# Towards Early Diagnosis of Epilepsy from EEG Data

**Diyuan Lu**[1,2,3] ELU@FIAS.UNI-FRANKFURT.DE

**Sebastian Bauer**[3,4] SEBASTIAN.BAUER@KGU.DE

**Valentin Neubert**[5] VALENTIN.NEUBERT@UNI-ROSTOCK.DE

**Lara Sophie Costard**[6] LARACOSTARD@RCSI.COM

**Felix Rosenow**[3,4] ROSENOW@MED.UNI-FRANKFURT.DE

**Jochen Triesch**[1,2,3] TRIESCH@FIAS.UNI-FRANKFURT.DE

[1]**Frankfurt Institute for Advanced Studies (FIAS), Frankfurt am Main, Germany**
[2]**Goethe University Frankfurt, Frankfurt am Main, Germany**
[3]**Center for Personalized Translational Epilepsy Research (CePTER), Frankfurt am Main, Germany**
[4]**Epilepsy Center Frankfurt Rhein-Main, University Hospital Goethe-University, Frankfurt am Main, Germany**
[5]**Oscar Langendorff Institute of Physiology, Rostock University Medical Center, Rostock, Germany**
[6]**Tissue Engineering Research Group, Royal College of Surgeons Ireland, Dublin, Ireland**

## Abstract

Epilepsy is one of the most common neurological disorders, affecting about 1% of the population at all ages. Detecting the development of epilepsy, i.e., epileptogenesis (EPG), before any seizures occur could allow for early interventions and potentially more effective treatments. Here, we investigate if modern machine learning (ML) techniques can detect EPG from intra-cranial electroencephalography (EEG) recordings prior to the occurrence of any seizures by a time frame of days or even weeks. We study a common form of epilepsy called mesial temporal lobe epilepsy (mTLE). Specifically, we use a rodent mTLE model where EPG is triggered by electrical stimulation of the brain, which induces hippocampal damages that resemble those in human patients. We propose a ML framework for EPG identification, which combines a deep convolutional neural network (CNN) with a prediction aggregation method to obtain the final classification decision. Specifically, the neural network is trained to distinguish five second segments of EEG recordings taken from either the pre-stimulation period or the post-stimulation period. Due to the gradual development of epilepsy, there is enormous overlap of the EEG patterns before and after the stimulation.

Hence, a prediction aggregation process is introduced, which pools predictions over a longer period. By aggregating predictions over one hour, our approach achieves an area under the curve (AUC) of 0.99 on the EPG detection task. This demonstrates the potential of ML for EPG prediction from EEG recordings.

## 1. Introduction

Identifying patients at high risk of developing epilepsy (epileptogenesis) is of great importance to allow early medical intervention and improve the effectiveness of anti-epileptogenic treatments. In many acquired epilepsy cases, there is a latent period between the brain injury and the onset of spontaneous recurring seizures. During this latent period, affected brain tissue is thought to transform such that it eventually can generate spontaneous seizures (Pitkänen and Engel, 2014). Over 30% of the patients will be pharmaco-resistant and continue to suffer from recurring seizures despite intake of medications (Kwan and Brodie, 2000). The more seizure episodes have occurred before the first clinical visit, the less effective of the treatment will be (Kwan and Brodie, 2000). Hence, identifying the presence of EPG before the epilepsy is fully established would be of great importance. However, the process of EPG is still not fully understood (Pitkänen et al., 2016). The precise time of onset of the brain being epileptogenic is untraceable (Pitkänen and Engel, 2014). However, it is safe to say that any anti-epileptogenic or disease-modifying therapies should be administered as early as possible (Löscher, 2019). Thus, discovering prominent features of EPG could facilitate early diagnosis and open the door for early interventions (Moshé et al., 2015).

Electroencephalography (EEG) is a common tool in the clinic due to its non-invasive and easy-to-deploy properties. However, detecting EPG from EEG data is challenging. Two reasons are the complexity of the mechanisms of EPG and the immense cross-subject variability, which result in different phenotypes of EEG signals. This makes reliable interpretation of EEG signals from previously unseen individuals difficult.

Some works have attempted to identify electrophysiological biomarkers of EPG based on various hand-selected features (Bentes et al., 2018; Rizzi et al., 2019; Milikovsky et al., 2017; Bragin et al., 2004, 2016). However, a manual selection of features may be biased and overlook useful information. Recently, fueled by advances in ML, impressive results have been achieved in a variety of domains by training on raw data and letting the learning algorithm identify useful features automatically. Such approaches can even outperform human experts (Hannun et al., 2019; Haenssle et al., 2018; Sarker et al., 2018).

Here, we recorded intracranial EEG signals from a rodent model of mesial temporal lobe epilepsy with hippocampal sclerosis (mTLE-HS) (Costard et al., 2019). In this model, epilepsy was induced by electrical perforant pathway stimulation (PPS) through depth electrodes. Continuous EEG recordings were obtained from the hilus of the dentate gyrus after the implantation of the electrode until the occurrence of the first spontaneous seizure (FSS). The EEG recordings were divided into two classes depending on the time of recording relative to the PPS stimulation. The samples recorded before the epilepsy-triggering PPS define the *baseline* (BL) class. The samples recorded after the PPS, but before occurrence of the FSS form the *epileptogenesis* (EPG) class. In the following, we propose a deep learning

(DL) framework to classify EPG vs. BL by training on raw EEG data in an end-to-end fashion.

