# Diagnosing Epileptogenesis with Deep Anomaly Detection

**Amr Farahat** AMR.FARAHAT@ESI-FRANKFURT.DE
*Frankfurt Institute for Advanced Studies (FIAS)*
*Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society*
*Frankfurt am Main, Germany*

**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

We propose a general framework for diagnosing brain disorders from Electroencephalography (EEG) recordings, in which a generative model is trained with EEG data from normal healthy brain states to subsequently detect any systematic deviations from these signals. We apply this framework to the early diagnosis of latent epileptogenesis prior to the first spontaneous seizure. We formulate the early diagnosis problem as an unsupervised anomaly detection task. We first train an adversarial autoencoder to learn a low-dimensional representation of normal EEG data with an imposed prior distribution. We then define an anomaly score based on the number of one-second data samples within one hour of recording whose reconstruction error and the distance of their latent representation to the origin of the imposed prior distribution exceed a certain threshold. Our results show that in a rodent epilepsy model, the average reconstruction error increases as a function of time after the induced brain injury until the occurrence of the first spontaneous seizure. This hints at a protracted epileptogenic process that gradually changes the features of the EEG signals over the course of several weeks. Overall, we demonstrate that unsupervised learning methods can be used to automatically detect systematic drifts in brain activity patterns occurring over long time periods. The approach may be adapted to the early diagnosis of other neurological or psychiatric disorders, opening the door for timely interventions.

## 1. Introduction

Epilepsy is a very common neurological disorder. Nearly 1% of the world's population will develop epilepsy at some point in their lives. Roughly 30% of these epilepsies will become drug-resistant (Kwan and Brodie, 2000), i.e., seizures cannot be controlled through medications. Epilepsy is often triggered by an initial brain injury, which is followed by a clinically silent so-called *latent* phase, during which the brain is undergoing a cascade of structural and functional changes. This process where the healthy brain transforms into an epileptic brain capable of generating spontaneous recurring seizures is called epileptogenesis (Löscher, 2019; Pitkänen and Engel, 2014). Importantly, the longer an epilepsy has been established, the more resistant to treatment it will be. Therefore, to issue early medical interventions and provide the potential epilepsy patients a better chance of living seizure-free lives, it may be helpful to identify epileptogenesis already before the first spontaneous seizure (FSS), which defines the beginning of an established epilepsy (Moshé et al., 2015).

EEG is a popular tool to measure brain activity at a high temporal resolution and it is often used in clinical settings and animal research (Löscher, 2019). The task of detecting epileptogenesis during the latent period, where there are no seizures yet, with EEG is very challenging and under-researched (Engel Jr and Pitkänen, 2020; Pitkänen et al., 2016), since it is often clinically silent. One contributing factor is that the data during this latent epileptogenesis phase is hard to acquire, especially in human patients. Usually, patients receive medical care only after experiencing at least one seizure. In animal epilepsy models, it is possible to acquire EEG data before the onset of the chronic seizures. However, due to a lack of well-established EEG biomarkers and well-annotated datasets, detecting epileptogenesis prior to the first spontaneous seizure remains a big challenge (Pitkänen et al., 2016; Engel Jr and Pitkänen, 2020).

Recent advances in machine learning (ML) offer promising directions for epilepsy research and have delivered encouraging results including seizure forecasting in canines with epilepsy (Nejedly et al., 2019), seizure forecasting and cyclic control in human patients (Stirling et al., 2021), epilepsy detection in clinical routine EEG data (Uyttenhove et al., 2020), as well as epileptogenesis detection and staging in animal epilepsy models (Lu et al., 2020b,a). Specifically, there have been several studies on biomarker discovery for identifying epileptogenesis focusing on high-frequency-oscillations (HFOs) (Bragin et al., 2004; Burnos et al., 2014), dynamics of theta band activity (Milikovsky et al., 2017), asymmetry of background EEG (Bentes et al., 2018), and nonlinear dynamics of EEG signals (Rizzi et al., 2019).

Generally, applying supervised ML to medical diagnosis problems is often hampered by the lack of large amounts of labeled training data. Therefore, we here consider a fully *unsupervised* learning framework that does not require any annotated data. Rather, the idea is to train a model to capture the statistics of normal healthy brain activity and use the model to subsequently detect systematic deviations from the healthy state. In our case, the types and the frequency of anomalous signals indicating the progression of epileptogenesis are not accessible and unpredictable during training. The signals are gradually evolving, which reflects the underlying changes taking place in the brain, evolving from a healthy brain to an epileptic one. This nature of the data renders a large amount of overlapping features between the healthy phase and the epileptogenic phase, which imposes grave difficulties for anomaly detection.

Inspired by the work from Schlegl et al. (2017) and Makhzani et al. (2015), we propose an adversarial autoencoder (AAE) network for anomaly detection in epilepsy progression. AAEs proposed by Makhzani et al. (2015) impose a prior distribution on to the latent codes learned by the encoder through the adversarial training. Here, we propose a flexible framework that makes use of different loss terms such as the reconstruction loss and the distance of the encoding distribution to the prior distribution to compute different anomaly scores.

Here, we would like to emphasize on one fundamental difference between our work and other works on seizure detection and prediction, i.e., there is **no seizure** yet in the data of interest in our work. We focus on detecting slow changes in brain activities before the very **first** unprovoked epileptic seizure aiming for early diagnosis of epilepsy (Fisher, 2015). This is a much more challenging problem that has only been recently addressed, but never with unsupervised methods (to the best of our knowledge).

Specifically, our contributions can be summarized as follows:

- We present an unsupervised adversarial autoencoder framework for detecting slowly evolving anomalies in brain activity.
- We validate our approach with data from a rodent epilepsy model and demonstrate good discriminative ability of signals from different phases of the epileptogenesis process.

