	[8]	[8]	[8]	[8]	0.946	0.298

filter to detect seizure onsets. We trained three networks with different first layer kernels to study how the model processes EEG in the early layers. At the segment-level, all the models had an AUC score of 0.87, and the best F1-score was 0.76 for the 3×131 model with a low decision threshold. Performance at the segment-level is lower than the performance reported in [15] which could be explained by a different segmentation strategy and a simpler model architecture in our study. Direct comparison is however not possible as this study constitutes the first DL work on the $REPO_2MSE$ dataset. The best F1-score at the seizure-level was 0.873 using the difference filter and the 3×131 model, because both the model and the method reduced the false positive rate while preserving high sensitivity. The sensitivity of our

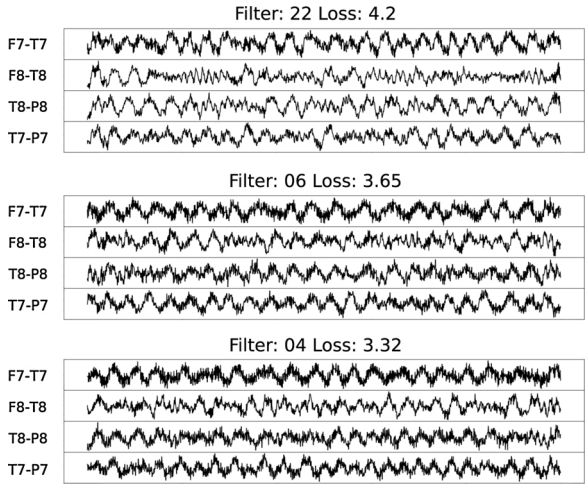

Fig. 10. Maximized inputs for kernels of the last convolutional block - 3×5 model. The three maximizing inputs eliciting the strongest activation after the last convolutional block of the 3×5 model are represented. This shows that the 3×5 model learned to make implicit representation of low frequency components of the EEG signal.

Table 12
Frequency components of top maximized inputs for last convolutional block of the 3×5 model. The preferred inputs have frequency components as low as 4Hz, showing that the model built a representation of low frequency components through pooling operations.

Filter idx	F7-T7	F8-T8	T8-P8	T7-P7	pred.	loss
22	5.0	5.0	4.0	5.0	0.999	4.174
6	5.0	5.0	4.0	5.0	0.999	3.661
27	5.0	5.0	5.0	5.0	0.983	3.294

patient-independent method is in line with related studies [13].

The difference filter was employed to potentially detect an evolution of the brain activity before the seizure onset label. Fig. 7 shows that this method can detect seizure onsets mostly right after the true label. Therefore, early drifts from the interictal to the immediate pre-ictal parts of the signal can not be identified by the model in the current segmentation paradigm. Classifying between ictal and interictal signals and using only 1 minute of interictal signal is most likely not suited to have a clear separation between the interictal and the immediate pre-ictal segments.

Misdetected seizures are generally consistent across all the three models suggesting that none of them was able to identify seizure events not identifiable by others. Because of patient inter-variability, general solutions are usually preferred to minimize the loss function, and signal with uncommon seizure patterns are likely to be misclassified.

5.2. *Decision interpretability*

We employed two visualization methods to explore how the kernels of the first layer were contributing to the final decision and to highlight ictal features on the input EEG. The first used gradient ascent to generate the preferred input to the first layer kernels and analysing the frequency components of the generated signals. The second used propagation of activation difference between an input and a baseline EEG to visualize decisive features to the ictal class. Results showed that most frequencies extracted in the first layer are strongly associated with the ictal class, while fewer ones are associated with the interictal class. Frequencies around 8 Hz lead to interictal classification, matching the common association of the alpha band with resting brain activity [32]. When increasing the amplitude of the input in the gradient ascent experiment, prediction of the network were strongly polarized toward the ictal class. Filters containing high-frequency components yield stronger activation in this case. Therefore, high amplitude is one main feature learned by the models and is possibly extracted with filters containing high-gamma frequencies. A previous study focusing on spectral bands for seizure classification showed that high-gamma frequencies were also important features to discriminate between pre-ictal and interictal segments [33]. Since high-gamma frequencies are key features for classification and can be detected with short convolutional windows, it can explain why the 3×5 model performs equally good at the segment level as the other models. Yet, high amplitude in the input signal is not sufficient nor necessary as some low amplitude patterns are detected as ictal in the DeepLIFT visualization, while some high amplitude patterns are not