To tackle the problem that a large portion of normal brain EEG patterns are also present in the EPG phase, we propose a prediction aggregation method where predictions from a longer time interval (e.g. one hour) are pooled together through a linear aggregation. We assume that the "EPG-typical" signals should be more frequent during the EPG phase compared to the baseline phase. This difference becomes apparent through the aggregation method. Specifically, we make the following contributions:

- We present the first attempt to identify the process of EPG with a deep neural network (DNN) trained on EEG time series data. This is a radical departure from the conventional (and hitherto not very successful) approach of attempting to predict individual seizures when the disease has already established itself.
- We propose a framework for EPG identification using massive amounts of EEG data from chronic recordings to maximally exploit the DNN's learning ability and minimize human effort in data labeling and feature engineering.
- We use a prediction aggregation method and demonstrate that it achieves high fidelity EPG detection in a rodent model.

**Generalizable Insights about ML in the Context of Healthcare**

Massive expert annotations are expensive and therefore often scarce in medical contexts. This poses tremendous difficulties for the application of ML. When large amounts of data can be collected but labelling by experts is infeasible, turning to a form of "cheap" labelling can be a way-out. In our case, detailed expert annotations are absent but the EEG signals are recorded continuously (24/7), which yields a large quantity of training data. We define the labels exclusively according to the relative time of the recording with respect to the PPS. This kind of label is cheap and easy to obtain but less informative, since in the EPG period large amounts of normal brain activity are still present, i.e., the data from the two classes are largely overlapping. To deal with this large overlap, we propose a prediction aggregation process to pool decisions over a long time window. We show here that even in the complete absence of expert annotations of specific events showing "EPG-typical" brain activity, the large data set in combination with the "cheap" labels allow us to build a powerful classification system. We suggest that many other medical problems where the application of ML is currently infeasible due to lack of detailed expert annotations could be tackled using similar methods. More generally, our approach of massive data collection to identify the earliest signatures of a developing disease may enable early diagnosis and intervention across a wide range of medical contexts.

## 2. Related Work

**EEG Analysis with Deep Learning** Modern ML techniques allow an end-to-end learning approach to the analysis of EEG data rather than relying on specific handcrafted features. In particular, DNNs have been applied to either frequency representations (Lu et al.,

2019; Thodoroff et al., 2016) or directly to raw EEG data in the time domain (Kiral-Kornek et al., 2018; Biswal et al., 2019; Avcu et al., 2019; Farahat et al., 2019; Bi and Wang, 2019). They have achieved promising results in seizure detection, seizure prediction, or even other neurological disorders such as Alzheimer's disease and Autism classification. For example, Zhou et al. (2018) compared the performance of a CNN on the EEG signal classification problem with time-domain and frequency-domain input and concluded that frequency-domain signals have greater potential for the task. Kiral-Kornek et al. (2018) demonstrated an accurate, automated patient-specific seizure prediction approach with a DNN trained on intracranial EEG data. Biswal et al. (2019) applied stacked CNNs and recurrent neural networks (RNNs) to extract temporal shift invariant features from EEG data. These features are used to classify multiple key EEG phenotypes. Avcu et al. (2019) developed an end-to-end solution for seizure onset detection. Bi and Wang (2019) applied a convolutional deep Boltzmann machine with EEG data in early diagnosis of Alzheimer's disease. Thodoroff et al. (2016) applied a deep RNN with a CNN to perform automated patient specific seizure detection with scalp EEG. A deep CNN was applied for EEG signal decoding during human decision making and demonstrates promising results (Farahat et al., 2019). These studies demonstrated the application of DL for EEG analysis.

Here, we want to emphasize the fundamental difference between seizure prediction and our task. The goal of epileptic seizure prediction is to predict the onset of individual seizures in an epileptic brain that already generates spontaneous seizures. The goal is typically to predict individual seizures several minutes before their occurrence, so the patient can be warned about the imminent seizure and take precautions. In contrast, we aim to detect if a brain is on its way to develop an epilepsy *before* the FSS has occurred, i.e. before an epilepsy is manifest. If this could be done several days or weeks before the FSS, this would allow for interventions that could slow down or even prevent the development of the disease, before spontaneous seizures occur.

**EPG Biomarkers in EEG** There have been several previous studies on biomarker discovery for identifying EPG. Bragin et al. (2004) found that the occurrence of high-frequency-oscillations (HFOs) is a strong indicator of future recurrent spontaneous seizures and the sooner HFOs occur, the shorter the EPG period will be. Andrade et al. (2017) found that a duration reduction of sleep spindles at the transition from stage III to rapid-eye-movement sleep indicates potential post-traumatic epilepsy in a lateral fluid-percussion rat model. In humans, it was shown that over 90% of the HFO area overlapped with the seizure onset zone for six patients (Burnos et al., 2014). Milikovsky et al. (2017) revealed that the dynamics of the theta band could predict future post-injury epilepsy and the seizure onset and thus could serve as a diagnostic biomarker for EPG. Lu et al. (2019) demonstrated that an increased delta band power, a decrease of theta band power as well as an increase of high gamma band power were correlated with the presence of EPG in a rat mesial temporal lobe epilepsy model. Rizzi et al. (2019) recently showed using concepts from nonlinear dynamics, that a reduction of the dimensionality of EEG/ECoG signals indicates the presence and the severity of EPG in three different rodent epilepsy models. Finally, Bentes et al. (2018) found that an asymmetry in background EEG signals and interictal epileptiform discharges can independently predict post-stroke epilepsy in a clinical study. However, so far a DL-based approach to EPG biomarker discovery in an end-to-end fashion has not yet been attempted.