### Generalizable Insights about Machine Learning in the Context of Healthcare

In medical applications, massive amounts of data have been collected, however, obtaining expert annotations is extremely expensive and often infeasible. Especially, during the early disease progression phase, e.g., the case of early diagnosis of epilepsy, where the background normal activities are dominating the collected data and only gradual changes of certain features are involved. Our approach provides the opportunity of modeling the normal (healthy) data in an easy-to-acquire clinical setting and of detecting the slow evolution of disease progression in the collected query data. We emphasize that our framework is very general and could be applied to other neurological and psychiatric disorders, supporting early diagnosis and intervention. Moreover, our ablation studies show the significance of using adverserial training to further restrict the prior distribution of the latent space of the autoencoders trained on normal (healthy) EEG data. It led the autoencoders to learn an approximation to the normal (healthy) data distribution that maximized the separability between the normal (healthy) and anomalous (unhealthy) data.

## 2. Related Work

Early diagnosis of epilepsy holds great potential, since it might enable timely treatments that could potentially alter or even halt the disease progression. However, analysing large scale EEG data to discover bio-markers of epilepsy progression is very challenging. Recently, there has been an increasing interest in this area. For example, Rizzi et al. (2019) applied nonlinear dynamics analysis of EEG signals via recurrence quantification analysis. They found a significant decrease of the so-called embedding dimension in early epileptogenesis that correlates with the severity of the ongoing epileptogenesis. Buettner et al. (2019)

identified two frequency sub-bands that are mostly effective in separating a healthy group from an epilepsy group with classic signal processing methods. Applying ML methods, Lu et al. (2020b) investigated the usage of raw EEG time series to distinguish mildly-injured and epileptogenic brain signals and demonstrated the potential of DNN-based methods in epileptogenesis detection. Furthermore, they extended the methods for staging the progression of epilepsy before the manifestation of the first spontaneous seizure (Lu et al., 2020a). In contrast to these supervised methods, we here propose an unsupervised anomaly detection approach, where the model is only trained with EEG signals that have been recorded prior to the disease-inducing injury in a rodent epilepsy model.

Anomaly detection (AD) describes a class of problems to detect samples that do not conform to the regularities of the training data. It can be addressed in a supervised learning, semi-supervised learning, or unsupervised learning fashion given the availability (or not) of sample labels (Gu et al., 2019). It can also be viewed as a one-class learning problem, where the training data are deemed to be the one class of interest. The models are trained to learn a classification boundary, either on a hyperplane (Schölkopf et al., 2001), or a hypersphere (Tax and Duin, 2004; Ruff et al., 2018) to separate anomalies from the nominal data (Shen et al., 2020; Ruff et al., 2019). Various AD methods are based on an encoder-decoder framework. In this framework, the model consists of two parts: an encoder and a decoder. The encoder maps the input into a lower-dimensional latent space representation, which the decoder uses to output a reconstructed version of the input. The reconstruction error between input and its reconstruction is usually used as the anomaly score, i.e., samples with high reconstruction error are deemed to be anomalous (Malhotra et al., 2016; Zhou et al., 2019). In addition, the error between the encoded latent vectors of the original input as well as that of the reconstructed input can be incorporated when defining the anomaly score (Kim et al., 2019). In the case where the knowledge of the anomalies is not accessible or is unpredictable during training, one can impose a regularizer on the learned latent distribution. Abati et al. (2019) propose to equip a deep autoencoder with a parametric density estimator, where the latent vector is generated in an autoregressive fashion. The overall model is trained to minimize the reconstruction error between the input and the output of the decoder network, as well as the log-likelihood of generating the latent vectors given the learned encoder network.

Adverserial autoencoders (AAEs) proposed by Makhzani et al. (2015) extend this notion of anomaly by imposing a prior distribution over the learned posterior by an encoder network through adversarial training. Specifically, an autoencoder is trained to reconstruct the input with low error, and an adversarial training process is applied to match the learned posterior distribution of the latent representation of the autoencoder to a prior distribution. One of the benefits of the AAE framework is the flexibility in choosing the prior distributions (Makhzani et al., 2015). The difference between AAEs and variational autoencoders (VAEs) is that VAEs use a KL-divergence term to impose a prior distribution on the latent code distribution, however AAEs achieve this by the adversarial training procedure. Schlegl et al. (2017) proposed a deep convolutional generative adversarial network trained to capture a manifold of normal anatomical variability in optical coherence tomography images of the retina based on the weighted sum of residual loss, a measure of reconstruction error, and discrimination loss. In Pidhorskyi et al. (2018), the proposed model consists of autoencoders under the adversarial training paradigm. Specifically, the probability distribution

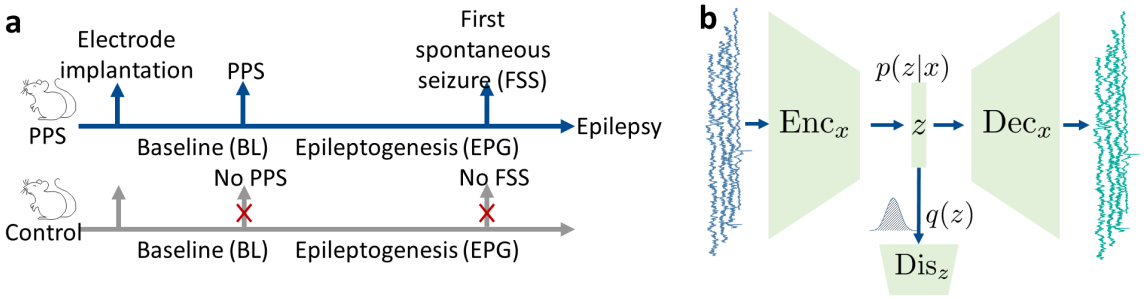


Figure 1: (a) Timeline of the experiment for the stimulated group (top) and the control group (bottom). Perforant pathway stimulation (PPS) is only performed on the PPS group but not the control group.(b) Proposed network structure. The backbone is a standard autoencoder, where the encoder $\text{Enc}_x$ encodes the input into a latent representation $z$ and the decoder $\text{Dec}_x$ reconstructs the input from the vector $z$. $\text{Dis}_z$ is a discriminator that distinguishes whether a sample $z$ is from the encoded representation or drawn from the prior distribution $q(z)$.

of the normal samples is learned through the encoder-decoder framework, and the anomaly score is computed through the evaluation of the probability of the test sample, i.e., normal samples will achieve high probabilities and anomalies will exhibit low probabilities.

