Staging Epileptogenesis with Deep Neural Networks

Diyuan Lu
elu@fias.uni-frankfurt.de
Frankfurt Institute for Advanced
Studies (FIAS)
Frankfurt am Main, Germany

Sebastian Bauer
Sebastian.Bauer@kgu.de
Epilepsy Center Frankfurt
Rhine-Main Neurocenter
Frankfurt am Main, Germany

Valentin Neubert
valentin.neubert@uni-rostock.de
Oscar-Langendorff-Institute for
Physiology
Rostock, Germany

Lara Sophie Costard
laracostard@rcsi.com
Royal College of Surgeons Ireland
Dublin, Ireland

Felix Rosenow
rosenow@med.uni-frankfurt.de
Epilepsy Center Frankfurt
Rhine-Main Neurocenter
Frankfurt am Main, Germany

Jochen Triesch
triesch@fias.uni-frankfurt.de
Frankfurt Institute for Advanced
Studies (FIAS)
Frankfurt am Main, Germany

ABSTRACT

Epilepsy is a common neurological disorder characterized by recurrent seizures accompanied by excessive synchronous brain activity. The process of structural and functional brain alterations leading to increased seizure susceptibility and eventually spontaneous seizures is called epileptogenesis (EPG) and can span months or even years. Detecting and monitoring the progression of EPG could allow for targeted early interventions that could slow down disease progression or even halt its development. Here, we propose an approach for staging EPG using deep neural networks and identify potential electroencephalography (EEG) biomarkers to distinguish different phases of EPG. Specifically, continuous intracranial EEG recordings were collected from a rodent model where epilepsy is induced by electrical perforant pathway stimulation (PPS). A deep neural network (DNN) is trained to distinguish EEG signals from before stimulation (baseline), shortly after the PPS and long after the PPS but before the first spontaneous seizure (FSS). Experimental results show that our proposed method can classify EEG signals from the three phases with an average area under the curve (AUC) of 0.93, 0.89, and 0.86. To the best of our knowledge, this represents the first successful attempt to stage EPG prior to the FSS using DNNs.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; Cross-validation; **Neural networks**; • **Applied computing** → **Bioinformatics**; • **Networks** → Network performance analysis.

KEYWORDS

epileptogenesis, deep neural network, machine learning, EEG, class activation map, feature visualization

1 INTRODUCTION

Epilepsy is one of the most common and disruptive neurological disorders affecting about 1% of the world's population. It is characterized by recurrent unprovoked seizures and is accompanied by various co-morbidities such as migraine, depression, dementia, etc. [12]. Over 30% of the patients will eventually develop refractory epilepsy, defined as inadequate control of seizures by any medication [15]. In acquired epilepsy, an initial precipitating injury (IPI) such as stroke, traumatic brain injury or encephalitis leads to structural and functional remodelling of neuronal networks resulting in the occurrence of spontaneous seizures after a clinically silent latent period [20]. This remodelling process is termed epileptogenesis (EPG). Traditionally, epilepsy is diagnosed and treated after at least one unprovoked seizure, which indicates that the EPG has already progressed to a relatively advanced stage. This latent period can last months or even years. Treating high-risk patients at the early stage of EPG, or even customizing the treatment based on the severity of EPG could result in more effective disease-altering or even disease-arresting outcomes.

Pathomechanisms of EPG are not fully understood and its detection remains a major challenge. Studying early EPG in human patients is extremely difficult, simply because the epilepsy is typically only detected after the FSS. Therefore, work on early EPG is typically restricted to animal models [2]. Furthermore, early EPG can comprise a complex cascade of changes to the brain after the initial brain insult and this cascade may strongly depend on the type of brain insult. Changes can include, e.g., inflammatory reactions or blood-brain-barrier damage [9]. Some of these brain changes may be reflected in the EEG in the form of interictal epileptiform discharges (IEDs, including sharp-waves, spikes, spike-and-waves complex.), high-frequency oscillations, slowing or alteration of sleep spindles. Correspondingly, there have been attempts to identify suitable EEG biomarkers for EPG using a wide range of approaches [1, 3, 6, 17–19, 21]. However, a reliable staging
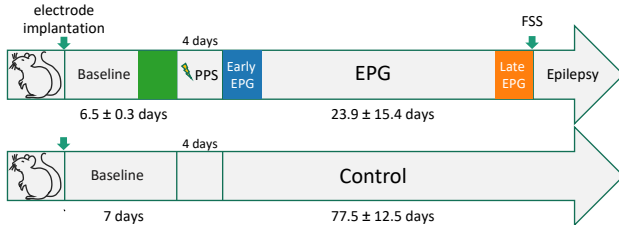
Figure 1: Timeline of the experiment. Shaded boxes indicate the different time periods where training and testing data are extracted. Upper row: PPS group. Lower row: control group (identical but without PPS). FSS: first spontaneous seizure. PPS: perforant pathway stimulation.

of EPG based on EEG measurements has not been demonstrated yet to the best of our knowledge.

Here, we use a rat epilepsy model, where EPG is induced by electrical perforant pathway stimulation (PPS) [8]. In previous work, we have shown that a DNN can be trained to distinguish EEG signals from baseline and EPG, i.e., before and after the PPS, with high specificity and sensitivity. Furthermore, we have demonstrated generalization to unseen rats [18]. Here we extend these results and present the first attempt to stage EPG using DNNs. In particular, we ask whether a DNN can also learn to distinguish early and late phases of EPG after the PPS but prior to the FSS, thereby allowing to estimate how "close" an individual may be to their FSS. The timeline of the experiment is shown in Fig 1. There are two groups of rats involved: a PPS group and a control group. The PPS group undergoes PPS and develops epilepsy before the end of the recording. The control group is not stimulated and they do not develop epilepsy before the end of the recording. Data from the control rats are used as a comparison to the PPS group. We demonstrate that our approach based on DNNs can successfully stage the EPG process and distinguish early from late EPG and that it generalizes to previously unseen rats.