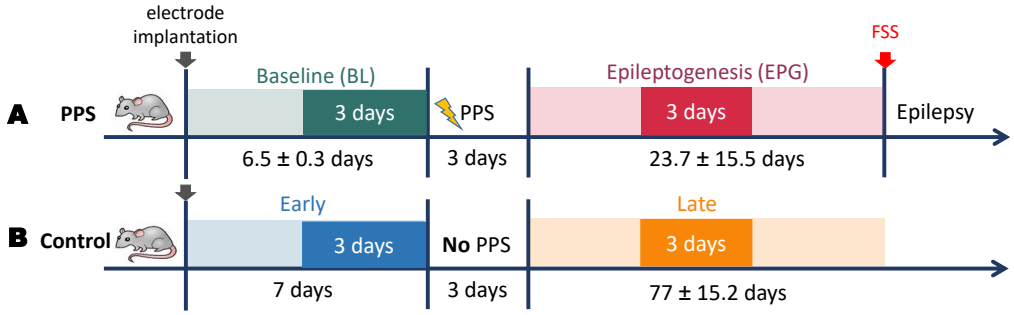


Figure 1: Schematic of the timeline of the experiment. A. time line for PPS-stimulated rats. B: time line for control rats. PPS: perforant pathway stimulation. FSS: first spontaneous seizure. The mean and standard deviation of the duration of EPG in the PPS group is $23.7 \pm 15.5$ days (min. 10 days, max. 56 days).

## 3. Methods

### 3.1. Dataset

We used intracranial EEG data recorded continuously (24/7) by a depth-electrode from a rodent mTLE-HS model, where epilepsy is induced by electrical PPS, as described in detail in Costard et al. (2019). The stimulated rats developed epilepsy after an average EPG phase of four weeks (range one to eight weeks). The EPG phase ended with the FSS.

The rat model provides an opportunity to study the progression of epilepsy and to discover potential biomarkers of EPG in the EEG. In this study, we included seven PPS-treated rats with continuous wireless EEG recordings. We also included three control rats which had electrodes implanted but did not undergo PPS and did not develop epilepsy by the end of the recording (limited by the lifetime of battery of the wireless transmitter). The time-lines for the PPS group and the control group are shown in Fig. 1. We denoted two phases of interest from the continuous recording, i.e., baseline (BL) and epileptogenesis (EPG) in the PPS group. In our study, we selected the last three days from the BL phase and three days from the EPG phase, highlighted in the colored boxes, and assigned them the labels "0" and "1", respectively. We selected the 7th, 8th and 9th day of EPG for training for all rats. Reasons for this choice are 1) to maintain the maximum time distance to acute symptomatic seizures which can occur within the first 1-3 days after the PPS, and 2) the rat with the shortest EPG duration developed its FSS on the 10th day after PPS and we wanted to keep the time window from which we get the class "1" signals the same across all rats.

**Preprocessing** The sampling rate of the EEG recordings was 512 Hz. A band-pass filter between 0.5 - 160 Hz and a notch filter at 50 Hz were applied to the raw data. In our experimental setting, the recorded EEG signals were susceptible to electric interference, which resulted in extremely high amplitude outliers. To fix this problem, we applied a MATLAB function, i.e., `filloutliers` [1] with the configuration `method = 'pchip';`

1. https://www.mathworks.com/help/matlab/ref/filloutliers.html

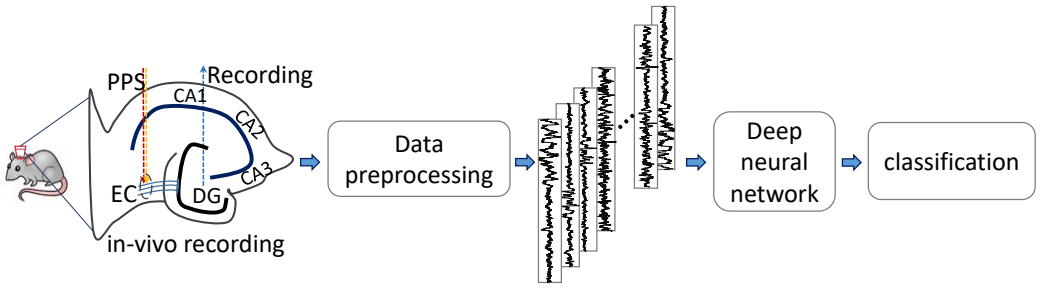


Figure 2: Workflow of our proposed framework. EC: entorhinal cortex, DG: dentate gyrus, CA: cornu ammonis, PPS: perforant pathway stimulation.

`movmethod = 'movmedian'; window = 50` to filter out these outliers. We obtained non-overlapping five-second segments from the continuous recordings. To clean up the data for training, those segments with more than 20% data loss due to weak wireless transmission were discarded. Then, those five-second segments were normalized via the z-score method from `scipy.stats.zscore` before being fed into the neural network. The workflow is shown in Fig 2.

Our proposed method consists of two parts: (a) a deep residual neural network and (b) a prediction aggregation process during the testing.