It is common that the aforementioned methods assume that during the training there are no anomalous samples. However, in our case, we do not enforce this assumption, and in fact, we expect during the training phase, the model will encounter close-to-anomalous samples due to the nature of the experiment setup. Whilst many anomaly detection problems require label information during training (Gu et al., 2019; Tax and Duin, 2004), our method is completely unsupervised.

## 3. Dataset

The dataset used in this study stems from intracranial EEG recordings with a single depth electrode from a rodent mesial temporal lope epilepsy with hioppocampal sclerosis (mTLE-HS) model, where epilepsy is induced by electrical perforant pathway stimulation (PPS) (Norwood et al., 2011; Costard et al., 2019). Two groups of animals were considered by Costard et al. (2019): (1) PPS-stimulated rats, which developed epilepsy after an average epileptogenesis duration of 24 days (standard deviation 15 days), (2) control rats that had the depth electrode implantation as in the PPS group, but did not undergo the PPS and did not develop seizures by the end of recording (recording time was limited by the lifetime of the battery of the wireless transmitter). Continuous EEG recordings were obtained from the time of implantation of the depth electrodes. On average, a week of pre-stimulation (baseline) period was recorded for all rats. The EEG was recorded at the sampling rate of 512 Hz and band-pass filtered between 0.5 Hz and 176 Hz. Additionally, a notch filter at 50 Hz was applied to all the recordings.

The animal cohort used in this study consists of seven PPS-stimulated rats and three control rats. It is worth noting that during the data acquisition, there are several sources of noise in the signals: (1) electronic interference to the wireless transmission, which results in occasional extremely high amplitude peaks, (2) data loss during the transmission, which results in unchanging values for certain periods. To handle the these problems, we applied an outlier filtering method from MATLAB: `filloutliers` [1] with the parameters `method = 'pchip'; movmethod = 'movmedian'; window = 50`. Furthermore, we discarded the segments that have more than 20% data loss, which resulted in around 5% of the total recordings being discarded. Due to lack of annotations of artifacts such as movements, muscle twitching, chewing, etc., we do not discard them specifically. The time span of the experiment and the different phases are shown in Figure 1a.

## 4. Methods

In this section, we describe the proposed adversarial autoencoder-based anomaly detection method in detail. The main idea is to train our model with only normal data from the training animals and measure the deviation of the test animal data from the learned distribution with an anomaly score based on two performance metrics: reconstruction error and distance of the latent code to the origin of the prior distribution. Code will be available online[2] for reproducability.

### 4.1. Proposed Model

We formulate our task as an unsupervised anomaly detection problem by learning only the distribution of the baseline EEG data through an adversarial autoencoder (AAE). The AAE is composed of three sub-networks: encoder, decoder, and discriminator (Figure 1b). The encoder is trained to map the input data into a lower-dimensional latent space $p(z|X)$, which the decoder uses to reconstruct the input $p(X|z)$. By being trained to discriminate between true samples from the prior distribution and the fake samples generated by the encoder, the discriminator generates a teaching signal to the encoder to generate a latent code that matches the prior distribution. This adversarial loss serves two purposes: first it acts as a regularizer for the training and second it is used as an additional performance metric as we explain later. Specifically, the discriminator is trained with the loss function:

$$\mathcal{L}_{\text{Dis}} = -\log(\text{Dis}(\mathbf{z})) - \log(1 - \text{Dis}(\text{E}(\mathbf{X}))) \,, \tag{1}$$

where $z$ are the true samples from the prior distribution and $X$ are the data samples. On the other hand, the encoder and the decoder are trained with the loss function:

$$\mathcal{L}_{AE} = \|\mathbf{X} - \text{Dec}(\text{Enc}(\mathbf{X}))\|^2 \tag{2}$$

and the encoder/generator is trained with the loss function:

$$\mathcal{L}_{\text{Gen}} = -\log(\text{Dis}(\text{Enc}(\mathbf{X}))) \,. \tag{3}$$

1. https://www.mathworks.com/help/matlab/ref/filloutliers.html
2. https://github.com/amr-farahat/Epileptogenesis

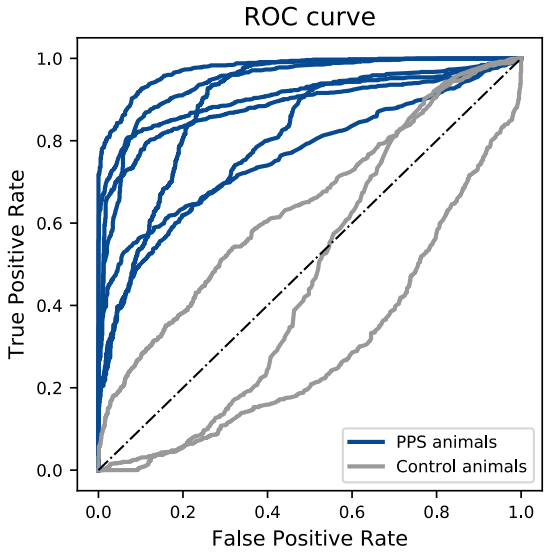


Figure 2: Receiver operating characteristic (ROC) curve for classifying baseline versus epileptogenesis periods for each animal in our dataset ($n = 10$). We show here the ROC curve for the $0.8\mathcal{R}$ anomaly score which is a weighted average of the $\mathcal{R}$ and $\mathcal{D}$ based anomaly scores. We use the count of supra-threshold one-second segments within one hour as an anomaly score. The threshold is selected to be the $99^{th}$ percentile of the training distribution of reconstruction errors and distances to the origin of the prior distribution of the latent space for the $\mathcal{R}$ and $\mathcal{D}$ based anomaly scores, respectively.