2 RELATED WORK

2.1 Deep Learning for EEG analysis

Deep Learning (DL) techniques are commonly used in the analysis of EEG data in medical research. Example applications include the detection of Alzheimer's disease [4], autism [5], or Parkinson's disease [10]. In the context of epilepsy, DL has been applied for abnormal brain activity detection [22, 29] as well as seizure detection and prediction [7, 14, 28, 30, 33]. Roy *et al.* proposed a hybrid CNN and gated recurrent units (GRU) in classifying normal and abnormal brain activity, which takes time series EEG data as input and outputs the probability of being normal and abnormal, which is one of the first steps to understand the state of the brain activity in order to improve the accuracy of the diagnosis and the quality of patient care [22]. Tjepkema *et al.* explored different combinations of CNNs and recurrent neural networks (RNNs) as classifier to identify IEDs from scalp EEG [29]. Zhou *et al.* proposed a CNN-based approach to classify EEG time series data from different states, i.e., ictal, preictal, and interictal for the purpose of seizure detection [33]. They also compared the performance with time series and

frequency-domain as input and found that frequency-domain input exhibits better potential for this task. Kiral-Kornek *et al.* proposed a DL-based approach for patient-specific seizure prediction by classifying intracranial EEG data in pre-ictal and interictal phases [14]. Thodoroff *et al.* proposed a neural network combining convolutional layers (conv-layers) with recurrent layers to detect seizure onset. Their network takes the image-based representation of EEG signals as input capturing spatial, spectral, and temporal features of patient-specific seizures [28]. Cho *et al.* compared the performance of different input modalities of EEG data with different DNN-based network architectures for seizure detection [7]. They concluded that the CNN with time-series EEG data, and the RNN with periodogram data resulted in the best performance. While these works have demonstrated the utility of DL for EEG analysis in the context of epilepsy, they have not addressed the challenging detection and staging of EPG prior to the FSS that we demonstrate here for the first time.

2.2 Interpretable DNNs

The interpretation of the reasoning of a neural network is crucial in medical applications, as it allows verification by human users and provides insights rather than just succumbing to a *black box*. Many studies have been done to address the interpretability of DNNs [13, 23, 25, 27, 31, 32]. Yosinski *et al.* developed a software tool for visualizing live feature extraction in the neural network by viewing the activation maps of different channels in different layers as well as by regularized optimization to generalize inputs that maximize the channel activation [31]. Simonyan *et al.* proposed to generate an input image that maximizes the output softmax probability of a given class. Meanwhile, a saliency map can be computed, which is the ranking of each pixel based on their contribution to the given class of a given sample [25]. Bach *et al.* proposed the Layerwise Relevance Propagation (LRP), which understands the learning of the network by decomposing the output in terms of the input dimensions in a fashion that relates to Taylor decomposition [23]. Sturm *et al.* applied the LRP technique to visualize the frequency contribution to the classification result with EEG data [27]. Zhou *et al.* proposed the concept of class activation map (CAM), which can identify important regions in the inputs by propagating back the weights of the dense softmax layer to the inputs [32]. CAM is easy to deploy and provides more focused and localized discrimination. In this work, we also leverage CAM with 1-*d* EEG data to better visualize the network properties and the learned features.

2.3 EEG-based Biomarkers of Epileptogenesis

Over the last decades several studies have attempted to find EPG biomarkers in EEG signals. Li *et al.* and Bragin *et al.* focused on high-frequency oscillations (HFOs) in a rat epilepsy model with kainic acid (KA) injection [6, 17]. They found that the sooner HFOs appear after the injection, the higher the rate of spontaneous seizures in the chronic phase, and the shorter the latent period is, the more spontaneous seizures will occur. Milikovsky *et al.* focused on theta band activity and showed that a decreased theta power can be a robust feature in identifying EPG in five animal epilepsy models [19]. Andrade *et al.* investigated the role of sleep-wake disturbance

in EPG and found that there is a decrease of the dominant frequency and the duration of sleep spindles in a traumatic brain injury epilepsy model with generalized seizures [1]. Bentes *et al.* found that in stroke patients, the asymmetry in the background activity with the occurrence of IEDs are independent indicators of post-stroke epilepsy in the first year after stroke [3]. Sheybani *et al.* found that in a mouse model of epilepsy with kainate injection, the spatial propagation of a subgroup of spikes across the brain can be a reliable indicator of EPG as well as epilepsy in the chronic phase [24]. Lu *et al.* trained a DNN with the Fourier transformation of the time-series EEG data from a rat epilepsy model and showed that a decrease of power in theta band and an increase of power in frequencies over 100 Hz can be reliable indicators of EPG [18]. Rizzi *et al.* investigated the nonlinear dynamics of EEG signals and found a significant decrease of the so-called embedding dimension in early EPG that correlates with the severity of the ongoing EPG [21]. Here, we use an unbiased deep learning approach to study the EPG process to subdivide it into different stages and identify potential biomarkers to distinguish early and late phases of EPG.

3 METHODS

3.1 Animal Model

We use a mesial temporal lobe epilepsy with hippocampal sclerosis (mTLE-HS) rodent model, where epilepsy is electrically induced through PPS. Details have been described in [8]. Continuous single-channel EEG recordings from a depth electrode implanted in the dentate gyrus are collected from each rat from the beginning of the implantation until the FSS, which indicates the manifestation of epilepsy. The 24/7 recordings enable us to continually monitor the entire EPG prior to the FSS. There are two groups of rats involved in this study, 1) seven rats had PPS and developed epilepsy before the end of recording, which we denote as PPS rats, 2) three rats did not get PPS stimulation and did not develop epilepsy by the end of recording, which we denote as control rats. In the PPS group, the average EPG phase is 4 weeks (range 1 – 7 weeks). The EPG phase is terminated by the FSS. The timelines for both group are shown in Fig. 1. Training data are taken from the three highlighted periods from PPS rats for the three-class classification task. We define the three classes to be the Baseline class (*BL*) – green, the *early EPG* class – blue, and the *late EPG* class – orange. The data from the control rats are used only for testing the model trained on the PPS group. The total available number of recordings from each rat is summarized in Table 1 and Table 2.