**Residual convolutional neural network** Our model is a DNN with 33 convolutional layers with residual connections and it is inspired by the work of Hannun et al. (2019). The network's structure is shown in Table 1. The concept of residual connections was first proposed by He et al. (2016a) for an image recognition task and has been widely used in a variety of tasks such as image segmentation (Huang et al., 2017; Lei and Todorovic, 2018; Liu et al., 2019), visual object detection (Mordan et al., 2018; Wang et al., 2019), and healthcare-related applications (Hannun et al., 2019; Sarker et al., 2018). The residual connection connects the pre-activation from one layer with the input of another previous layer in an additive fashion skipping several layers in between. Then, the non-linear activation is applied to the sum to compute the input for the next layer. The collection of the computations between one residual connection is termed a block (ResBlock). The output of the network is a softmax layer taking the flattened feature maps as input and outputting a probability distribution over the two possible classes.

Before we started our official classifier training, we performed a hyper-parameter exploration for our specific task with a small randomly selected data set. A drop-out rate of 0.25 yielded the best performance among the values 0.2, 0.25, 0.3, 0.5, and 0.65. The number of blocks that performed best was 15 among 5, 7, 11, and 15. A filter size of 32 worked the best among values of 3, 9, 11, 16, 32, and 64. We tried ReLU and leaky ReLU as the nonlinear activation function and no significant difference was observed, so we chose the ReLU activation for this work. A starting number of 16 filters yielded better results than 8 and 32. After the network hyper-parameter exploration, we fixed the choices for further experiments.

Table 1: The network structure used in our work. The **Config** column show the filter size (always 32) and the number of filters we use in each convolutional layer. The number of filters is increased every four blocks by a factor of 2. Every other block sub-samples its input by a factor of 2, indicated by the value of **stride**. Here, the batch size at the first dimension is omitted in the output shape column

| **Name** | **Config** | **Stride** | **Factor i** | **Output shape** |
|---|---|---|---|---|
| Conv layer 0 | $\begin{bmatrix} 32\times 1, & 16\times 2^i \end{bmatrix}$ | 1 | 0 | [2560, 1, 16] |
| ResBlock 0 | $\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$ | 1 | 0 | [2560, 1, 16] |
| ResBlock 1 | $\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$ | 2 | 0 | [1280, 1, $16\times 2^i$] |
| ResBlock 2 | $\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$ | 1 | 0 | [1280, 1, $16\times 2^i$] |
| ResBlock 3 | $\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$ | 2 | 0 | [640, 1, $16\times 2^i$] |
| ResBlock 4 | $\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$ | 1 | 1 | [640, 1, $16\times 2^i$] |
| ResBlock (5,..., 8) | $\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$ | (2, 1, 2, 1) | (1, 1, 1, 2) | [320, 1, $16\times 2^i$] |
| ResBlock (9,..., 12) | $\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$ | (2, 1, 2, 1) | (2, 2, 2, 3) | [80, 1, $16\times 2^i$] |
| ResBlock (13, 14) | $\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$ | (2, 1) | (3, 3) | [20, 1, $16\times 2^i$] |
| Dense | 2 | | | [2] |

We adopted the pre-activation design from He et al. (2016b). The convolutional layer had a filter width of 32. The number of filters increased by a factor of 2 in every four blocks starting from 16. The feature maps were down-sampled in every other block with a stride of 2. To keep the dimensionality compatible, the max-pooling branch shared the same stride value as in the second convolutional layer in each block.

**Prediction aggregation** We hypothesize that the EPG phase may be better characterized by a change of distribution of different waveforms rather than a specific waveform that can be identified in every individual segment. Therefore a reliable classification can only be achieved by pooling information from many data segments. Our method is inspired by Smyth and Wolpert (1999). For each segment, the network outputs how likely this segment is taken from each class. We linearly aggregate the predictions for multiple consecutive segments to obtain the final classification result.

Considering the data pairs, the EEG segments are $x_{(h,i)}$ and the associated labels are $y_{(h,i)}$ in one continuous hour $h$, where $i = 1, \ldots, N$ and $N$ is the total number of the samples in this hour. The softmax output of these samples is given by $\hat{y}_{(h,i)} = f(x_{(h,i)}, \text{model})$ and it is in shape $[N, 2]$ where 2 is the number of classes in our supervised scheme. The aggregated

Table 2: Performance **without** (5 second) and **with** one hour of aggregation. Data are presented as mean ± standard deviation. SEN: sensitivity, SPE: specificity, AUC: area under the curve

| Aggregation length | Task | SEN | SPE | AUC |
|---|---|---|---|---|
| 5 second | Task A | $0.73 \pm 0.25$ | $0.77 \pm 0.17$ | $0.86 \pm 0.07$ |
| | Task B | $0.57 \pm 0.42$ | $0.43 \pm 0.42$ | $0.50 \pm 0.08$ |
| 1 hour | Task A | $\mathbf{0.94} \pm 0.05$ | $\mathbf{0.96} \pm 0.04$ | $\mathbf{0.99} \pm 0.01$ |
| | Task B | $0.63 \pm 0.45$ | $0.37 \pm 0.45$ | $0.45 \pm 0.06$ |

prediction for hour $h$ is given by $\hat{y}_h = \sum_{i=1}^{N} \hat{y}_{(h,i)} = \sum_{i=1}^{N} f(x_{(h,i)}, \text{model})$, and in shape of $[1, 2]$. In a final step, we normalize $\hat{y}_h$ along the column axis. The resulting number is interpreted as a class probability and used to compute corresponding performance metrics.