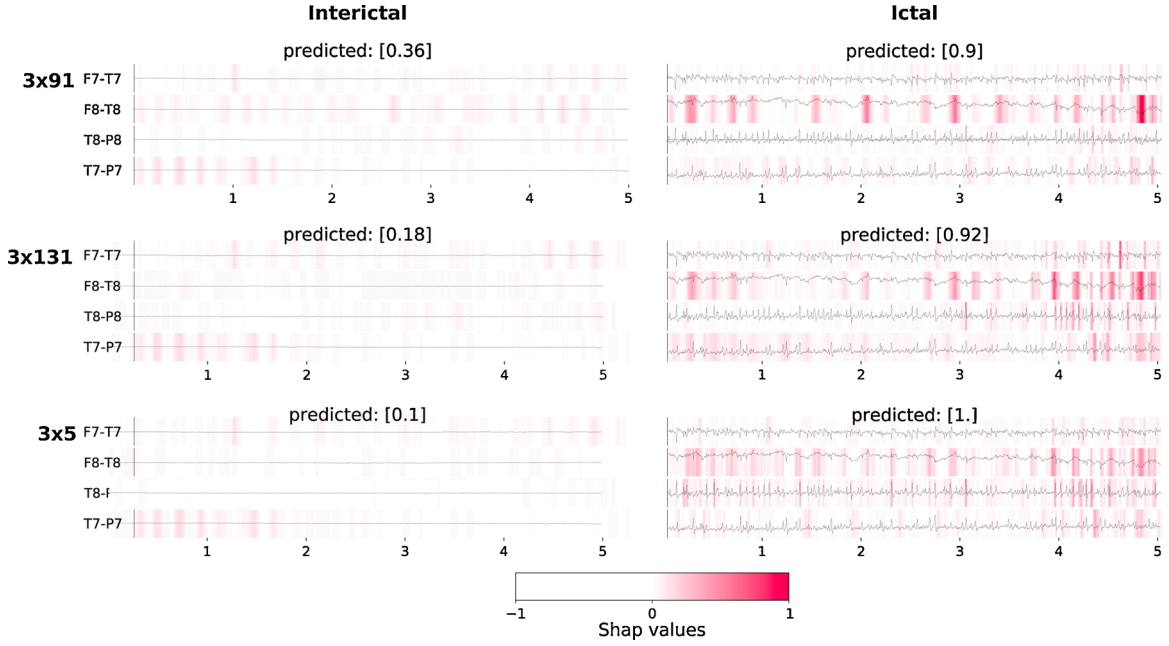

Fig. 11. Comparison of approach visualization across models and time. Both windows are taken from the same signal and fed to all three networks. Shap values are then computed and over-layed on the EEG signal after smoothing. Red bands correspond to cluster of positive SHAP values and indicate decisive ictal features.

detected as such. High-gamma frequencies may also be responsible for detecting different shapes of spikes in the EEG signals. It was expected that most features contributed to an ictal classification as they are easily identifiable to the naked-eye but learning resting EEG features was also useful as a counter-balancing information. Our analysis does not exclude that other components of the maximized inputs, such as waveform, phase and spatial combination can be key features of ictal classification.

Generalized seizures are often characterized by synchronous brain activity which can be observed on EEG signals. In Fig. 8, some maximized inputs show a correlation in the shape of the signals between different channels. This phenomenon is also observed on the highlighted features of the DeepLIFT method as observed in Fig. 11. Since the first layer kernels span 3 channels at each pass on the input data, the identification of correlated patterns across channels is likely to occur, especially with large kernel sizes in the time dimension. However, it is arguable whether the correlation is explicitly learned as a feature or if the same features are detected in different channels.