3.2 EEG Data Preprocessing

The data acquisition was achieved through wireless EEG transmitters with a sampling rate of 512 Hz and a band-pass filter between 0.5 - 160 Hz as well as a notch filter at 50 Hz. Occasionally, EEG artifacts can appear as extreme amplitude values and signal loss due to electronic interference and weak transmission. To combat this problem, we first applied a MATLAB function, i.e., `filloutliers` [1] with the parameters `method = 'pchip'`; `movmethod = 'movmedian'`; `window = 50` to filter out unrealistic extreme values. Then, the continuous recordings are divided into

---

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

Table 1: Summary of the data collections from PPS rats in hours (hrs).

rat ID	PPS 1	PPS 2	PPS 3	PPS 4	PPS 5	PPS 6	PPS 7
BL (hrs)	162	160	149	82	163	164	157
EPG (hrs)	700	508	400	140	1568	173	648

Table 2: Summary of the data collections from control rats in hours (hrs).

rat ID	Ctr 1	Ctr 2	Ctr 3
in total (hrs)	1536	2140	2248

five-second long non-overlapping segments. To manage data loss, we discarded any five-second segments with more than 20 % data loss. As a result, we discarded around 5% of the total recordings. The remaining segments were eligible for the DNN training.

3.3 DNN Architecture

We use a deep residual neural network with 16 blocks with residual connections (res-block), as shown in Fig. 2, inspired by [11]. The model takes five-second long EEG segments as input and outputs the probability over three classes, i.e., BL, early EPG, and late EPG. We keep the design of each res-block as in [11], where each res-block consists of two conv-layers, batch-normalization, dropout, and ReLU non-linear activation. The number of channels in the first conv-layer and the first block is 16, and it increases by a factor of 2 in every four blocks. There are two branches in each block: one goes through convolution, batch-normalization, ReLU activation and dropout; the other, called skip connection, simply goes through max-pooling. They are combined in an additive manner at the end of the block before passing through the batch-normalization and ReLU activation. To reduce the dimensionality of the feature maps, we use a stride of two in the second conv-layer and the max-pooling layer in every other block starting from the second block. The output of the last conv-layer is fed to the global average pooling (GAP) operation, which is followed by a dense layer with three output units with softmax non-linear activation. The dropout rate is 0.2 everywhere in the graph.

3.4 Class Activation Map

Proposed by Zhou *et al.*, the class activation map is a method to visualize the "importance" of different regions of the input for the classification decision. It takes advantage of the global average pooling (GAP) after the last conv-layer, and assigns different weights to each squashed feature map. To be specific, the k-th feature map from the last conv-layer, denoted as f_k, which has shape [*height*, *width*]. The GAP layer takes the mean activation of each f_k, and the resulting k-th feature map F_k is $\frac{1}{N} \sum_{i,j} f_k(i, j)$, where N is the total number of elements of f_k. It reduces the dimension by the factor of *height* $\times$ *width*. Then, for a given class c, the input to the softmax layer, S_c, is a weighted linear combination of all the feature
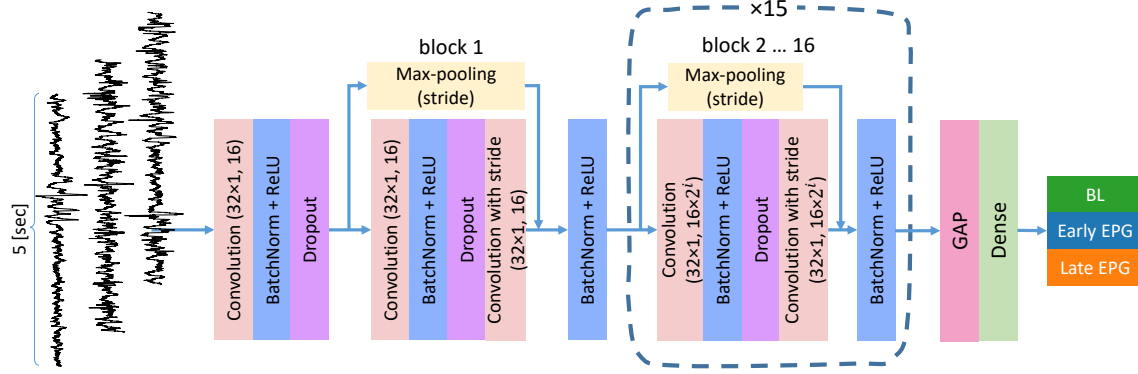
Figure 2: The DNN structure used in this study. The network takes a mini-batch of five-second segments as input and outputs the probability over the three classes. GAP: global average pooling. BL: Baseline

maps, which is computed by

$$S_c = \sum_k w_k^c \frac{1}{N} \sum_{i,j} f_k(i,j) \propto \sum_{i,j} \sum_k w_k^c f_k(i,j), \quad (1)$$

where w_k^c denotes the *importance* of f_k for class c. Finally, the softmax probability for class c can be computed as $\frac{\exp S_c}{\sum_c \exp S_c}$. Then, when the training is finished, the class activation map for class c at position (i,j), $\mathrm{CAM}_c(i,j)$, is given by

$$\mathrm{CAM}_c(i,j) = \sum_k w_k^c f_k(i,j). \quad (2)$$

Hence, $S_c = \sum_{i,j} w_k^c \mathrm{CAM}_c(i,j)$, and the weights w_c are fixed after the training. Then, $\mathrm{CAM}_c(i,j)$ indicates the importance of the activation at the position (i,j) contributing to the class c.

3.5 DNN Training and Evaluation

We apply a seven-fold leave-one-out cross-validation (LOO-CV) scheme, where the network is trained with the data from six out of seven rats in the PPS group. Specifically, in each fold, we withhold the data from one rat as the test set, and the data from other six rats form the training and the validation sets with a train-validation-split of 8:2. This procedure is repeated seven times, and each time we hold out a different rat for testing. This is highly relevant to test the generalization ability of the classifier to unseen data from unseen subjects. We randomly select 25 hours from a three-day window from each phase for training and validation, shown as the shaded periods in Fig. 1. The choice of 25 hours is a reasonable trade-off between computational cost and performance from empirical experience, since we also experiment using all data from the three day periods and it increases the total training time by a factor of three and no significant improvement regarding the classification performance is found. Our DNN model is implemented in Tensorflow and trained with an NVIDIA GeForce RTX 2080 Ti GPU and one epoch of training takes 35 minutes on average. After the network is trained, we test it with all the data from those three-day periods (shown in Fig. 1) of the previously withheld rat. We report results as the average across all seven LOO test trials.