Input data are one-second EEG segments collected as described in Section 3. The encoder model is a residual convolutional neural network (He et al., 2016) that consists of two blocks each composed of four residual units. Each residual unit is formed of two convolutional layers with kernel size $= 3 \times 3$ followed by batch normalization (Ioffe and Szegedy, 2015) and RELU activation functions. The number of kernels gradually doubles from 64 to 512 every two residual units and the signal gets downsampled at the beginning of each block with `stride` $= 2$. At last, we have a convolutional layer with a kernel of size $1 \times 1$ to collapse the feature maps into the 128-dimensional latent code. The decoder model follows the same architecture, but with the use of transposed convolutions to upsample the latent code into the original 512-dimensional input size. The discriminator model is a fully connected network formed of two hidden layers each with 1000 units and followed by a leaky RELU activation function with $\alpha = 0.2$. The output layer is formed of one unit with a sigmoid activation function for binary classification.

The model is trained in two phases: a reconstruction phase and a regularization phase. In the reconstruction phase, both the encoder and the decoder are updated to minimize the reconstruction loss (Equation 2). In the regularization phase, the discriminator is first updated to distinguish between the true samples drawn from the prior distribution and the samples generated by the encoder (Equation 1). Then, the encoder/generator is updated to

fool the discriminator (Equation 3). We balance the contributions of both $\mathcal{L}_{AE}$ and $\mathcal{L}_{Gen}$ to the trainable weights of the encoder/generator by a weighting parameter that we set to 0.99 and 0.01 respectively. All parts of the model are updated with the Adam optimizer (Kingma and Ba, 2014) with base learning rate = 0.0002, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The prior distribution is a multivariate normal distribution with $\mu = 0$ and $\sigma = 0.1$ (see more in Section 5.2). We used MATLAB for preprocessing the data and used python for creating and training the models (specifically using the TensorFlow library (Abadi et al., 2015)) and performing the post-hoc analysis of the results. It takes approximately 1.5 hours to train one epoch of 330-360 hours of EEG data.

After training, the AAE can be used to scan the query data to look for deviations from the training data distribution. In the data space, anomalous data are expected to have high reconstruction errors. On the other hand, depending on the nature of the changes in the brain activity due to the disease process (global or local changes in the signal), anomalous data can be expected to either lie in the the low or high probability density areas of the prior distribution used for training (Schreyer et al., 2019). For that reason, we additionally test the value of using the distance of the latent code to the origin of the prior distribution to define anomalous data.

### 4.2. Cross-validation Scheme

We adopt a leave-one-out (LOO) cross-validation scheme where we iterate over the list of all animals (seven PPS rats and three control rats) and in each iteration, we withhold the data from the test animal completely and train the model on the normal data collected from all other (nine) animals. Since we aim for the model to capture the features of a normal EEG signal, we only use the data from the baseline period of the PPS groups. Additionally, we include the data from the control animals from the entire recording period. Note that it is shown that in longitudinal EEG recordings, various noise sources will be introduced due to the degradation of the implanted depth electrodes and changes in the electrode-tissue interface near the electrode (Kappenman and Luck, 2010; Straka et al., 2018). Hence, it is important to include the data from the control animals covering weeks of recording time in order to make sure that the model utilizes epileptogenesis-related features for discriminating between baseline and epileptogenesis periods and not the artifacts induced by the long-term recording. Specifically, we randomly selected 30 hours from the baseline period of each PPS animal and 75 hours from the whole recording period from each control animal to create the training dataset for each test animal in a LOO cross-validation scheme.

### 4.3. Detection Process

After training the full model on the training data from 9 out of 10 animals, we tested the ability of the trained model to distinguish between baseline and epileptogenesis periods of the data from the withheld test animal. Note that animals in the control group did not undergo the PPS. In order to keep the terms “baseline” and “epileptogenesis” consistent between the PPS group and the control group, we use the following notation for the control group. Baseline: one week period after the electrode implantation; epileptogenesis: starting 10 days after the electrode implantation. During testing, we apply the trained model to scan the data from the whole recording period of the test animal and compute the following

metrics for each one-second segment $\mathbf{x}$: the reconstruction error ($\mathcal{R}$) and the distance of the latent code to the origin of the prior distribution ($\mathcal{D}$):

$$\mathcal{R}(\mathbf{x}) = \|\mathbf{x} - \mathrm{Dec}(\mathrm{Enc}(\mathbf{x}))\|^2 \tag{4}$$

$$\mathcal{D}(\mathbf{x}) = \|0 - \mathrm{Enc}(\mathbf{x})\|^2 \tag{5}$$

We set a threshold ($\lambda$) for these metrics based on the statistics of the training data, which is the $99^{th}$ percentile of the distribution of $\mathcal{R}$ and $\mathcal{D}$ computed from the training data. They are denoted by $\lambda_{\mathcal{R}}$ and $\lambda_{\mathcal{D}}$, respectively.

