TODOs

Exploring the Science of Seizure Timing: from Deterministic to Probabilistic Forecasts

Thesis submitted in partial fulfillment of the requirements for the degree of "MASTER OF SCIENCE"

By

NOAM SIEGEL

Submitted to the Department of Computer Sciences at
Ben-Gurion University of the Negev

JULY 2022

Beer Sheva

Exploring the Science of Seizure Timing: from Deterministic to Probabilistic Forecasts

Thesis submitted in partial fulfillment of the requirements for the degree of "MASTER OF SCIENCE"

By

NOAM SIEGEL

This work was carried under the supervision of

DR. OREN SHRIKI

AND

DR. DAVID TOLPIN

The Department of Computer Sciences

Approved by the advisors: _____ Oren   Shriki _____

Approved by TODO:

_____

JULY 2022

BEER SHEVA

Acknowledgements

First and foremost, I would like to thank...

To

"The theory of probabilities is at bottom nothing but common sense reduced to calculus"

Abstract

My abstract

Contents

1. Introduction

NS: write c1Introduction

NS: add illustration for introduction

People who suffer from epilepsy undergo recurring *seizures* - episodes that include loss of consciousness, loss of motor control, and unusual paroxysmal neuronal firing. These episodes are also termed *ictal* episodes. The uncertainty associated with seizure occurrences is deemed to be the leading cause of fear, stress and other comorbidities in patients with epilepsy. Therefore, a reliable method for estimating the likelihood of a near-term seizure is in the interest of epilepsy patients and their caregivers.

We tackle the problem of inferring the probability of an ictal episode at a given time, *conditioned* on observing a window of real-time EEG data. To the best of our knowledge, most machine-learning solutions take a fully supervised approach, therefore depending on labeled datasets for learning the discrimination function. This increases the costly dependence on expert-level domain knowledge. Even clinical-grade annotations, considered to be the "gold standard" in seizure documentation, suffer from observation bias. NS add citation

To overcome this problem, we propose a weakly-supervised Bayesian framework for likelihood estimation of epileptic seizures. We introduce a multilevel probabilistic model of seizure occurrences and clinical annotations (dataset labels). The model takes into account the possibility of missed seizures unrecorded by the annotator. We implemented an algorithm to compute the inference, namely Bayesian Seizure Likelihood Estimator (BSLE). Lastly, we validate the model and examine it's applicability to a practical seizure warning system.

2. Deterministic Seizure Prediction

It is accepted by the epilepsy community that predicting the oncoming of a seizure is a worthy goal [Kelley, Jacobs, and Lowenstein, 2009, Dumanis, French, Bernard, Worrell, and Fureman, 2017]. The outcome of this chapter is the implementation and evaluation of an end-to-end pattern recognition system for EEG data from patients with epilepsy.

2.1. Chapter Overview

The paper which inspired this chapter is [Mirowski, Madhavan, LeCun, and Kuzniecky, 2009]. They show one of the first works to use convolutional neural networks, the same components that boosted the success of deep-learning in image processing. In this case, the classifiers were trained on patterns created by hand-engineered features from segments of 5 minute-long EEG recordings (for example, see figure 2.5), to distinguish between preictal and interictal patterns.

While Mirowski et al. [2009] trained several classifiers (logistic regression, SVM and LeNet5-CNN), on a set of synchronicity-based features, we trained a wider variety of machine learning models, using reusable, modular, scalable python code. We performed 5-fold cross validation of 9 types of classifiers on each of 5 nonlinear and bivariate feature datasets, for a total of 675 model fittings over 3 patients.

> NS: write c2deterministic AUC scores, compared to Miroswki AUC scores. And number of patients with IoC.

2.1.1. Evaluation Metrics

In this project, we report the following metrics for all features-classifiers pairs:

1. ROC AUC

2. Precision

3. Recall

4. fit time

5. score time

And for some select feature-classifier pairs we report the ROC curve as well.

2.1.2. Epilepsy

Epilepsy is characterized by pathological electric activity occurring in the brain. In some individuals, this pathology can vary with time; epilepsy can form in a nonepileptic brain, exhibit changing etiologies, and for some patients be resolved, allowing them seizure-freedom [Kandel, Schwartz, Jessell, Siegelbaum, Hudspeth, and Mack, 2000].

The classification of epilepsy into types is an ongoing debate. The epilepsies are grouped hierarchically according to seizure types, epilepsy types and epilepsy syndromes, and longitudinally according to etiologies (see Figure I. in [Scheffer, Berkovic, Capovilla, Connolly, French, Guilhoto, Hirsch, Jain, Mathern, Moshé, et al., 2017]).