To evaluate the performance, we compute the receiver operating characteristic (ROC) curve in the multi-class scenario, where the

ROC curve is computed for each class in a one-vs-all manner. The area under the ROC curve is a scalar value indicating the goodness of the trained classifier. Several other performance metrics including precision, recall, and F1-score are also computed. These metrics are given by:

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\text{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}},$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative numbers, respectively. We also compare our results with several baseline network structures: a feed-forward neural network (FNN), a deep convolutional neural network (DCNN) [26], EEGNet [16], and one variant of our proposed model with only four blocks, which we denote as Proposed-4block.

The FNN used in this work is a straight forward multi-layer perceptron with three dense layers equipped with 1024, 256, and 128 units per layer. Each dense layer is regularized with L_2 penalty with a factor of 0.01 and followed by a batch-normalization layer and a dropout (rate=0.5) layer. The nonlinear activation is ReLU in this model.

Sors *et al.* proposed the DCNN for sleep staging with single-channel EEG. Compared to the original architecture, we made several changes to adapt to the training data format we have in our experiment. First, due to our input being shorter (five-second segments under 512 Hz sampling rate, which yields 2560 data points per sample) than theirs (15 000 data points), we reduced the number of conv-layer from twelve to nine: five (instead of six) conv-layers with 128 output channels and four (instead of six) conv-layers with 256 output channels. Each conv-layer has stride 2 to sub-sample the feature map. The architecture is conv (7×1, 128, stride 2) – conv (7×1, 128, stride 2) – conv (7×1, 128, stride 2) – conv (7×1, 128, stride 2) – conv (7×1, 128, stride 2) – conv (5×1, 256, stride 2) – conv (5×1, 256, stride 2) – conv (5×1, 256, stride 2) – conv (3×1, 256,

stride 2) – flatten – fully-connected (units=100) – fully-connected (units=3). We kept other training parameters identical to the original paper.

Lawhern *et al.* proposed the original EEGNet for EEG classification in multiple brain-computer interfaces. The EEG snippets used in their evaluation are multi-channel event related potential (ERPs) recorded from surface EEG setups, band-pass filtered between 1-40 Hz, downsampled to 128 Hz, and focused on 1 to 2 seconds around the event onset. The original EEGNet demonstrates good generalization to EEG classification among different experiment diagrams even though the total number of parameters is two orders of magnitude smaller than the baseline methods evaluated in their work. To adapt EEGNet to our task, we made several changes to the architecture while keeping layers such as batch-normalization, dropout, exponential linear unit (ELU) activation function, and average pooling unchanged: 1) We expanded the width of the convolutional filter from 64 to 256, which is half of our sampling rate as suggested in the original paper. 2) We used three instead of two layers of convolution while omitting the depth-wise convolution, since our data is single-channel. Unfortunately, the classification accuracy of this modified EEGNet (henceforth denoted EEGNet1) does not exceed chance-level. One contributing factor might be the low number of trainable parameters. In total, EEGNet1 only has 223 323 learnable parameters, which is considerably fewer than our proposed model. To make the total number of trainable comparable to ours, we increased the number of conv-layers and the number of filters in each layer. This is essentially equivalent to a relatively shallow CNN (7 conv-layers compared to 33 layers in our proposed model) with very wide convolutional filters, which we denote as EEGNet2. The resulting structure of EEGNet2 is conv $(256 \times 1, 16)$ – batch-normalization – conv $(256 \times 1, 16)$ – batch-normalization + ELU + average-pooling + dropout – conv $(256 \times 1, 32)$ – batch-normalization + ELU + average-pooling + dropout – conv $(256 \times 1, 32)$ – batch-normalization + ELU + average-pooling + dropout – conv $(256 \times 1, 64)$ – batch-normalization + ELU + average-pooling + dropout – conv $(256 \times 1, 64)$ – batch-normalization + ELU + average-pooling + dropout – flatten – fully-connected (units=3). As a result, the EEGNet2 has a total number of 4 195 107 parameters, which is comparable to that of our proposed model (4 200 048). However, the results show that with the same amount of training data and training time, both versions of EEGNets, i.e., EEGNet1 and EEGNet2 perform at chance-level. Thus, their performance measures were omitted in the performance report.

4 EXPERIMENTS AND RESULTS

Table 3 shows the performance summary of our proposed model in comparison to the baseline methods. The reported performance metrics are averaged for each class as well as a macro-average of all classes across all LOO test trials. Our proposed method obtains the best performance in almost all evaluated metrics compared to the baseline methods. Notably our proposed-4block model still obtains better performance than FNN and DCNN, even though the number of trainable parameters is more than 20 times smaller. Compared to the full-size proposed model, the Proposed-4block model suffers from a slight performance degradation. From the

class-wise performance, we can see that, in general, the BL class is easier for the networks to classify as shown by the highest average performance among the three classes in all models.

4.1 Prediction Aggregation and ROC Analysis

To gather statistics of the estimated class membership over a longer time period, we apply a prediction aggregation technique as proposed in our previous study [18]. Essentially, we apply a linear average aggregation of the resulting softmax probability across multiple consecutive five second data segments such that the probabilities of each class are accumulated across a longer period of time. Figure 3 shows the averaged AUCs of the three classes across all LOO test trials **with** and **without** the prediction aggregation (Fig. 3A and Fig. 3B) as well as the effect of the pooling length used in the prediction aggregation (Fig. 3C). In general, the network can distinguish BL segments better than the other two classes as shown by the highest average AUC under the ROC curve among the three classes, with or without the prediction aggregation. Prediction results for the control group are only marginally better than chance, suggesting that the network really detects changes in brain activity patterns due to the PPS, rather than any changes triggered by the initial electrode implantation that are independent of the PPS. Prediction aggregation over one hour increases the average AUC of the baseline, early, and late EPG classes by 0.1, 0.12, and 0.11, respectively.

To study the benefits of aggregation in more detail, we compute the AUCs for various aggregation lengths in each LOO test trial, i.e., 5 seconds, 30 seconds, one, two, five, ten, 20, 30, and 60 minutes. The average AUC as a function of the aggregation lengths is depicted in Fig 3C. It reflects the inter-rat variability in the three-class classification with our proposed network, i.e., the AUC starts at different levels of confidence without prediction aggregation (the first data points from all rats). The figure shows that with an increasing pooling length, the average AUC increases in all LOO test trials. To be specific, with one hour of aggregation, the average AUC improved by 0.12 (a maximum of 0.18 and a minimum of 0.06). Hence, aggregating the softmax output from the network across multiple consecutive segments captures trends across a longer period, which is essential for distinguishing different classes in our task. Aggregation over even longer time periods (>1 hour) might be able to further improve performance.