In order to aggregate the evidence from longer recording time periods and at the same time simulate a clinical setting, we compute the number of suprathreshold segments within a certain time window ($\mathcal{T}$ = one hour), for both the baseline and the epileptogenesis data of the test animal and consider this number as the anomaly score ($\mathcal{S}$).

$$\mathcal{S}_{\mathcal{R}}(\mathcal{T}) = \sum_{i=1}^{n} I_{\mathcal{R}i} \tag{6}$$

where

$$I_{\mathcal{R}i} = \begin{cases} 1 & \text{if} \quad \mathcal{R}(\mathbf{x_i}) > \lambda_{\mathcal{R}} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and

$$\mathcal{S}_{\mathcal{D}}(\mathcal{T}) = \sum_{i=1}^{n} I_{\mathcal{D}i} \tag{8}$$

where

$$I_{\mathcal{D}i} = \begin{cases} 1 & \text{if} \quad \mathcal{D}(\mathbf{x_i}) > \lambda_{\mathcal{D}} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $n$ is the number of one-second segments in time window $\mathcal{T}$, e.g., 3600 in one hour.

Consequently, we evaluate the ability of this aggregated anomaly score to distinguish between baseline and epileptogenesis data by computing the receiver operating characteristic (ROC) curve and calculating the area under the curve (AUC). We compute the ROC-AUC with the aggregated $\mathcal{R}$ and $\mathcal{D}$ metrics. Moreover, we investigate whether a weighted average of both anomaly scores would lead to better classification results.

## 5. Results

### 5.1. Epileptogenesis Detection

The main goal of this study is to investigate the potential of using electrical brain activity in an unsupervised way for predicting brain disorders and follow their development as the brain activity deviates from its baseline distribution. We trained an AAE on the baseline intracranial EEG data collected from PPS rats before stimulation and from control rats in a leave-one-out cross-validation scheme. For each test animal, we used the corresponding model to scan its whole data and record the average reconstruction error for each one-second

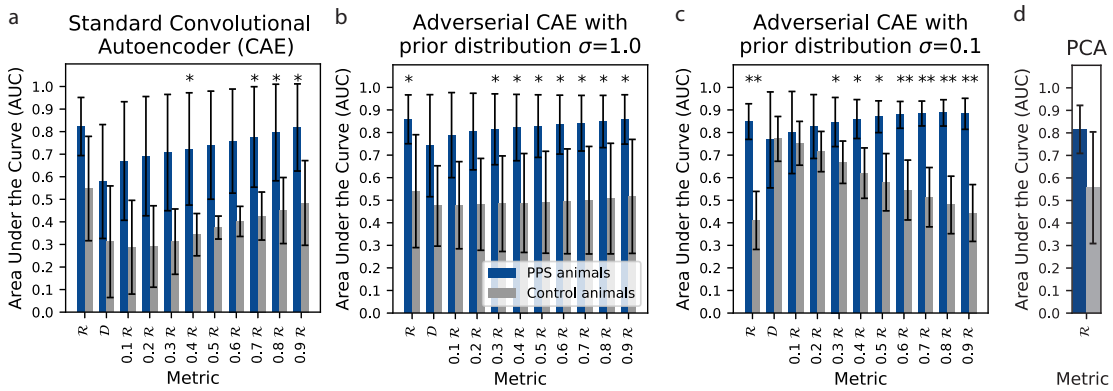


Figure 3: Average area under the curve (AUC) for the PPS animals and control animals for different definitions of the anomaly score and different models. (a) Standard convolutional autoencoder (CAE). (b,c) Adversarial CAE with different prior distributions. (d) Principal Component Analysis (PCA). In each panel, $\mathcal{R}$ denotes the reconstruction error metric, $\mathcal{D}$ denotes the metric based on distance of the latent code to the origin of its prior distribution. The remaining columns consider weighted averages of $\mathcal{R}$ and $\mathcal{D}$; the weight of $\mathcal{R}$ is indicated. The asterisks above the bars denote that the difference between PPS and control animals is statistically significant according to a Mann–Whitney $U$ test (*:$p < 0.05$, **:$p < 0.02$).

segment in the data space. Additionally, we recorded the distance of the latent code to the origin of its prior distribution. We considered different metrics to compute the anomaly score: the reconstruction error in the data space ($\mathcal{R}$) and the distance to the origin of the Gaussian prior distribution in the latent space ($\mathcal{D}$). Using each of these metrics, we computed an anomaly score by counting the number of supra-threshold segments within one hour. The threshold was computed as the $99^{th}$ percentile of the training distribution of this metric. We randomly sampled 1000 hours from each of the baseline and the epileptogenesis periods of the test animal, computed the anomaly scores for them, and calculated the receiver operating characteristic (ROC) curve for discriminating between the two periods for each test animal in the dataset. We also computed additional anomaly scores as the weighted averages of the anomaly scores computed based on the $\mathcal{R}$ and $\mathcal{D}$ metrics which we denote $x\mathcal{R}$ where $x \in [0, 1]$ and represents the weight assigned to the $\mathcal{R}$-based anomaly score where the $\mathcal{D}$-based metric is assigned the weight $1-x$ (see Figure 2 for the ROC curve based on the $0.8\mathcal{R}$ metric as it was our best performing anomaly score and Figure 3c for the average area under the curve (AUC) for all anomaly scores). We observe that control animals have their ROC curves around the diagonal which is expected since they were not exposed to PPS and therefore there should not be a significant difference between their baseline and hypothetical epileptogenesis periods. On the other hand, while there is variability among PPS animals, all their ROC curves lie above the diagonal, which denotes above chance discrimination performance. This is also reflected in the significant difference between the average AUC of PPS and control animals (Figure 3c first two bars). Contrarily, we note that the anomaly score based on the $\mathcal{D}$ metric alone does not show a difference between animal

groups (Figure 3c third and fourth bars), which means it is not a good metric for computing the anomaly score for discriminating between baseline and epileptogenesis periods. Next in the ablation study, we examine the value of the adversarial loss as a regularizer.