2.1.3. EEGs: electric potential recordings

Electroencephalography (EEG) is a form of neuroimaging which is capable of sensing local-field-potentials near cortical neural populations, electrical activity in the brain. The EEG is an effective tool in epilepsy seizure monitoring and understanding because of its sensitivity to macroscopic neural network dynamics, such as the balance of excitatory and inhibitory processes, or measures of synchronicity among different sensor locations .

> NS
> add citation

The various terms local field potentials (LFPs), electrocorticography[1] (ECoG) and electroencephalography (EEG) all refer to measurements of electric potential: in nerve tissue, on the directly exposed surface of the brain, or on the surface of the scalp, respectively. These types of recording apparatuses measure electric potential which originates from summed electric activity of populations of individual cells (e.g., neurons).

Traditionally, video-electroencephalograms (vEEG) are the central diagnostic tool used by professionals to assess epilepsy in patients. The vEEG records simultaneously both optical video information and EEG data, which allows clinicians to observe simultaneously the subject and EEG dynamics. Due to the independence between the modalities, EEG is often prevalent without video surveillance. In these cases, diagnostics rely more heavily on EEG analysis.

2.1.4. Data

EEGs may be intracranial or placed noninvasively on the scalp's surface. In any case, the resulting data is of the matrical form

$$X \in \mathbb{R}^{c \times T \cdot sfreq} \tag{2.1}$$

where c is the number of channels, *sfreq* is the sampling frequency measured in Hz and T is the time length of the recording measured in s (see figure **??**). The entries of X are typically measured in μVs. iEEG has a higher signal-to-noise ratio, but sEEG has the advantage of noninvasiveness.

---

[1]also termed intracranial EEG.

Data collection

Typical publicly available epilepsy-seizure-prediction datasets consists of long-term raw EEG recordings, supplemented with annotations provided by approved experts [Handa, Mathur, and Goel, 2021]. In these datasets, the occurrence of seizures is reported as a list of ictal (seizure) intervals

$$I := \{Ictal_i\} = \{\langle t^i_{start} | S_i | t^i_{end} | \rangle\} \tag{2.2}$$

where t_{start} is the seizure onset, t_{end} the seizure offset, and S_i, when available, are the additional details reported such as seizure etiology. In this project, we used data from the Epilepsiae Dataset [Ihle, Feldwisch-Drentrup, Teixeira, Witon, Schelter, Timmer, and Schulze-Bonhage, 2012]. Table 2.1 lists raw recording length, sample frequency, number of surface electrodes and size-on-disk per patient.

Table 2.1. Data used in this chapter.

Patient Name	Length	Frequency	Electrodes	SOD
pat_3500	92.8 h	1024 Hz	32	21 GB
pat_3700	79.8 h	512 Hz	32	8.8 GB
pat_7200	94.6 h	1024 Hz	29	21 GB

2.2. Methods

In the paper which inspired this chapter, Mirowski et al. [2009] trained three variants of a neural network, namely support vector machines, logistic regression and convolutional neural network. They evaluated the classifiers on 6 types of bivariate features from the fields of correlation statistics, nonlinear signal analysis, wavelet spectral analysis, and chaos theory. Of these, we selected 5 features (cf. section 2.2.3) which cover all of the feature-category types, and used the implementations in the *mne-features* module given by [Schiratti, Le Douget, Le van Quyen, Essid, and Gramfort, 2018].

2.2.1. Preprocessing

Although it is common to apply spectral filters to EEG data to reduce line noise etc., we chose to focus on the effect of classifier and features choice on predictive performance, and thus kept preprocessing to a minimum. First, EEG channel selection was performed to select a spatially far-reaching distributed subset of 19 channels, common to all patients:

['Fp1', 'Fp2', 'F3', 'F4', 'C3', 'C4', 'P3', 'P4', 'O1', 'O2', 'F7', 'F8', 'Fz', 'Cz', 'Pz', 'T7', 'T8', 'P7', 'P8'] .

This was done to reduce data dimension as well as standardization amongst patients. The raw data was resampled to 256 Hz to reduce disk space and processing time. It was then shifted and scaled to zero mean and unit standard deviation, per patient.

2.2.2. Binary Labelling for Seizure Prediction

A practical way to characterize time-varying brain dynamics is through a sliding window approach (see Fig. 1. in [Lehnertz, Geier, Rings, and Stahn, 2017]). An EEG time series matrix X (see eq. 2.1) is partitioned temporally into non-overlapping windows $x_1, ..., x_N$. Each window x_i is given a label y_i, to form a dataset $D = \{(x_1, y_1), ..., (x_N, y_N)\}$. Optionally, each window is transformed via a feature extraction function $f(x)$, which yields a feature-formed dataset $D_f = \{(f(x_1), y_1), ..., (f(x_N), y_N)\}$.