4.2 Disease Progression

EPG is a gradual process, but above we treated EPG detection and staging as a discrete classification problem by defining (somewhat arbitrarily) the first three days after the stimulation as the early EPG phase, and the last three days before the FSS as the late EPG phase. The data from the period in between these two phases has not been considered so far. In the following, we analyze samples from this intermediate period and study how the network, which has been trained to distinguish Baseline, early and late EPG phases, will classify them. Specifically, we consider the estimated probability for each class, denoted as the *class score*, throughout the whole recording period from a randomly picked pre-trained model from the LOO cross-validation scheme, which we call "Pretrained-1" model. Here, we are interested in the general trend rather than the

Table 3: Performance across all leave-one-out test trials with one hour of prediction aggregation. Evaluation metrics are reported in class-wise average and overall average for each model. Numbers are shown in *mean ± std*.

Model	Class	Precision	Recall	F1-score	Accuracy	# trainables
FNN	0	0.51 ± 0.08	0.67 ± 0.13	0.57 ± 0.06	0.44 ± 0.06	2 920 963
	1	0.49 ± 0.07	0.65 ± 0.07	0.55 ± 0.02	0.43 ± 0.04	
	2	0.43 ± 0.06	0.63 ± 0.16	0.49 ± 0.04	0.39 ± 0.03	
	average	0.47 ± 0.03	0.65 ± 0.04	0.53 ± 0.03	0.42 ± 0.03	
DCNN	0	0.47 ± 0.13	0.54 ± 0.25	0.46 ± 0.18	0.46 ± 0.13	1 607 187
[26]	1	0.43 ± 0.32	0.40 ± 0.30	0.41 ± 0.31	0.41 ± 0.08	
	2	0.35 ± 0.23	0.35 ± 0.28	0.33 ± 0.23	0.40 ± 0.07	
	average	0.42 ± 0.11	0.43 ± 0.22	0.40 ± 0.18	0.42 ± 0.07	
Proposed-4block	0	0.70 ± 0.14	0.88 ± 0.04	0.78 ± 0.10	0.66 ± 0.14	82 912
	1	0.43 ± 0.06	0.68 ± 0.18	0.53 ± 0.10	0.41 ± 0.05	
	2	0.51 ± 0.05	**0.82 ± 0.13**	0.62 ± 0.01	0.47 ± 0.02	
	average	0.55 ± 0.04	0.79 ± 0.09	0.64 ± 0.05	0.51 ± 0.05	
Proposed model	0	**0.85 ± 0.17**	**0.96 ± 0.02**	**0.90 ± 0.10**	**0.84 ± 0.17**	4 200 048
	1	**0.69 ± 0.12**	**0.81 ± 0.17**	**0.74 ± 0.14**	**0.64 ± 0.15**	
	2	**0.71 ± 0.33**	**0.74 ± 0.32**	**0.72 ± 0.31**	**0.71 ± 0.22**	
	average	**0.75 ± 0.15**	**0.84 ± 0.12**	**0.78 ± 0.14**	**0.73 ± 0.14**	
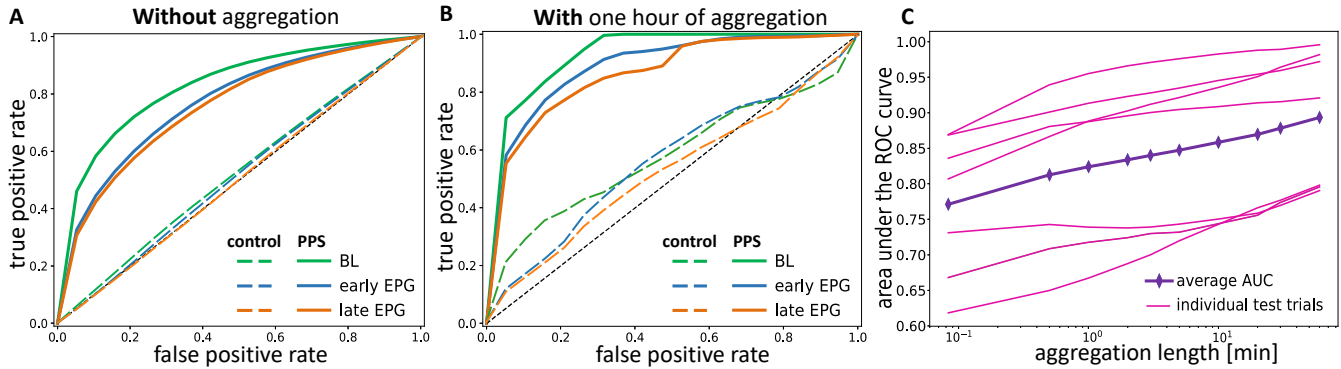
Figure 3: Network performance across all test trials within the PPS and the control group. A. Average ROC curves of multiple classes without aggregation within the PPS group and the control group. The AUC for the three classes of PPS rats are 0.83, 0.77 and 0.75 (solid lines) and those of the control rats are 0.52, 0.51, and 0.50 (dashed lines). B. Average ROC curves of multiple classes with aggregation over one continuous hour within the PPS and the control group. The AUC of the three classes for the PPS rats (solid lines) are 0.93, 0.89, and 0.86, and those of the control rats are 0.58, 0.56, and 0.53 (dashed lines). C. The AUC as a function of the aggregation length in all individual PPS LOO test trials (magenta lines) and the average AUC of all classes across all trials (purple with diamonds). ROC: receiver operating characteristic. AUC: area under the curve.

classification accuracy, so the training data were also included. The progression of class scores from two example **PPS** rats and one **control** rat are shown in Fig. 4. One of the PPS rats (PPS 1) has a relatively long EPG duration (30 days) and the other (PPS 4) has a short EPG duration (6 days). The control rat (Ctr 1) has 64 days of recordings in total.