Both visualization methods indicate that the model focuses on amplitude differences, the spikes of high frequency, and the contrast with the alpha band activity and low amplitude phases, which all relate to the current expertise in seizure detection. Additionally, the optimal analysis window W^* in the Bayesian aggregation method is between 12.5 and 15 seconds, which also corresponds to the typical length of screen windows used by neurologists when reading EEG. Additionally, Hartmann and colleagues showed that every layer specializes in different frequency ranges. Our results show that indeed frequency components maximizing filters of the last convolutional block are different than of the first block in the 3×5 model. However, late layers of the model extract low frequency components while the opposite is shown in [24].

The visualization methods can also help to understand the performance differences between the three models. Results showed an antithetical behavior between the 3×131 and the 3×5 models in several steps of the methodology. First, at the segment-level, the 3×131 model showed a higher precision, while the 3×5 model showed a higher sensitivity for a similar F1-score. This was also highlighted by different output probability distributions, where the 3×5 model was more shifted towards the positive class than the 3×131 one. At the seizure-level, this behavior was counter-balanced by the optimized hyper-parameters, but would still be observed when fixing the hyperparameters in both aggregation methods. Maximized inputs bring additional arguments to the difference between both models. The 3×131 model can detect frequencies as low as 2 Hz in the first layer. Since frequencies in the alpha band are associated with interictal features, the 3×131 model can extract more information of the interictal class, increasing its precision. On the contrary, the 3×5 model only detects frequencies above 51 Hz at the first layer. The 3×5 model is then focusing primarily on ictal features in early layers. The higher sensitivity of 3×5 model is also verified when visualizing the decisive features in the input signals as SHAP values show sharper and more numerous bands contributing to the ictal class. Depending on the application needed, one can exploit the contrasting behavior of the two models to either reduce the false positive rate or to increase the sensitivity.

6. Conclusions

The goal of this study was to develop a DL-based methodology for online seizure event characterization able to handle inter-patient variability, and to explore some parameters of the model behavior from the interpretability point of view, including the problem of moving from a segment-level classification to a seizure-level classification. We demonstrated that the kernel size in the first layer is not significantly affecting the model performance, but a larger kernel size enables the study of the model behavior more thoroughly. We also provided insights on the features learned by the model by first observing the behavior of the first layer kernels and their maximized inputs and by highlighting the learned features back on the EEG input signal. Regarding the detection performance, our methodology was successfully able to generalize patient inter-variability for the majority of the population, and we found that the optimal time scale required for seizure-level classification is similar to that used by human experts when reading EEG signals. Moreover, the resulting model may be implemented in a wearable device with low energy requirements. Future developments should focus on the causality between important frequency components and the decision probability at the different internal states of the network and on handling classification of different sub-populations of seizures within a patient cohort to improve the generalization of the methodology.

Conflict of interest statement

All authors declared no conflicts of interest.

Acknowledgement

This work has been supported by the H2020 DeepHealth Project (GA No. 825111) and by theH2020 RECIPE project(GA No. 801137). The REPO$_2$MSE cohort was funded by the French Ministery of Health (PHRC national 2009) and sponsored by Hospices Civils de Lyon, and involved the following investigators: Philippe Ryvlin (Lyon, PI), Sylvain Rheims (Lyon, co-PI), Philippe Derambure (Lille), Edouard Hirsch (Strasbourg), Louis Maillard (Nancy), Francine Chassoux (Paris-St-Anne), Arnaud Biraben (Rennes), Cécile Marchal (Bordeaux), Luc Valton (Toulouse), Fabrice Bartolomei (Marseille), Jérôme Petit (La Teppe), Vincent Navarro (Paris-Salpétrière), Philippe Kahane (Grenoble), Bertrand De Toffol (Tours), Pierre Thomas (Nice).