In binary classification for seizure prediction, each window is matched against the seizure intervals (see eq. 2.2) to create the appropriate label:

$$y_i := \begin{cases} 0 & \text{if } preictal(x_i) \\ 1 & \text{if } interictal(x_i) \end{cases} \tag{2.3}$$

Here, $ictal(x_i)$ means that x_i temporally overlaps with some ictal interval I_j. Similarly, $preictal(x_i)$ means that x_i overlaps with a specific τ_P time interval preceding an ictal interval I_j[2]. Finally, $interictal(x_i)$ means that x_i does not closely preceed or proceed any known seizure interval, with thresholds τ_a before and τ_b after, respectively[3].

2.2.3. Bivariate Feature Extraction Methods

The deterministic classification method relies on hand-crafted, manually engineered, features. Specifically, following [Mirowski et al., 2009], we focus on bivariate measures of synchronicity between pairs of EEG channels (see figure 2.1 for examples).

For each patient, the recording is segmented into 5 minute windows. Each window is segmented into 60 frames, each 5 seconds long. Each 5 second frame is reduced into a vector of length $c \cdot (c - 1)/2$ (where c is number of channels), via one of the feature extraction methods described below. Each 5 minute window is regarded a single pattern with a single label (preictal vs. interictal).

Maximal Linear Cross Correlation (figure 2.2)

In order to quantify the similarity of two signals $\{x_i\}$ and $\{y_i\}$ the maximum of a normalized cross-correlation function is taken as a measure for lag synchronization [Rosenblum, Pikovsky, and Kurths, 1997]:

$$C_{max} = \max_{\tau} \left| \frac{C_{xy}(\tau)}{\sqrt{C_{xx}(0) \cdot C_{yy}(0)}} \right| \tag{2.4}$$

---

[2]In this project, τ_P is taken to be 1 hour.
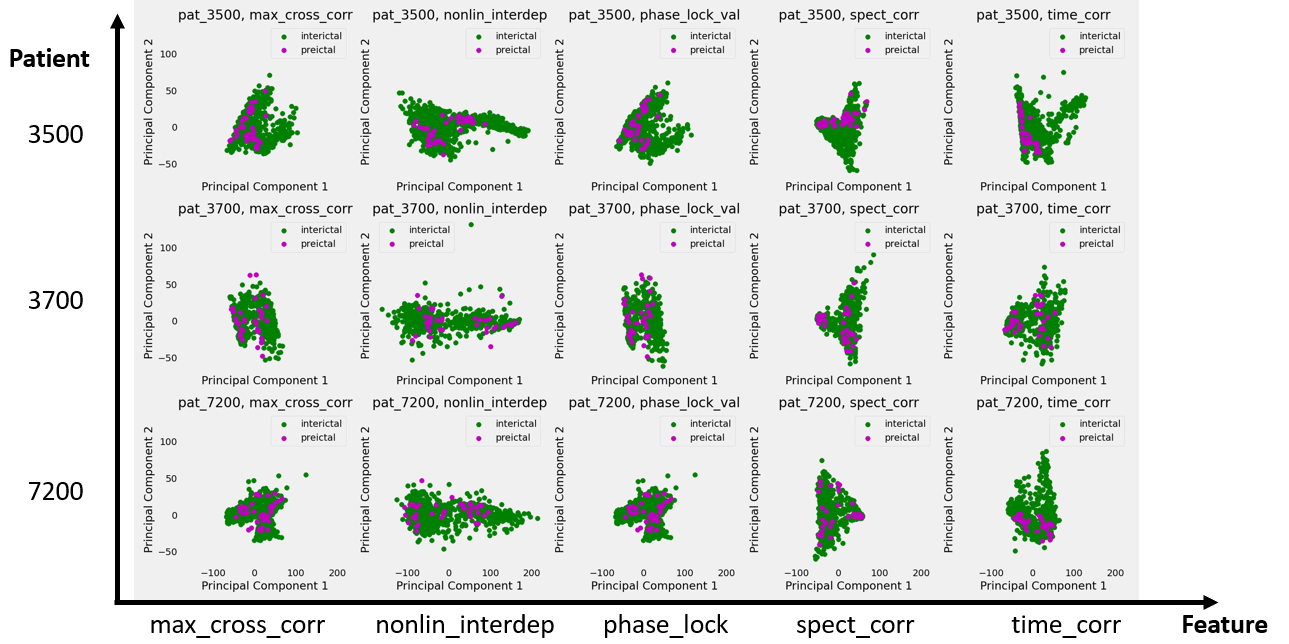[3]τ_a and τ_b are each taken to be 4 hours.

Figure 2.1. Extracted Features

Transformed datasets are visualized in the reduced 2-dimensional PCA space. Each point represents a 5-minute window (i.e., a pattern), and it's color denotes class label (interictal or preictal).