Several findings are evident in the data for the PPS rats in Fig. 4A,B. First, the Baseline score is high during the entire baseline period and drops to small values during the EPG phase. Second, with the beginning of the EPG phase, the early EPG score increases and then gradually decreases towards the late EPG phase. Third,

conversely, the late EPG score is low during baseline and the beginning of EPG and then gradually increases towards the late EPG phase. Fourth, in some animals we observe a circadian rhythm in the early and late EPG scores during the transition period between early and late EPG (compare Fig. 4A). These findings are in sharp contrast to those for the control rats. In their case, the late EPG score remains low throughout the entire recording period, in line with these animals not developing epilepsy during the experiment (compare Fig. 4C).
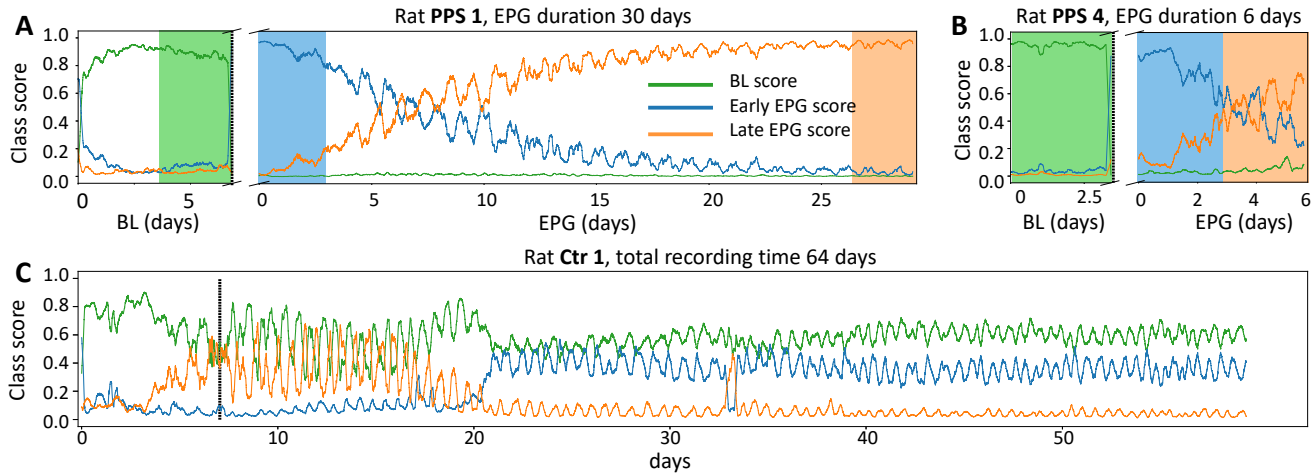
Figure 4: Class scores from two PPS rats (A,B) and one control rat (C) during the entire recording. The vertical black dashed lines indicate the time when the PPS rats started receiving PPS, while control rats did not.

4.3 Feature Representation

The interpretation of EEG signals is always challenging, since they are highly variable — especially across subjects. Analyzing and understanding the discriminative features learned by a DNN model can give valuable insights as to what distinguishes the classes. This can be particularly helpful in medical applications, where the differences between classes many not be easily spotted — even by the expert eye. Here, we present the feature representations learned by the network. Using the Pretrained-1 model, we passed all the data from all seven rats through the network and computed the average activation of the last conv-layer for each class, shown on the top left of Fig. 5. We can see that there is a group of feature channels that are very active. Most importantly some of these feature channels are most active for one class but not the others and some extract features that contribute to more than one class. Next, we identified several channels that were highly active for each class and plotted the EEG segments that maximally activate them. Interestingly, we found several feature channels responding to very distinctive features such as spikes in channel 3, spike-and-slow-waves in channel 9, spindles and HFOs in channel 15, theta rhythm in channel 16, delta wave plus low beta in channel 21, etc. From this we can conclude that before the softmax layer, the network has already extracted class-specific features that are clinically meaningful.

To further elucidate which parts of the input contribute most to the classification of the different EPG stages, we leverage the CAM visualization while manipulating the assigned labels for the EEG segments. Taking Pretrained-1 model, we freeze the weights and for a given sample, we assign in turn the three different labels. Then, by computing the CAM of the given sample under the assigned label, we trace back which parts of the given five second input segment most support (> 80-th percentile) the assigned classification. The results are shown in Fig. 6. Indeed, the CAMs for the sample vary depending on the given label. There are several interesting features that the network has discovered. First, the BL class is most

supported by low-amplitude waves, and many downwards deflections. Second, sharp waves contribute to both EPG classes, but the difference lies in the width of the wave forms. While an early EPG classification is supported by narrow spikes, or spike-like waves, a late EPG classification is supported by somewhat wider sharp waves.

5 CONCLUSION

We have proposed a DNN model for single-channel intracranial EEG classification to better understand the progression of epileptogenesis (EPG). Specifically, our aim was to stage the EPG process prior to the first spontaneous seizure (FSS), which could facilitate early intervention **before** an epilepsy becomes manifest. In previous work, we had already shown that a DNN can learn to distinguish EEG data from before and after the epilepsy-inducing stimulation with high discrimination and generalization ability [18]. Here, we have sought to answer a) whether we can further distinguish different stages of EPG before the FSS, and b) what EEG features would be representative for each stage. To this end, we have trained a DNN model with five-second EEG segments recorded from three phases in a rodent epilepsy model [8]: three days before the PPS (Baseline, BL), three days shortly after the PPS (early EPG), and three days immediately before the FSS (late EPG). We have evaluated our approach in a LOO scheme to test the generalization ability of the model to data from unseen rats. To pool evidence over larger time windows, we applied a prediction aggregation method as in previous work [18]. We also compared the performance of our model to four other models, specifically an FNN model, a DCNN model [26], the well-known EEGNet [16], and a reduced version of our model with 50 times fewer parameters. In an extensive performance evaluation, we showed that our proposed model yielded the best results and could distinguish different EPG stages with high accuracy. Furthermore, we showed that the network learns to extract meaningful EEG features to perform the classification.