References

[1] World Health Organization. Fact sheet on epilepsy. 2019. URL https://www.who.int/news-room/fact-sheets/detail/epilepsy.
[2] Aghaei-Lasboo A, Fisher RS. Methods for measuring seizure frequency and severity. Neurol Clin 2016;34(2):383–94. https://doi.org/10.1016/j.ncl.2015.11.001.
[3] Maganti RK, Rutecki P. EEG and epilepsy monitoring. Continuum (Minneap Minn) 2013;19:598–622. https://doi.org/10.1212/01.CON.0000431378.51935.d8.
[4] Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J. Deep learning-based electroencephalography analysis: a systematic review. J Neural Eng 2019;16(5):051001. https://doi.org/10.1088/1741-2552/ab260c.
[5] Adadi A, Berrada M. Explainable AI for healthcare: from black box to interpretable models. In: Bhateja V, Satapathy SC, Satori H, editors. Embedded systems and artificial intelligence, advances in intelligent systems and computing. Singapore: Springer; 2020. https://doi.org/10.1007/978-981-15-0947-6_31.
[6] Sopic D, Aminifar A, Atienza D. e-Glass: a wearable system for real-time detection of epileptic seizures. 2018 IEEE international symposium on circuits and systems (ISCAS), IEEE, Florence 2018:1–5. https://doi.org/10.1109/ISCAS.2018.8351728.
[7] Fisher RS, Blum DE, DiVentura B, Vannest J, Hixson JD, Moss R, et al. Seizure diaries for clinical research and practice: limitations and future prospects. Epilepsy Behav 2012;24(3):304–10. https://doi.org/10.1016/j.yebeh.2012.04.128.
[8] Blachut B, Hoppe C, Surges R, Stahl J, Elger CE, Helmstaedter C. Counting seizures: the primary outcome measure in epileptology from the patients' perspective. Seizure 2015;29:97–103. https://doi.org/10.1016/j.seizure.2015.03.004.
[9] Conradsen I, Beniczky S, Hoppe K, Wolf P, Sams T, Sorensen HBD. Seizure onset detection based on one sEMG channel. 2011 annual international conference of the IEEE engineering in medicine and biology society. Boston, MA: IEEE; 2011. p. 7715–8. https://doi.org/10.1109/IEMBS.2011.6091901.
[10] Ming-Zher Poh, Loddenkemper T, Swenson NC, Goyal S, Madsen JR, Picard RW. Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor. 2010 annual international conference of the IEEE engineering in medicine and biology 2010:4415–8. https://doi.org/10.1109/IEMBS.2010.5625988.
[11] Ryvlin P, Cammoun L, Hubbard I, Ravey F, Beniczky S, Atienza D. Noninvasive detection of focal seizures in ambulatory patients. Epilepsia 2020. https://doi.org/10.1111/epi.16538. epi.16538.
[12] Tsiouris KM, Pezoulas VC, Zervakis M, Konitsiotis S, Koutsouris DD, Fotiadis DI. A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. Comput Biol Med 2018;99:24–37. https://doi.org/10.1016/j.compbiomed.2018.05.019.
[13] Truong ND, Nguyen AD, Kuhlmann L, Bonyadi MR, Yang J, Ippolito S, et al. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. Neural Netw 2018;105:104–11. https://doi.org/10.1016/j.neunet.2018.04.018.