**Training procedure** We applied leave-one-out (LOO) cross validation to test the generalization ability of our approach for both the PPS group and the control group. Specifically, in each fold the data from one rat were completely withheld as the test set, and the data from the other six rats form the training and the validation sets. For training and validation, we randomly selected 25 hours from each phase and from each rat and applied a train-validation-split of 9:1. The choice of 25 hours represents a trade-off between computation cost and performance, chosen empirically. We tried training with the whole three-day recordings, and the computation time was increased by a factor of three without obvious performance improvement. After the network was trained, we tested it with the data from the previously held-out rat. The procedure was repeated seven times in the PPS group and three times in the control group and results were averaged for each group.

## 4. Experiments and Results

### 4.1. Experiment Design

To evaluate our methods ability to identify EPG, we designed two tasks: Task A is designed to classify BL vs. EPG signals in PPS rats as shown in Fig 1A. Task B is a control designed to classify signals recorded in the early and late implantation phases in the set of control rats as shown in Fig 1B.

**Task A: BL vs. EPG classification in PPS rats** This is our main task in which we want to distinguish EEG signals from BL and EPG phases. In this task, we applied seven-fold LOO cross validation with the data from the seven PPS-stimulated rats.

**Task B: *early* vs. *late* classification in control rats** In this control task we want to rule out the possibility that differences between BL and EPG in Task A could be simply due to systematic changes in the tissue after electrode implantation that have nothing to do with the EPG triggered by PPS. Therefore, we study control rats that do not undergo PPS (see Fig 1B) and analyze if there are systematic differences between the EEG signals

recorded from the early and late implantation phases. We applied a three-fold LOO cross validation scheme with the same network configuration as in Task A.

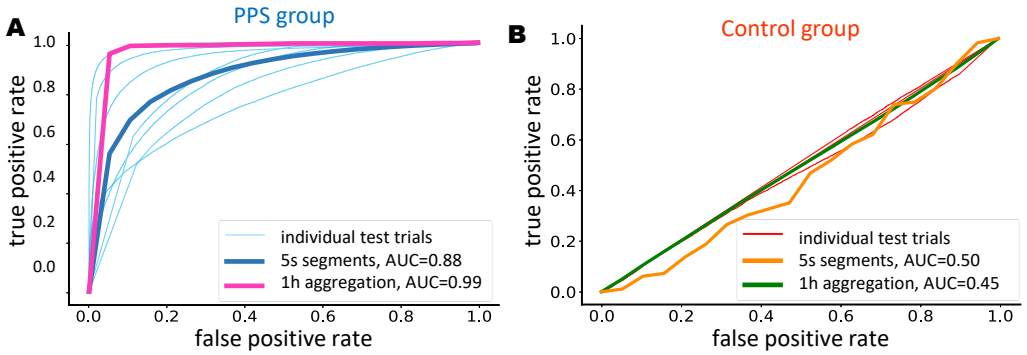


Figure 3: Receiver operating characteristic (ROC) curves. **A**: **The PPS group** (seven rats). Individual ROC curves from all LOO test trials (thin light blue), the average ROC curve without prediction aggregation (thick blue) and the average ROC curve with aggregation in a continuous stretch of one hour (thick pink). **B**: **The control group**. Individual ROC curves (thin light orange), the average ROC curve without aggregation (thick orange) and the average ROC curve with aggregation over one hour (thick green) from all LOO test trials in the control group. AUC: area under the curve. PPS: perforant pathway stimulation. LOO: leave-one-out

### 4.2. Results

#### 4.2.1. ROC Analysis

The average ROC curves of all the leave-one-out test trials in each task are shown in Fig 3. The AUC values are computed in two scenarios: a) each five second segment is viewed independently and the AUC is calculated based on the prediction of all the five second segments, b) the predictions of multiple consecutive five second segments are aggregated together through a linear stacking. In Fig 3A, we show the ROC curves in individual LOO test trials, and the averaged ROC curves with and without prediction aggregation. Our method could discern signals from both phases with an average AUC under the ROC curve of 0.88. It suggests that the neural network has learned features that are informative for the correct classification. With the proposed prediction aggregation over one hour, the average AUC achieves 0.99, which shows that the proposed approach can reliably discern EEG signals from the BL and the EPG phase. In contrast, for the control group, the *early* vs. *late* phase classification, the network does not show clear discriminative ability. The average AUCs from all the test trials with and without the prediction aggregation are 0.50 and 0.45, respectively. The detailed performance measurements such as sensitivity (SEN) $= \frac{TP}{TP+FN}$, specificity (SPE) $= \frac{TN}{TN+FP}$ and the AUC are shown in Table. 2, where *TP, TN, FP, FN* denote true positive, true negative, false positive and false negative, respectively.

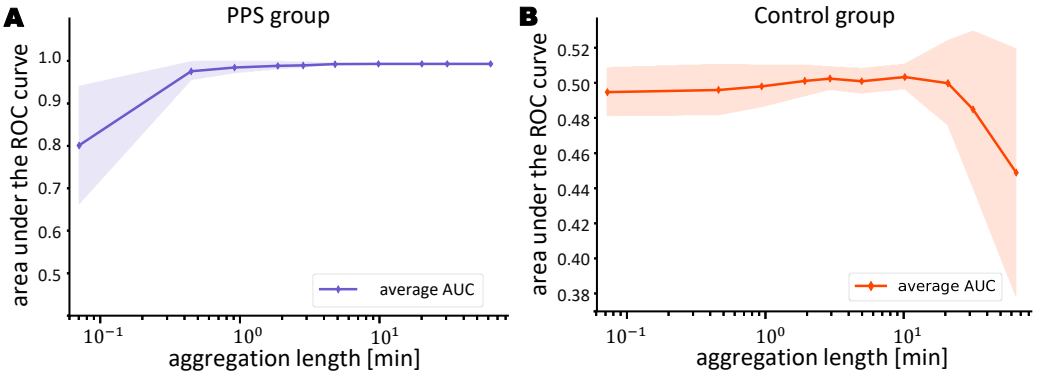


Figure 4: The average AUC over all leave-one-out test trials as a function of the aggregation length for the two groups. The shaded area represents one standard deviation.