### 5.2. Ablation Study

In Figure 3c, we noticed that the discriminative ability of the model using only the $\mathcal{R}$ metric is better than that with only the $\mathcal{D}$ metric. This is reflected in the AUC from control animals being around the chance level for the $\mathcal{R}$ metric and significantly above the chance level for the $\mathcal{D}$ metric. This suggests that the differences between the normal and anomalous data in the data space are too subtle for the encoder to push them into the low-density areas in the lower-dimensional latent space.

To further investigate the relevance of different loss components of the proposed method to the final epileptogenesis detection task, we performed ablation studies. To this end, we trained a standard convolutional autoencoder (CAE) (Figure 3a) and an adverserial CAE with a standard Gaussian prior distribution ($\sigma = 1.0$) (Figure 3b) rather than our proposed method with $\sigma = 0.1$ (Figure 3c). We notice that even though the $\mathcal{D}$ metric did not prove useful alone for computing an anomaly score that maximizes the separability between baseline and epileptogenesis periods, adding the adversarial loss acted indirectly as a regularizer that boosted the discriminability of the $\mathcal{R}$-based anomaly score as evident by the high variability of the average AUC of the PPS and control animal groups in case of the standard CAE (Figure 3a first two columns). Average AUC of PPS animals improved from 0.82 with $std = 0.13$ to 0.85 with $std = 0.08$. Additionally, using the weighted average of both $\mathcal{R}$ and $\mathcal{D}$ based anomaly scores improved the average AUC of PPS animals from 0.85 with $std = 0.08$ to 0.89 with $std = 0.06$ ($0.8\mathcal{R}$) but only when training with prior distribution with $\sigma = 0.1$ while there was no improvement for the standard CAE or when training with prior distribution with $\sigma = 1.0$. This can be explained by the fact that at the beginning of training with standardized inputs and random weights, the encoder already produces a latent code that approximates samples from a standard Gaussian distribution. Consequently, the discriminator does not get the chance to learn the prior distribution and send a teaching signal to the encoder/generator. Therefore, making the problem harder for the encoder/generator by restricting the standard deviation of the prior distribution has a better regularizing effect on the trained models.

Moreover, we compared to a linear baseline for reconstruction-based anomaly detection by using Principal component analysis (PCA) to reduce the dimensionality of the data to 128 components and then project back to the data space and compute a reconstruction error. We merely obtain average AUC for PPS animals of 0.82 with $std = 0.11$ which does not show statistically significant difference to the control group AUCs (Figure 3d). This is comparable to the standard CAE results, but falls short to our best achieved results with the adversarial CAE with a gaussian prior distribution with $\sigma = 0.1$ (Average AUC = 0.89 with $std = 0.06$).

### 5.3. Time Course of Epileptogenesis

We have shown so far that the $\mathcal{R}$-based anomaly score was successful at differentiating between EEG signals recorded during the baseline period and the epileptogenesis period