[14] Ullah I, Hussain M, Qazi E-u-H, Aboalsamh H. An automated system for epilepsy detection using EEG brain signals based on deep learning approach. Expert Syst Appl 2018;107:61–71. https://doi.org/10.1016/j.eswa.2018.04.021.
[15] Yuan Y, Xun G, Jia K, Zhang A. A multi-view deep learning framework for EEG seizure detection. IEEE J Biomed Health Inform 2019;23(1):83–94. https://doi.org/10.1109/JBHI.2018.2871678.
[16] Yuan Y, Xun G, Ma F, Suo Q, Xue H, Jia K, et al. A novel channel-aware attention framework for multi-channel EEG seizure detection via multi-view deep learning. 2018 IEEE EMBS international conference on biomedical & health informatics (BHI). Las Vegas, NV, USA: IEEE; 2018. p. 206–9. https://doi.org/10.1109/BHI.2018.8333405.
[17] de Aguiar K, Franca FMG, Barbosa VC, Teixeira CAD. Early detection of epilepsy seizures based on a weightless neural network. Annual international conference of the ieee engineering in medicine and biology society. IEEE engineering in medicine and biology society. annual conference 2015 2015:4470–4. https://doi.org/10.1109/EMBC.2015.7319387.
[18] Antoniades A, Spyrou L, Took CC, Sanei S. Deep learning for epileptic intracranial EEG data. 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP). Vietri Sul Mare, Salerno, Italy: IEEE; 2016. p. 1–6. https://doi.org/10.1109/MLSP.2016.7738824.
[19] Avcu MT, Zhang Z, Shih Chan DW. Seizure detection using least eeg channels by deep convolutional neural network. ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). Brighton, United Kingdom: IEEE; 2019. p. 1120–4. https://doi.org/10.1109/ICASSP.2019.8683229.
[20] Bai S, Kolter J, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. 2018 03.
[21] Klatt J, Feldwisch-Drentrup H, Ihle M, Navarro V, Neufang M, Teixeira C, Adam C, Valderrama M, Alvarado-Rojas C, Witon A, Le Van Quyen M, Sales F, Dourado A, Timmer J, Schulze-Bonhage A, Schelter B. The EPILEPSIAE database: an extensive electroencephalography database of epilepsy patients, 53; 2012. p. 1669–76.
[22] Shoeb AH, Guttag J. Application of machine learning to epileptic seizure detection. In: In Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10); 2010. p. 975–82.
[23] Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glasstetter M, Eggensperger K, Tangermann M, et al. Deep learning with convolutional neural networks for EEG decoding and visualization. Hum Brain Mapp 2017;38(11):5391–420. https://doi.org/10.1002/hbm.23730.
[24] Hartmann KG, Schirrmeister RT, Ball T. Hierarchical internal representation of spectral features in deep convolutional networks trained for EEG decoding. 2018 6th international conference on brain-computer interface (BCI). GangWon: IEEE; 2018. p. 1–6. https://doi.org/10.1109/IWW-BCI.2018.8311493.
[25] Sturm I, Lapuschkin S, Samek W, Müller K-R. Interpretable deep neural networks for single-trial EEG classification. J Neurosci Methods 2016;274:141–5. https://doi.org/10.1016/j.jneumeth.2016.10.008.
[26] Rheims S, Alvarez BM, Alexandre V, Curot J, Maillard L, Bartolomei F, Derambure P, Hirsch E, Michel V, Chassoux F, Tourniaire D, Crespel A, Biraben A, Navarro V, Kahane P, De Toffol B, Thomas P, Rosenberg S, Valton L, Bezin L, Ryvlin P. Hypoxemia following generalized convulsive seizures. Neurology 2019;92(3):e183–93. https://doi.org/10.1212/WNL.0000000000006777.
[27] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015 Feb.
[28] Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. SIGKDD Explor 2010;12(1):49–57. https://doi.org/10.1145/1882471.1882479.
[29] Welch P. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust 1967;15(2):70–3. https://doi.org/10.1109/TAU.1967.1161901.
[30] Lundberg SM, Lee S-I, et al. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, editors. Advances in neural information processing systems 30. Curran Associates, Inc; 2017. p. 4765–74.
[31] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. 2017 04.
[32] Moini J, Piran P. Chapter 6 - cerebral cortex. In: Moini J, Piran P, editors. Functional and Clinical Neuroanatomy. Academic Press; 2020. p. 177–240. https://doi.org/10.1016/B978-0-12-817424-1.00006-9.
[33] Park Y, Luo L, Parhi KK, Netoff T. Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. Epilepsia 2011;52(10):1761–70. https://doi.org/10.1111/j.1528-1167.2011.03138.x.