Various challenges will need to be overcome, in order to translate our findings to human patients. First, the rodent model we have
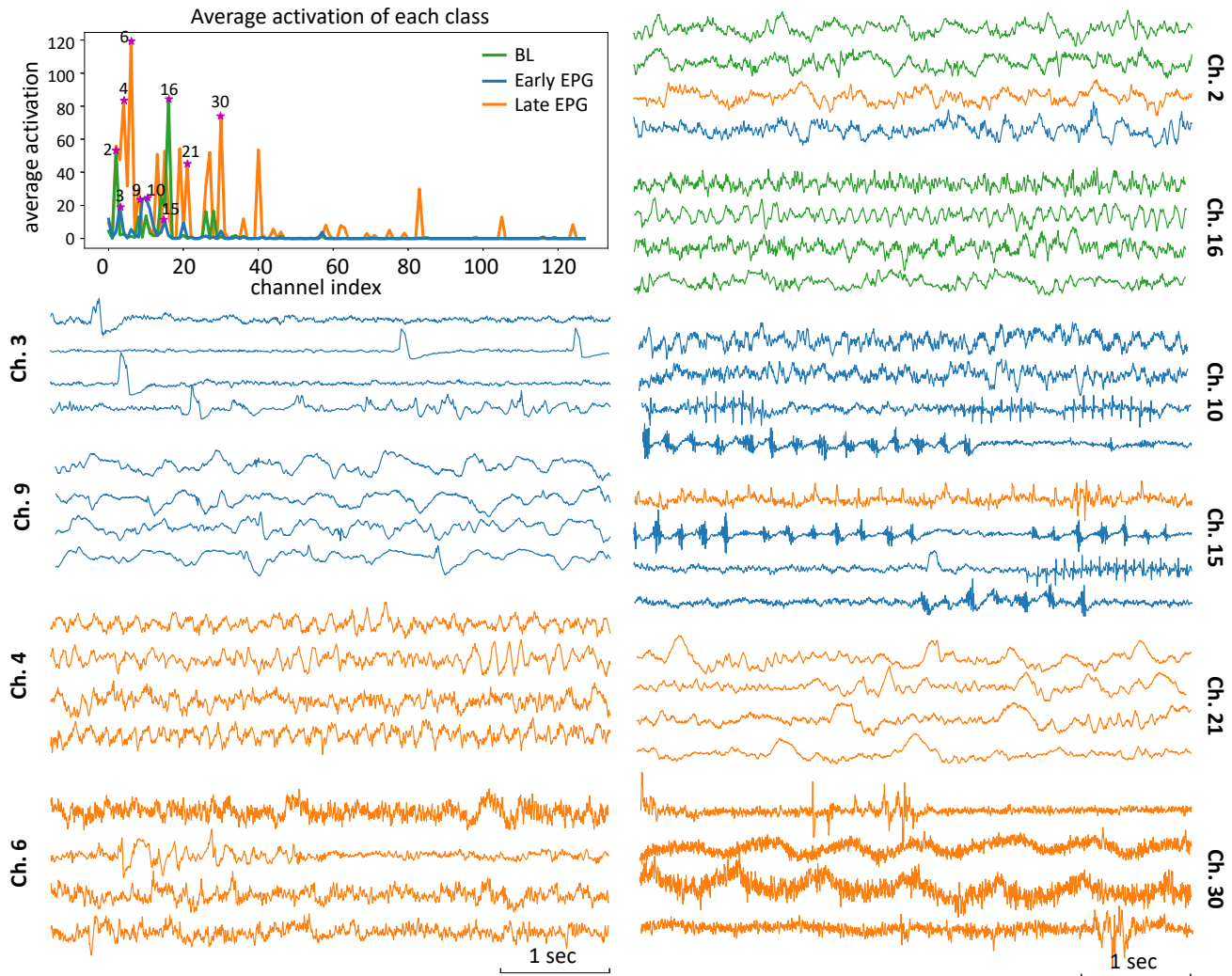
Figure 5: Normalized average activation of the last conv-layer by class (top left). Examples of five second EEG samples that maximize the activation of certain channels in the last conv-layer. Color indicates a sample's class label. Scale bar represents 1 second.

used provides quasi ideal conditions, supplying high quality, 24/7 intracranial recordings directly from the affected brain region. It is unclear whether similar results could be achieved with surface EEG recordings from a diverse set of human patients. The second challenge is that epilepsy is typically diagnosed only *after* the FSS. In order to attempt early detection of EPG as we have demonstrated here in human patients, one would have to obtain recordings from patients *before* the FSS. This requires monitoring a population of patients with a sufficiently high risk of developing epilepsy, which is challenging. Third, our approach relies on a large data set comprising around-the-clock recordings over several weeks for each individual. Acquiring similar data from a (homogeneous) patient population would be very difficult. It is an open question, how much data would be required to allow accurate classification and good generalization. Fortunately, in our experiments, pooling data over one hour already provided very good results. Such a time span

appears manageable in clinical practice. Finally, even if EPG could be detected and staged reliably in human patients at risk of developing epilepsy, it is far from clear which forms of early intervention would be effective in modifying or halting the disease development. In fact, such interventions will likely have to depend on the specific type of epilepsy and be adapted to individual patients. In the future, machine learning may also support physicians in this challenging task.
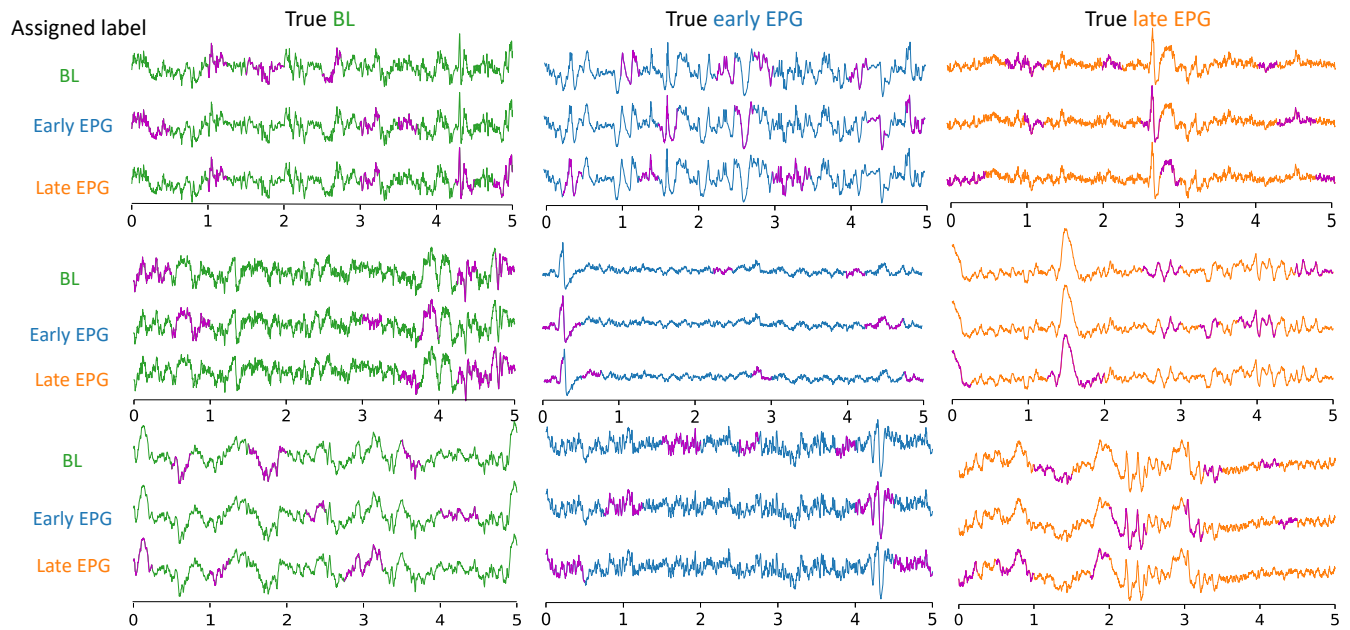
ACKNOWLEDGMENTS