126   Where

$$C_{xy} = \begin{cases} \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} x_{i+\tau} y_i, & \text{for } \tau >= 0 \\ C_{yx}(-\tau), & \text{for } \tau < 0 \end{cases} \quad (2.5)$$

127

128   is the linear cross-correlation function. C_{max} is confined to the interval $[0,1]$ with high values indicating
129 that the two signals have a similar course in time (though possibly shifted by a time lag τ) while dissimilar
130 signals will result in values close to zero.

Phase Locking Value (figure 2.3)

131

132 Introduced in [Lachaux, Rodriguez, Martinerie, and Varela, 1999], the *Phase Locking Value* measures
133 synchronicity between eeg channels in different locations in the brain. First, for each channel i, the instan-
134 taneous phase $\sigma_i^a(t)$ of the analytical signal $x_i^a(t)$ is extracted. Then, for each pair (i, j) of channels, we
135 compute the modulus of the time averaged phase difference mapped onto the unit circle:

$$PLV_{ij} = \left| \frac{1}{T} \sum_t e^{i(\phi_i^a(t) - \phi_j^a(t))} \right| \quad (2.6)$$
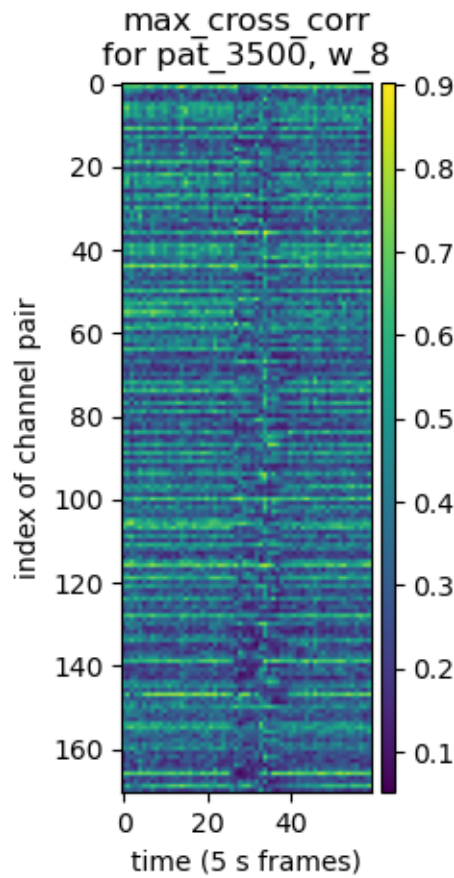
136

Figure 2.2. Maximal Linear Cross-Correlation

Correlation Coefficients and Eigenvalues in the Time and Frequency Domains (figure 2.4)

We compute the correlation coefficients between each pair of EEG channels, along with the eigenvalues of the correlation coefficients matrix, in both the time and frequency domains. This provides two more sets of measures of synchronicity across EEG channels.

Nonlinear Interdependence (figure 2.5)

The *non-linear interdependence* measure for generalized synchronization between two EEG signals $\{x_i\}$ and $\{y_i\}$ is presented in [Mormann, Andrzejak, Elger, and Lehnertz, 2007]. First, the two signals are represented as trajectories in a state space, via time-delay embedding. Then, an asymmetric statistic measuring the Euclidean distance, in reconstructed state-space, between trajectories $\{\vec{x}_i\}$ and $\{\vec{y}_i\}$ is calculated. See [Mirowski, LeCun, Madhavan, and Kuzniecky, 2008] for more details.
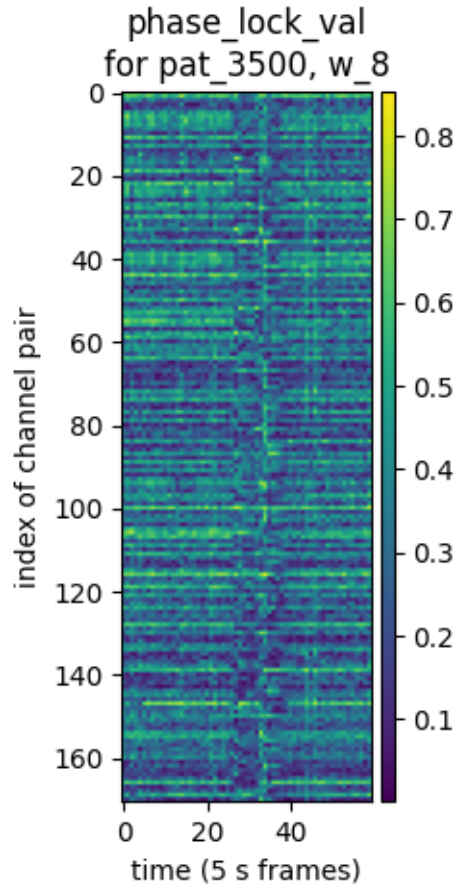
Figure 2.3. Phase Locking Value

2.2.4. Classifiers

We trained and tested 9 of classifiers implemented in the *scikit-learn* [Buitinck, Louppe, Blondel, Pedregosa, Mueller, Grisel, Niculae, Prettenhofer, Gramfort, Grobler, Layton, VanderPlas, Joly, Holt, and Varoquaux, 2013] ML toolkit (v1.0.1), listed in table 2.2. We chose a variety of classifiers from different families (i.e., neural networks, ensembles and decision trees).