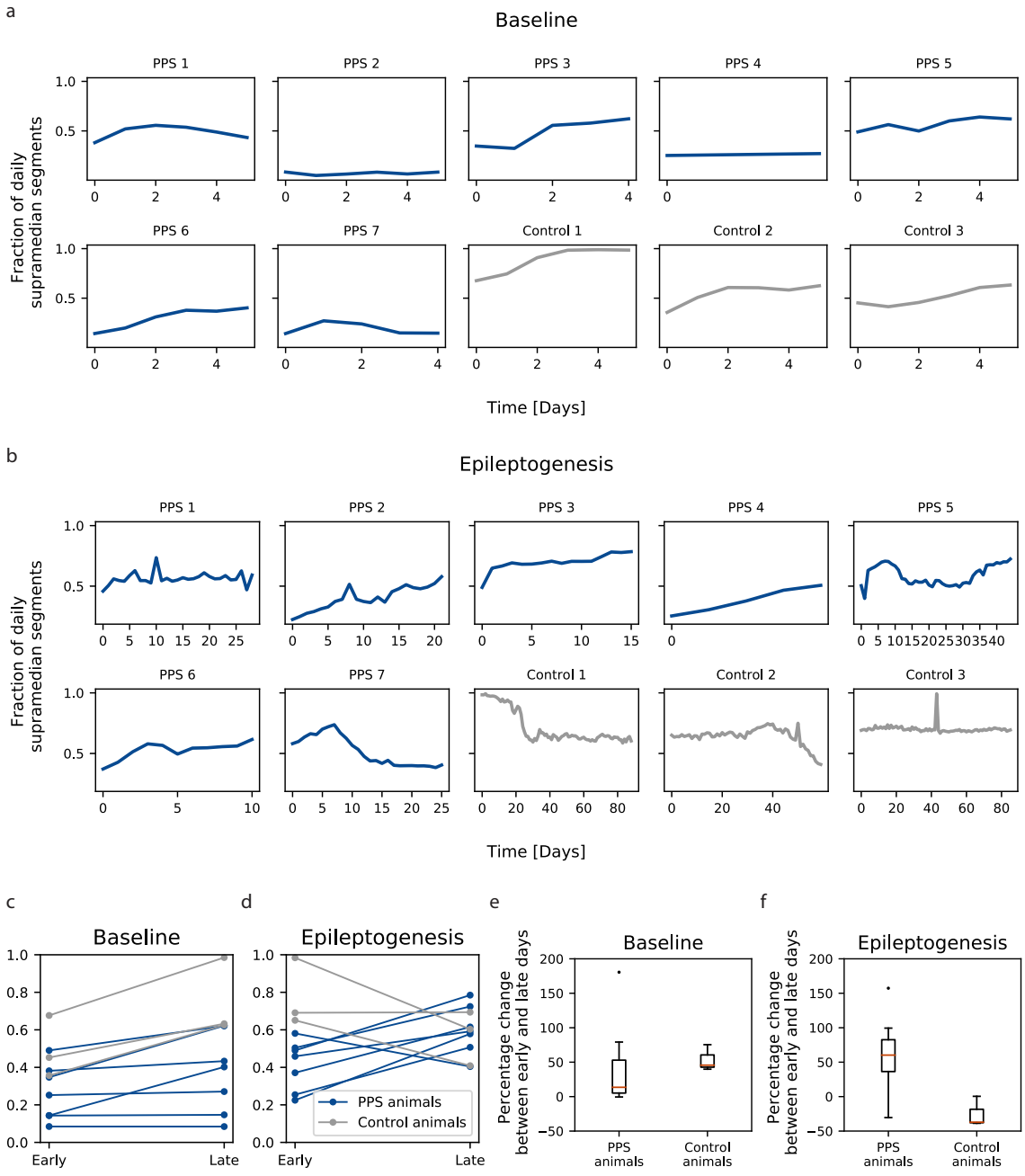


Figure 4: Fraction of one-second segments that exceeds the median of the training distribution of reconstruction errors (supra-median fractions) for each animal in our dataset ($n = 10$) for the baseline (a) and epileptogenesis (b) periods. Comparing supra-median fractions between the first (early) and last (late) day of the baseline (c) and epileptogenesis (d) periods for each animal in our dataset ($n = 10$). Average percentage of change between early and late supra-median fractions for each animals group (PPS or control) for baseline (e) and epileptogenesis (f) periods.

after PPS. Next, we examined what the temporal evolution of reconstruction errors of the EEG signal can reveal about the epileptogenesis process. For each full 24-hour day in the baseline and epileptogenesis periods, we computed the fraction of one-second segments that have a reconstruction error that exceeds the median of the reconstruction error training distribution (fraction of daily supra-median segments in Figure 4). We notice that the time course of supra-median fractions is complex and variable across animals in both periods. However, it is less variable in the baseline period specifically when we consider the difference between the control and PPS animal groups (Figure 4a and quantified in Figure 4c by contrasting the first and the last full-days of the whole period). On the one hand, all animals tend to have either stable or slightly increasing daily supra-median fractions across the whole baseline period. On the other hand, in the epileptogenesis period (Figure 4b and quantified in Figure 4c), control animals tend to have stable or decreasing daily supra-median fractions. This is in contrast to PPS animals, which mostly, with the exception of only one animal (PPS 7), have increasing daily supra-median fractions. Additionally, we computed the percentage change in the daily supra-median fractions between the first and last day for baseline and epileptogenesis periods for each animal in our dataset. Looking at the averages across animal groups for each period (Figure 4e and f), we observe that both animal groups have comparable percentage change in daily supra-median fractions in the baseline period ($p$-value is 0.18 with Mann–Whitney U test). In contrast, PPS animals show significantly higher percentage change than control animals in daily supra-median fractions during the epileptogenesis period ($p$-value is 0.02 with Mann–Whitney U test). These results show that the epileptogenesis process causes alternations to the brain that are reflected in its electrical activity, which is in turn reflected in the ability of the model to reconstruct this electrical signal. These changes in brain activity get progressively stronger and consequently, the reconstruction errors increase.

## 6. Discussion

Machine learning techniques have been transforming many domains of investigation, in particular those that require detecting patterns in vast amounts of data. Healthcare applications have been at the top of the list of these domains, specifically when it comes to diagnosing diseases or rehabilitating patients by training machine learning models on labeled biomedical data like X-Rays (Rajpurkar et al., 2017), magnetic resonance imaging (MRI) (Lundervold and Lundervold, 2019), EEG (Lu et al., 2020a; Farahat et al., 2019), and electrocardiograms (ECG) (Hannun et al., 2019). Machine learning algorithms trained on large amounts of data can discover new patterns in the data, e.g., diagnostic biomarkers, that may be too subtle to be detected by humans. However, one problem is that the data collected in the medical domain are usually imbalanced. There is a scarcity of abnormal data that corresponds to certain diseases and disorders relative to normal data from healthy subjects. Also, collecting data from patients is subject to regulations that protect the privacy of the patients which makes it harder to obtain.

One potential approach to overcome this problem of scarcity of abnormal data is to leverage the abundance of normal data by training machine learning models to learn the distribution of normal data and then survey the query data for deviations from this learned distribution. Clinically, this approach can work as a screening procedure for individuals with

risk factors who can then be further evaluated by professionals. Technically, this approach has the advantage of only requiring the relatively cheap data of healthy subjects. However, this approach is challenging when the deviations from normal data caused by the disease process are subtle (especially early in the disease) and develop gradually over a long period of time, which is the case in epileptogenesis.

In this study, we have given a proof of concept that such an approach can be implemented through an adversarial convolutional autoencoder model. We trained the model on normal EEG data collected from a rodent epilepsy model and used it in an anomaly detection paradigm to screen the data of test animals to discriminate between the data collected before and after PPS, i.e., to detect a developing epilepsy. The anomaly scores were computed based on how the reconstructed signal deviates from the original signal and could be viewed as a proxy of how the epileptogenesis process develops over time after PPS. This is important as anticipating epilepsy before the FSS could urge medical intervention that significantly improves the patients' long-term quality of life (Moshé et al., 2015). Note that we chose the time window of our anomaly score computation to be one hour — which is clinically feasible — to act as a simulation for a clinical routine.