Figure 6: Identifying informative regions in the five-second long input via CAM. The main color of a trace corresponds to its true label. The areas highlighted in magenta most strongly support the assigned classification (>80-th percentile).

REFERENCES

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

[2] AJ Becker. 2018. Animal models of acquired epilepsy: insights into mechanisms of human epileptogenesis. *Neuropathology and applied neurobiology* 44, 1 (2018), 112–129.

[3] 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. 2018. Early EEG predicts poststroke epilepsy. *Epilepsia open* 3, 2 (2018), 203–212.

[4] Xiaojun Bi and Haibo Wang. 2019. Early Alzheimer's disease diagnosis based on EEG spectral images using deep learning. *Neural Networks* 114 (2019), 119–135.

[5] William J Bosl, Helen Tager-Flusberg, and Charles A Nelson. 2018. EEG analytics for early detection of autism spectrum disorder: a data-driven approach. *Scientific reports* 8, 1 (2018), 1–20.

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

[7] Kyung-Ok Cho and Hyun-Jong Jang. 2020. Comparison of different input modalities and network structures for deep learning-based seizure detection. *Scientific Reports* 10, 1 (2020), 1–11.

[8] 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. 2019. Electrical stimulation of the ventral hippocampal commissure delays experimental epilepsy and is associated with altered microRNA expression. *Brain Stimulation* 12, 6 (2019), 1390–1401.

[9] Jerome Engel Jr and Asla Pitkänen. 2020. Biomarkers for epileptogenesis and its treatment. *Neuropharmacology* 167 (2020), 107735.

[10] Faraz Faghri, Sayed Hadi Hashemi, Hampton Leonard, Sonja W Scholz, Roy H Campbell, Mike A Nalls, and Andrew B Singleton. 2018. Predicting onset, progression, and clinical subtypes of Parkinson disease using machine learning. *bioRxiv* (2018), 338913.

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

[12] Mark R Keezer, Sanjay M Sisodiya, and Josemir W Sander. 2016. Comorbidities of epilepsy: current concepts and future perspectives. *The Lancet Neurology* 15, 1 (2016), 106–115.

[13] Pieter Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. 2018. Learning how to explain neural networks: PatternNet and PatternAttribution. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings* (2018), 1–12.

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

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

[16] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering* 15, 5 (2018), 056013.

[17] Lin Li, Mayur Patel, Joyel Almajano, Jerome Engel Jr, and Anatol Bragin. 2018. Extrahippocampal high-frequency oscillations during epileptogenesis. *Epilepsia* 59, 4 (2018), e51–e55.

[18] Diyuan Lu, Sebastian Bauer, Valentin Neubert, Lara Sophie Costard, Felix Rosenow, and Jochen Triesch. 2019. 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)*. IEEE, 1–6.

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

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

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

[22] Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. 2019. ChronoNet: a deep recurrent neural network for abnormal EEG identification. In *Conference on Artificial Intelligence in Medicine in Europe*. Springer, 47–56.

[23] Bach Sebastian, Binder Alexander, Montavon Grégoire, Klauschen Frederick, Müller Klaus-Robert, Samek Wojciech, and Suarez Oscar Deniz. [n.d.]. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *Plos One* 10, 7 ([n. d.]), e0130140.

[24] Laurent Sheybani, Gwenaël Birot, Alessandro Contestabile, Margitta Seeck, Jozsef Zoltan Kiss, Karl Schaller, Christoph M Michel, and Charles Quairiaux. 2018. Electrophysiological evidence for the development of a self-sustained large-scale epileptic network in the kainate mouse model of temporal lobe epilepsy. *Journal of Neuroscience* 38, 15 (2018), 3776–3791.

[25] K Simonyan, A Vedaldi, and A Zisserman. 2017. Deep inside convolutional networks: visualising image classification models and saliency maps. CoRR 2013; abs/1312.6034. *arXiv preprint arXiv:1312.6034* (2017).

[26] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. 2018. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control* 42 (2018), 107–114.

[27] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2016. Interpretable deep neural networks for single-trial EEG classification. *Journal of neuroscience methods* 274 (2016), 141–145.

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

[29] Marleen C Tjepkema-Cloostermans, Rafael CV de Carvalho, and Michel JAM van Putten. 2018. Deep learning for detection of focal epileptiform discharges from scalp EEG recordings. *Clinical neurophysiology* 129, 10 (2018), 2191–2196.

[30] KD Tzimourta, AT Tzallas, N Giannakeas, LG Astrakas, DG Tsalikakis, and MG Tsipouras. 2018. Epileptic seizures classification based on long-term EEG signal wavelet analysis. In *Precision medicine powered by pHealth and connected health.* Springer, 165–169.

[31] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding Neural Networks Through Deep Visualization. *Computer Science* (2015).

[32] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2921–2929.

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