2.3. Results

A total of 117 classifier-datasets pairs were trained and evaluated with 5-fold cross validation, yielding a total of 585 model fits.
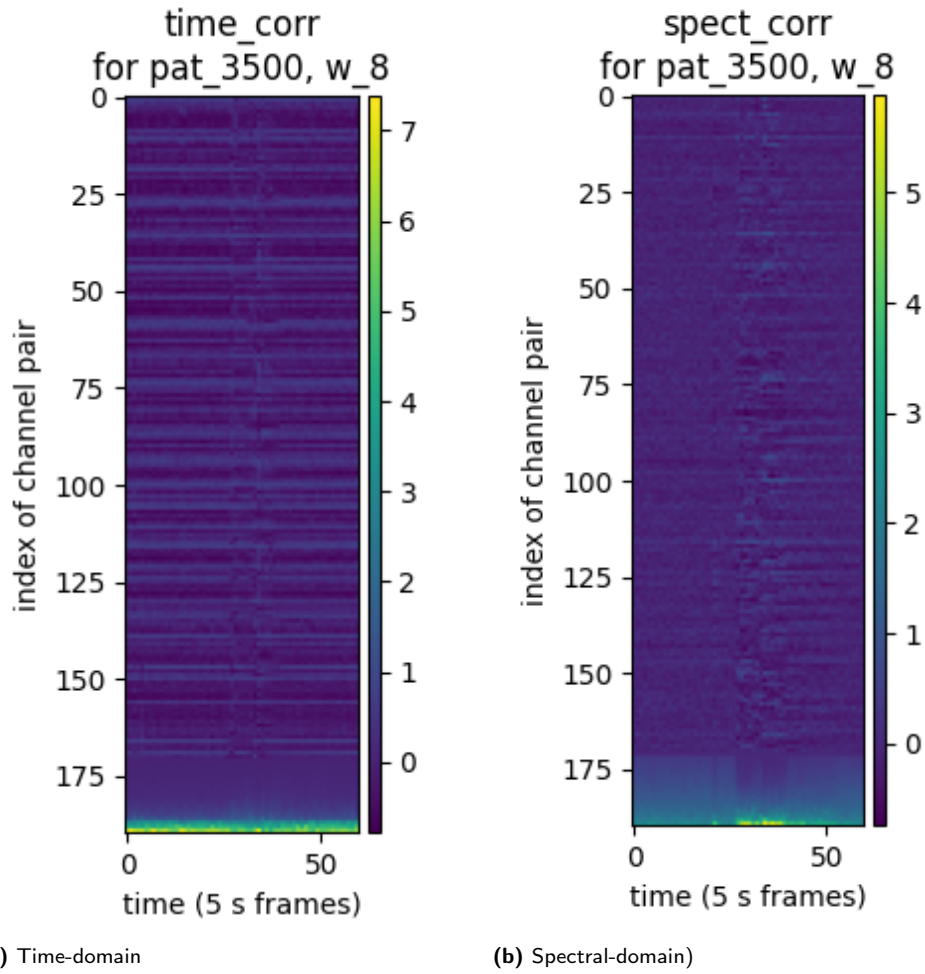
(a) Time-domain   (b) Spectral-domain)

Figure 2.4. Inter-channel correlation coefficients, with eigenvalues appended at the bottom.

2.3.1. Classifier and Feature Comparison

Figures 2.6 and 2.7 show the mean and standard-deviation for each classifier on each feature set, for each patient. It is found that the Linear SVM and Neural Net (MLPClassifier) are the top two performers consistently, sometimes followed closely by the Nearest Neighbor Classifier.