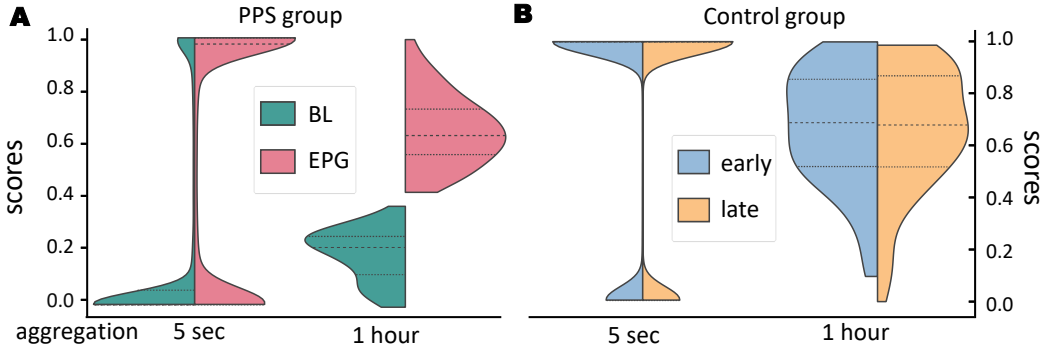


Figure 5: Example distributions of scores from both classes. **A**: **PPS group**. **(left)** without aggregation. The mean and variance of the two distributions, i.e., from all BL segments and all EPG segments, are different but overlapping. **(right)** with one hour of aggregation. **B**: **Control group**. **(left)** without aggregation. **(right)** with one hour of aggregation. BL: baseline. EPG: epileptogenesis

#### 4.2.2. Aggregation Effect

To further investigate the effect of aggregation, we computed the AUC value in each test trial with various intervals, i.e., five seconds, 30 seconds, one, two, five, ten, 20, 30, 60 minutes. The average AUC across all the test trials in the PPS group as a function of the aggregation lengths is shown in Fig 4A. It shows a clear trend of an increasing AUC and a decrease of standard deviation with a longer aggregation length. Thus, the prediction aggregation from multiple consecutive segments is essential for a strong performance in the PPS group. In contrast, in the control group, the aggregation not only did not help increase but reduced the average AUC, as depicted in Fig 4B.

We also tested if the seven neural networks trained on the PPS group would discriminate the *early* and *late* phase EEG patterns from the control animals. If so, this would suggest

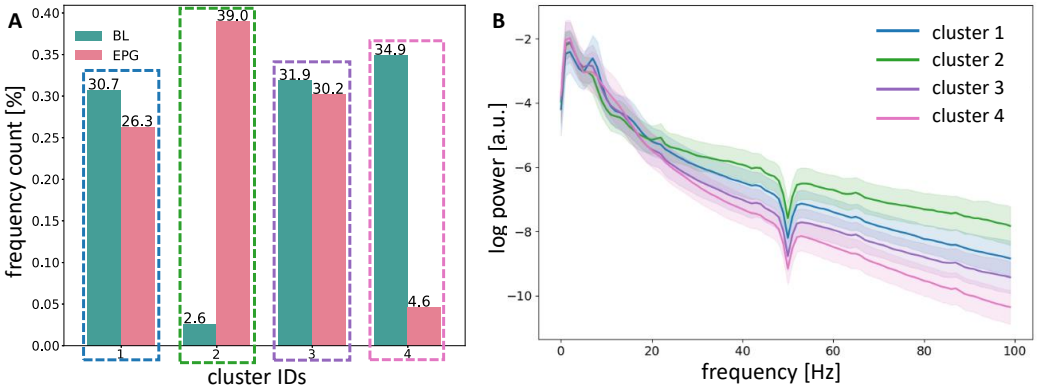


Figure 6: Clustering of high EPG score examples. **A**: percentage count of each class in each cluster. **B**: mean spectra of each cluster. The shaded area represents the standard deviation.

that these networks learn to discover changes in the EEG patterns across time that are triggered by the surgical procedure but are independent of the PPS and the ensuing EPG. However, we found that these networks could not discriminate *early* and *late* EEG patterns from the control group (mean AUC = 0.53, std. dev. = 0.12) and over 82% of all test samples from both *early* and *late* phases are classified as BL. This is additional evidence that the networks have learned to detect changes in EEG patterns that are induced by the PPS.