**Limitations** The main goal of this study was to test the potential of an unsupervised deep anomaly detection paradigm in detecting subtle changes in brain electrical activity as a consequence of a brain-altering disease process. Despite the success of the approach, it still falls short of a fully supervised approach. In particular, using the same dataset, a previous supervised approach achieved an average AUC = 0.93 for distinguishing between baseline and epileptogenesis in PPS rats (Lu et al., 2020a), in contrast to 0.89 for our approach. This is expected as in our approach, the model does not have access to any epileptogenesis data. Another difference is that the authors of that study used five-second segments instead of one-second segments used here. However, we also experimented with five-second segments and obtained similar results. Nevertheless, given the advantages of our approach mentioned earlier, it is worth pursuing and with further advances in unsupervised and self-supervised learning techniques, we expect further improvements.

Another limitation of our approach is that we computed anomaly scores on relatively short one-second (or five-second) EEG segments. While our approach aggregates these scores over longer periods of one hour, it does not look for patterns at these longer time scales. This choice was motivated by the fact that identified frequency bands that effectively differentiate healthy subjects from epileptics in the epileptogenesis period lay above 1 Hz (Buettner et al., 2019). However, we can not exclude the possibility that there is additional valuable information in lower frequency bands that are not usually considered in EEG analysis.

A final limitation is the relatively small number of individuals considered in this study.

**Outlook** In the future, we plan to pursue two broad directions with this approach. First, we aim to translate the results to human patients at risk of developing epilepsy. Second, we would like to test the generality of the approach by applying it to other neurological or psychiatric disorders. In particular, several psychiatric disorders are characterized by the alternation of episodes of different "states". Examples are bipolar disorder or schizophrenia. Detecting transitions between these states early and automatically could improve the

management of such disorders. Critically for both research directions is to investigate the applicability of the approach to non-invasive surface EEG recordings.

## Acknowledgments

This work was supported by the China Scholarship Council (No. [2016]3100), the LOEWE Center for Personalized Translational Epilepsy Research (CePTER), and the Johanna Quandt Foundation.

## References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 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.

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.

Ricardo Buettner, Janek Frick, and Thilo Rieg. High-performance detection of epilepsy in seizure-free eeg recordings: A novel machine learning approach using very specific epileptic eeg sub-bands. In *ICIS*, 2019.

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.

Jerome Engel Jr and Asla Pitkänen. Biomarkers for epileptogenesis and its treatment. *Neuropharmacology*, 167:107735, 2020.

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.

Robert S Fisher. Redefining epilepsy. *Current opinion in neurology*, 28(2):130–135, 2015.

Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. Statistical analysis of nearest neighbor methods for anomaly detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 10923–10933. Curran Associates, Inc., 2019.

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):65–69, 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, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

Emily S Kappenman and Steven J Luck. The effects of electrode impedance on data quality and statistical significance in erp recordings. *Psychophysiology*, 47(5):888–904, 2010.

Ki Hyun Kim, Sangwoo Shim, Yongsub Lim, Jongseob Jeon, Jeongwoo Choi, Byungchan Kim, and Andre S Yoon. Rapp: Novelty detection with reconstruction along projection pathway. In *International Conference on Learning Representations*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

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. Staging epileptogenesis with deep neural networks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10, 2020a.

Diyuan Lu, Sebastian Bauer, Valentin Neubert, Laura Sophie Costard, Felix Rosenow, and Jochen Triesch. Towards early diagnosis of epilepsy from eeg data. In *Machine Learning for Healthcare Conference*, pages 80–96. PMLR, 2020b.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016.

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.

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

Petr Nejedly, Vaclav Kremen, Vladimir Sladky, Mona Nasseri, Hari Guragain, Petr Klimes, Jan Cimbalnik, Yogatheesan Varatharajah, Benjamin H Brinkmann, and Gregory A Worrell. Deep-learning for seizure forecasting in canines with epilepsy. *Journal of neural engineering*, 16(3):036031, 2019.

Braxton A Norwood, Sebastian Bauer, Sven Wegner, Hajo M Hamer, Wolfgang H Oertel, Robert S Sloviter, and Felix Rosenow. Electrical stimulation-induced seizures in rats: a "dose-response" study on resultant neurodegeneration. *Epilepsia*, 52(9):e109–e112, 2011.

Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in Neural Information Processing Systems*, 31:6822–6833, 2018.

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.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

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.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, pages 4393–4402, 2018.

Lukas Ruff, Yury Zemlyanskiy, Robert Vandermeulen, Thomas Schnake, and Marius Kloft. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, 2019.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Marco Schreyer, Timur Sattarov, Christian Schulze, Bernd Reimer, and Damian Borth. Detection of accounting anomalies in the latent space using adversarial autoencoder neural networks. *arXiv preprint arXiv:1908.00734*, 2019.

Lifeng Shen, Zhuocong Li, and James Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. *NIPS 2020*, 2020.

Rachel E Stirling, Mark J Cook, David B Grayden, and Philippa J Karoly. Seizure forecasting and cyclic control of seizures. *Epilepsia*, 62:S2–S14, 2021.

Malgorzata M Straka, Benjamin Shafer, Srikanth Vasudevan, Cristin Welle, and Loren Rieth. Characterizing longitudinal changes in the impedance spectra of in-vivo peripheral nerve electrodes. *Micromachines*, 9(11):587, 2018.

David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

Thomas Uyttenhove, Aren Maes, Tom Van Steenkiste, Dirk Deschrijver, and Tom Dhaene. Interpretable epilepsy detection in routine, interictal eeg data using deep learning. In *Machine Learning for Health*, pages 355–366. PMLR, 2020.

Bin Zhou, Shenghua Liu, Bryan Hooi, Xueqi Cheng, and Jing Ye. BeatGAN: Anomalous Rhythm Detection using Adversarially Generated Time Series. In *IJCAI*, pages 4433–4439, 2019.