2.3.2. Training Efficiency

Figure 2.8 shows the mean time it took to fit each classifier to each feature-set, for each patient, over 5-fold CV. It is shown that the neural network, ensemble method (AdaBoost) and support-vector-machines consistently take longer than the decision tree, naive-Bayes, quadratic discriminant analysis, and random forest classifiers.
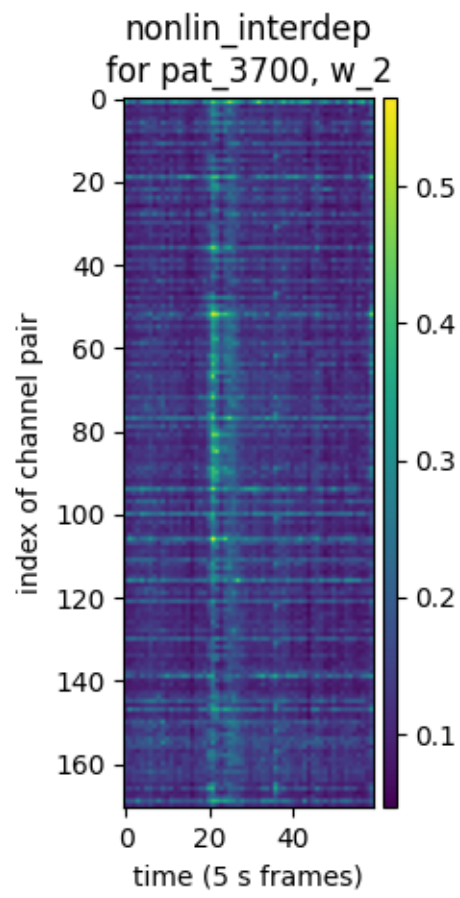
Figure 2.5. Nonlinear Interdependence

Table 2.2. Classifiers used in chapter

| Classifier Name	Parameters
Nearest Neighbors	`K=3`
Linear SVM	`kernel="linear", C=0.025`
RBF SVM	`gamma=2, C=1`
Decision Tree	`max_depth=5`
Random Forest	`max_depth=5, n_estimators=10`
Neural Net	`alpha=1, max_iter=1000`
AdaBoost	default
Naive Bayes	default
QDA	default

2.3.3. Results for Best Classifiers

Since the Linear SVM performed the best in the previous experiments, we compared the results from the Linear SVM classifier on different feature sets. Notably, the classifier performed very similarly on all feature sets, with `spect_corr` slightly outperforming with respect to the ROC AUC score. See figures 2.9, 2.10 and 2.11.

2.4. Discussion

Mirowski et. al. evaluated their pipeline on 6-channel, intracranial data from 21 epilepsy patients. In this project, we use 19-channel data from surface EEG, from 3 patients. They reported 71% sensitivity with 0 false alarms, on 15 out of the 21 patients they assessed. We trained 117 classifiers in different combinations of classifier-dataset. For each patient, the ROC curves for the highest-roc-auc-scoring classifier are presented in figures 2.12, 2.13, and 2.14. As can be seen, only Patient 3700 achieves over 70% sensitivity at the 0 false alarms threshold. For Patients 3500 and 7200, the sensitivities at 0 false alarms are 42% and 40%, respectively. In my opinion, the leading factor in explaining this performance difference is the sensor location: intracranial in the original paper, and extracranial (surface) in our project. Another difference is the number of channels. Perhaps with the increase in channel pairs from 15 pairs to 171 pairs, the sample space grows significantly such that much more data is needed.

2.5. Conclusion

The problem of quantifying the likelihood of a seizure occurrence, namely seizure susceptibility, is an open challenge in the epilepsy research community [?]. The aim of this project is to showcase the predictive power of different classifiers and synchronicity-based features. We trained 117 classifiers in different combinations

Mean ROC-AUC Score
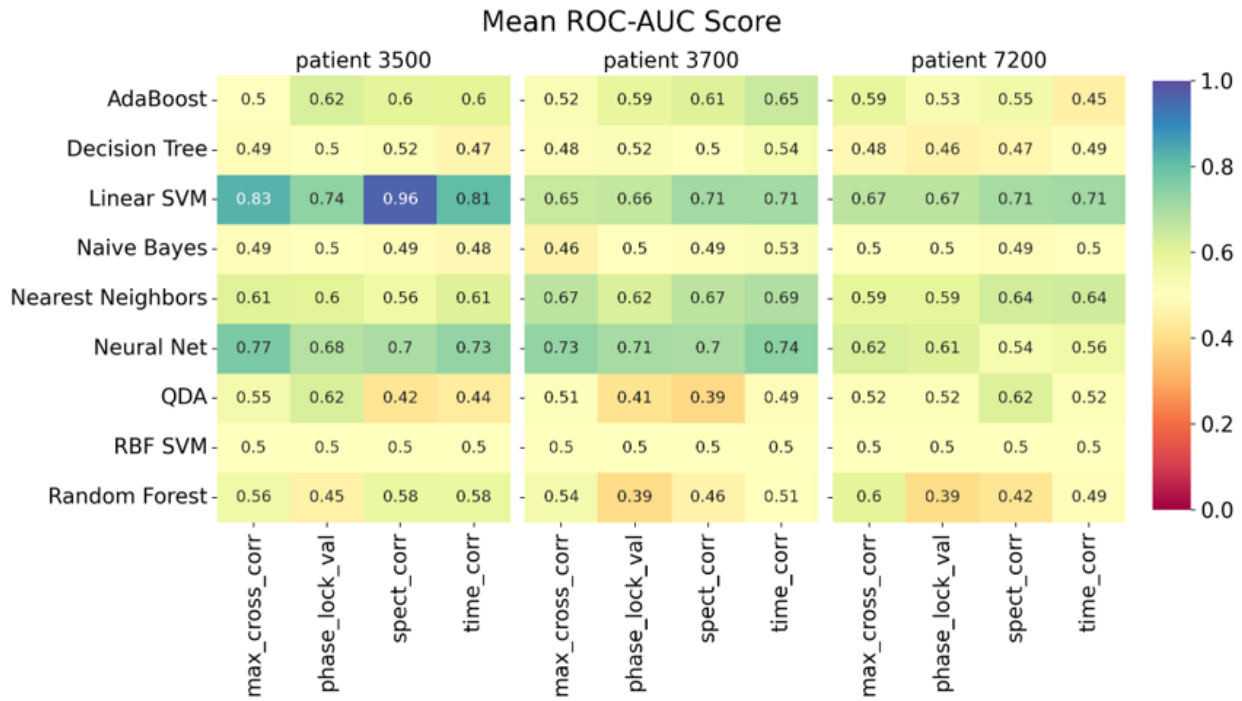
Figure 2.6. Mean AUC-ROC scores for 5-fold cross validation.