To visualize how exactly the prediction aggregation improves the discriminative ability of the model, we compute the distribution of scores assigned by the network to all test segments. Notably, the **score** is defined as the softmax output of the segment being EPG. Ideally, scores for BL segments should be close to zero, and EPG segments should have close-to-one scores. For simplicity, we only show the distributions of one representative LOO test trial from each group, as presented in Fig 5. The difference of the distributions within the same aggregation length is evaluated with the ANOVA test and the Wilcoxon rank sum test. In Fig. 5A, the distributions are significantly different in both cases for this rat (the ANOVA test, p-value $\leq 10^{-25}$, the Wilcoxon rank sum test, p-value $\leq 10^{-17}$). Results for other PPS rats are similar (not shown). To measure the sizes of differences between two distributions within the same aggregation length, we also computed Cohen's $d$ effect size (Rice and Harris, 2005). In the two examples shown, $d = 0.94$ and 2.91, respectively. Average $d$ values for the whole PPS-stimulated group with and without aggregation are 0.85 and 1.24, respectively. Cohen suggested that an effect size absolute value over 0.8 is considered large. Notably, there is still a considerable overlap between BL and EPG segments, i.e., in the BL period there are a certain number of segments classified as EPG and vice versa. When we aggregate over one hour, the effect of the distribution shift is magnified. In contrast, in the control group, the distributions of scores from one representative test trial with and without aggregation, are shown in Fig 5B, are not significantly different (the ANOVA test, p-values $\geq 0.5$, the Wilcoxon rank sum test, p-values $\geq 0.4$) with an effect size $d = 0.004$ and 0.012, respectively. The other two LOO test trials in the control group exhibit the same pattern.

#### 4.2.3. K-Means Clustering Analysis

In order to obtain a better understanding of the characteristics of the learned features, we conducted k-means clustering analysis on very certain samples collected from LOO test-trials. Here, a certain sample is defined as one whose softmax probability is larger than a threshold (set to 0.999). The k-means clustering analysis is based on the Euclidean distance between two samples power spectra. Specifically, we cluster the log-power spectrum of examples into four clusters, where the number of four is determined by the elbow-theory (Kodinariya and Makwana, 2013). From the frequency count plot, see Fig. 6A, we can see that the majority of the cluster No. 2 stems from the EPG class and that of the cluster No. 4 is from the BL class. From the mean spectra of each cluster, we can see that the EPG-dominant cluster has higher power in the frequency range over 20 Hz to 100 Hz. Specially, in this cluster, there is strong power around 22 Hz and its harmonics. On the other hand, the mean power spectrum of the BL-dominant cluster, cluster No. 4, has a faster decay towards higher frequencies.

## 5. Discussion

In recent years, ML could capitalize on the availability of big medical data sets. However, acquiring expert annotations for such data is impractical in many applications, representing a challenge for ML approaches. Here, we have tried to answer the question if an emerging epilepsy might be detectable from EEG signals even before the first seizure occurs. For this, we have used a rodent model of epilepsy (Costard et al., 2019), where epileptogenesis (EPG) is triggered through PPS. While massive amounts of training data are available from the BL (pre-stimulation) and the EPG (post-stimulation) periods, these data are only labeled by their time of recording. On the one hand, there might be large amounts of EPG-like signals present in the BL phase because there is brain injury involved in implanting the electrode. On the other hand, normal brain activities are still present in the EPG phase. Thus, we can expect short segments of EEG recordings to be often indistinguishable. A reliable classification requires pooling data over longer time windows. To achieve this, we have proposed a DNN approach with a prediction aggregation method. Our method is trained in an end-to-end fashion on five second segments and we have observed massive performance gains when aggregating predictions over one hour (improvements of 21%, 19%, and 13% in SEN, SPE, and AUC, respectively). Therefore, we have demonstrated a viable method for automatically predicting epilepsy from EEG recordings prior to the first epileptic seizure. This opens the door for early interventions to modify or even arrest the progression of the disease (Löscher, 2019). Furthermore, EEG patterns that the network has identified as being predictive of EPG may point towards new biomarkers of the disease. As a plausible alternative approach to our network architecture, a recurrent neural network (RNN) could be considered. However, our preliminary investigations have shown that RNN training requires more structure exploration and hyper-parameter search and our results leave little room for improvement on the data set presented here.

**Limitations** From the perspective of practical utility, a good biomarker for identifying EPG in a clinical setting should be noninvasive. In contrast, the data in our study were recorded with a depth electrode, which has a much higher signal-to-noise-ratio compared

to surface EEG recordings. For training a similar model to predict EPG in humans, the collection of surface EEG data from human patients would be necessary. As an immediate next step, we plan to extend our results to a group of human patients, who will undergo EEG (surface or intracranial) recording in the hospital after suffering a brain injury but before epilepsy is manifest. With sufficient training data from these and non-epileptic patients, we could envision a machine-learning-assisted diagnostic tool for the early detection of a developing epilepsy in human patients.

## References

Pedro Andrade, Jari Nissinen, and Asla Pitkänen. Generalized seizures after experimental traumatic brain injury occur at the transition from slow-wave to rapid eye movement sleep. *Journal of neurotrauma*, 34(7):1482–1487, 2017.

Mustafa Talha Avcu, Zhuo Zhang, and Derrick Wei Shih Chan. Seizure Detection Using Least Eeg Channels by Deep Convolutional Neural Network. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1120–1124. IEEE, 2019.

Carla Bentes, Hugo Martins, Ana Rita Peralta, Carlos Morgado, Carlos Casimiro, Ana Catarina Franco, Ana Catarina Fonseca, Ruth Geraldes, Patrícia Canhão, Teresa Pinho e Melo, et al. Early EEG predicts poststroke epilepsy. *Epilepsia open*, 3(2):203–212, 2018.

Xiaojun Bi and Haibo Wang. Early Alzheimers disease diagnosis based on EEG spectral images using deep learning. *Neural Networks*, 114:119–135, 2019.

Siddharth Biswal, Cao Xiao, M Brandon Westover, and Jimeng Sun. EEGtoText: Learning to Write Medical Reports from EEG Recordings. In *Machine Learning for Healthcare Conference*, pages 513–531, 2019.