STD ROC-AUC Score
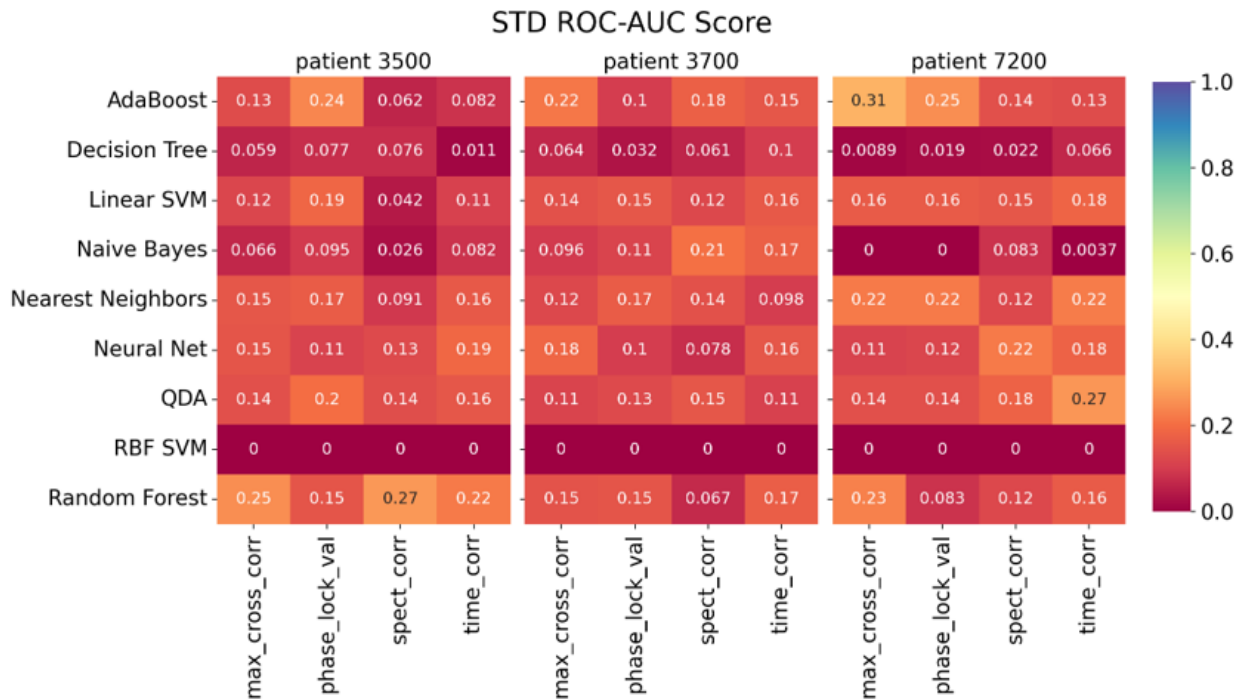
Figure 2.7. Standard deviation of AUC-ROC scores for 5-fold cross validation.