Anatol Bragin, Charles L Wilson, Joyel Almajano, Istvan Mody, and Jerome Engel Jr. High-frequency oscillations after status epilepticus: epileptogenesis and seizure genesis. *Epilepsia*, 45(9):1017–1023, 2004.

Anatol Bragin, Lin Li, Joyel Almajano, Catalina Alvarado-Rojas, Aylin Y Reid, Richard J Staba, and Jerome Engel Jr. Pathologic electrographic changes after experimental traumatic brain injury. *Epilepsia*, 57(5):735–745, 2016.

Sergey Burnos, Peter Hilfiker, Oguzkan Sürücü, Felix Scholkmann, Niklaus Krayenbühl, Thomas Grunwald, and Johannes Sarnthein. Human intracranial high frequency oscillations (HFOs) detected by automatic time-frequency analysis. *PloS one*, 9(4), 2014.

Lara S Costard, Valentin Neubert, Morten T Venø, Junyi Su, Jørgen Kjems, Niamh MC Connolly, Jochen HM Prehn, Gerhard Schratt, David C Henshall, Felix Rosenow, et al. Electrical stimulation of the ventral hippocampal commissure delays experimental epilepsy and is associated with altered microrna expression. *Brain Stimulation*, 12(6): 1390–1401, 2019.

Amr Farahat, Christoph Reichert, Catherine M Sweeney-Reed, and Hermann Hinrichs. Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization. *Journal of neural engineering*, 16(6):066010, 2019.

Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8): 1836–1842, 2018.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016b.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Isabell Kiral-Kornek, Subhrajit Roy, Ewan Nurse, Benjamin Mashford, Philippa Karoly, Thomas Carroll, Daniel Payne, Susmita Saha, Steven Baldassano, Terence O'Brien, et al. Epileptic seizure prediction using big data and deep learning: toward a mobile system. *EBioMedicine*, 27:103–111, 2018.

Trupti M Kodinariya and Prashant R Makwana. Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6):90–95, 2013.

Patrick Kwan and Martin J Brodie. Early identification of refractory epilepsy. *New England Journal of Medicine*, 342(5):314–319, 2000.

Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6742–6751, 2018.

Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 82–92, 2019.

Wolfgang Löscher. The holy grail of epilepsy prevention: Preclinical approaches to antiepileptogenic treatments. *Neuropharmacology*, 167:107605, 2019.

Diyuan Lu, Sebastian Bauer, Valentin Neubert, Lara Sophie Costard, Felix Rosenow, and Jochen Triesch. A Deep Residual Neural Network Based Framework for Epileptogenesis Detection in a Rodent Model with Single-Channel EEG Recordings. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE, 2019.

Dan Z Milikovsky, Itai Weissberg, Lyn Kamintsky, Kristina Lippmann, Osnat Schefenbauer, Federica Frigerio, Massimo Rizzi, Liron Sheintuch, Daniel Zelig, Jonathan Ofer, et al. Electrocorticographic dynamics as a novel biomarker in five models of epileptogenesis. *Journal of Neuroscience*, 37(17):4450–4461, 2017.

Taylor Mordan, Nicolas Thome, Gilles Henaff, and Matthieu Cord. Revisiting multi-task learning with ROCK: a deep residual auxiliary block for visual detection. In *Advances in Neural Information Processing Systems*, pages 1310–1322, 2018.

Solomon L Moshé, Emilio Perucca, Philippe Ryvlin, and Torbjörn Tomson. Epilepsy: new advances. *The Lancet*, 385(9971):884–898, 2015.

Asla Pitkänen and Jerome Engel. Past and present definitions of epileptogenesis and its biomarkers. *Neurotherapeutics*, 11(2):231–241, 2014.

Asla Pitkänen, Wolfgang Löscher, Annamaria Vezzani, Albert J Becker, Michele Simonato, Katarzyna Lukasiuk, Olli Gröhn, Jens P Bankstahl, Alon Friedman, Eleonora Aronica, et al. Advances in the development of biomarkers for epilepsy. *The Lancet Neurology*, 15 (8):843–856, 2016.

Marnie E Rice and Grant T Harris. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and human behavior*, 29(5):615–620, 2005.

Massimo Rizzi, Claudia Brandt, Itai Weissberg, Dan Z Milikovsky, Alberto Pauletti, Gaetano Terrone, Alessia Salamone, Federica Frigerio, Wolfgang Löscher, Alon Friedman, et al. Changes of dimension of EEG/ECoG nonlinear dynamics predict epileptogenesis and therapy outcomes. *Neurobiology of disease*, 124:373–378, 2019.

Md Mostafa Kamal Sarker, Hatem A Rashwan, Farhan Akram, Syeda Furruka Banu, Adel Saleh, Vivek Kumar Singh, Forhad UH Chowdhury, Saddam Abdulwahab, Santiago Romani, Petia Radeva, et al. SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 21–29. Springer, 2018.

Padhraic Smyth and David Wolpert. Linearly combining density estimators via stacking. *Machine Learning*, 36(1-2):59–83, 1999.

Pierre Thodoroff, Joelle Pineau, and Andrew Lim. Learning robust features using deep learning for automatic seizure detection. In *Machine learning for healthcare conference*, pages 178–190, 2016.

Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7289–7298, 2019.

Mengni Zhou, Cheng Tian, Rui Cao, Bin Wang, Yan Niu, Ting Hu, Hao Guo, and Jie Xiang. Epileptic seizure detection based on EEG signals and CNN. *Frontiers in neuroinformatics*, 12(95), 2018.