of classifier-datasets. Although some classifiers performed better than others, we did not find any of the feature sets to be remarkably better than the others, especially when generalizing across all 3 patients.
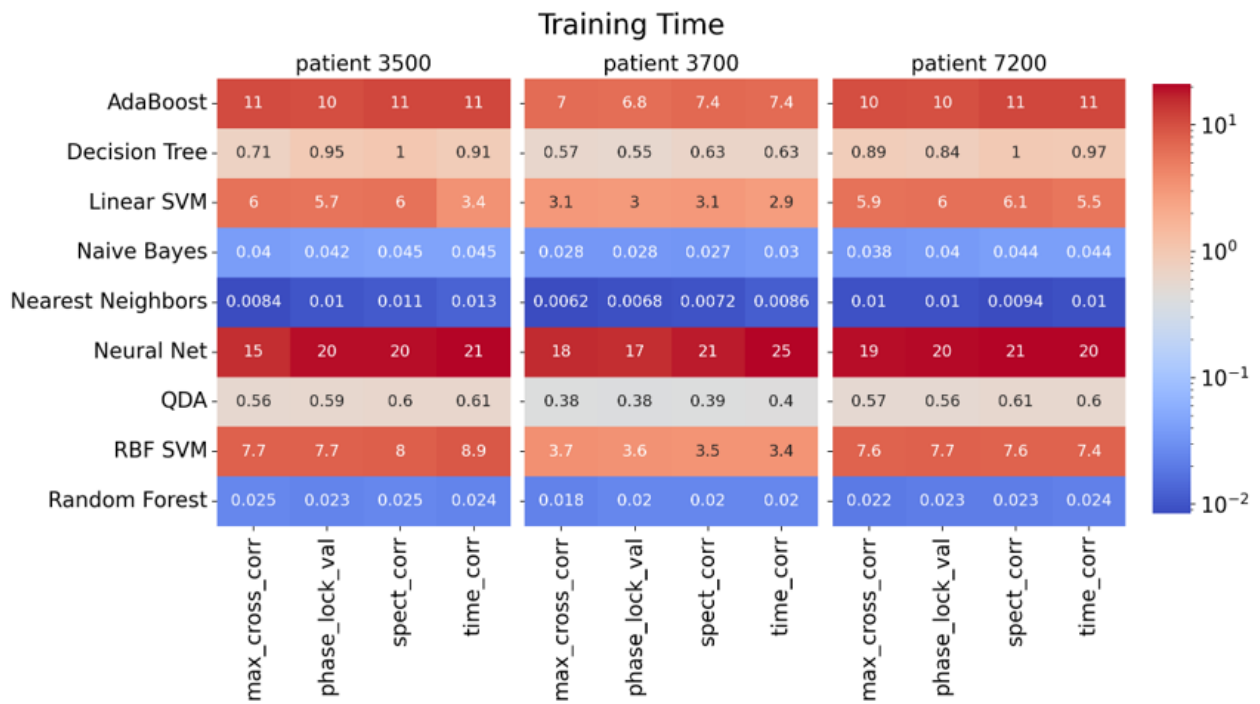
Figure 2.8. Mean training times (in seconds) for fitting the classifiers to the datasets.

NS: write comparison between c2deterministic and Mirowski et al. work
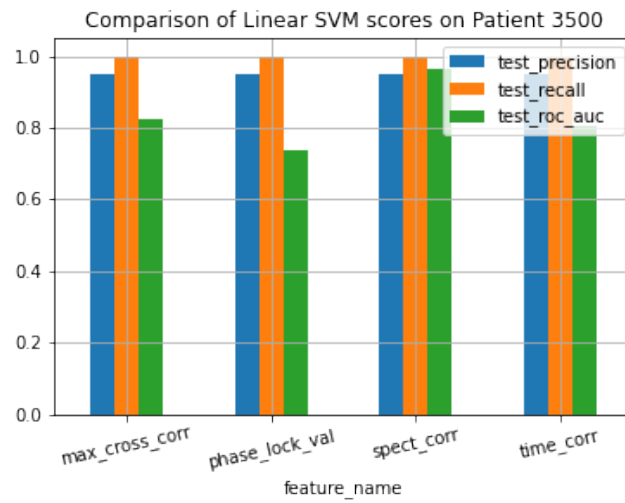
Figure 2.9. Comparison of predictive performance of Linear SVM for Patient 3500 on different feature datasets
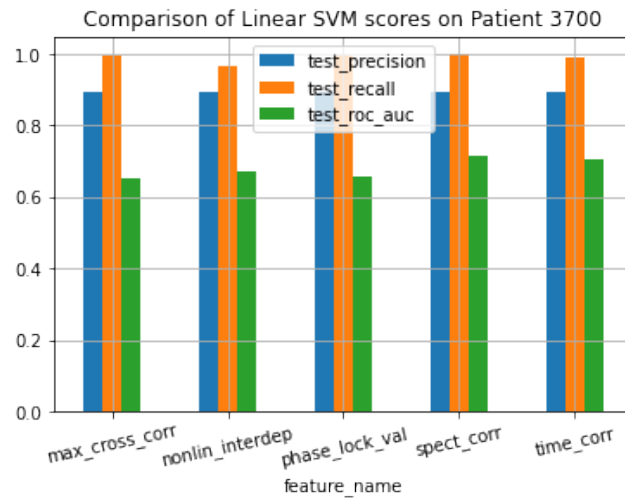
Figure 2.10. Comparison of predictive performance of Linear SVM for Patient 3700 on different feature datasets
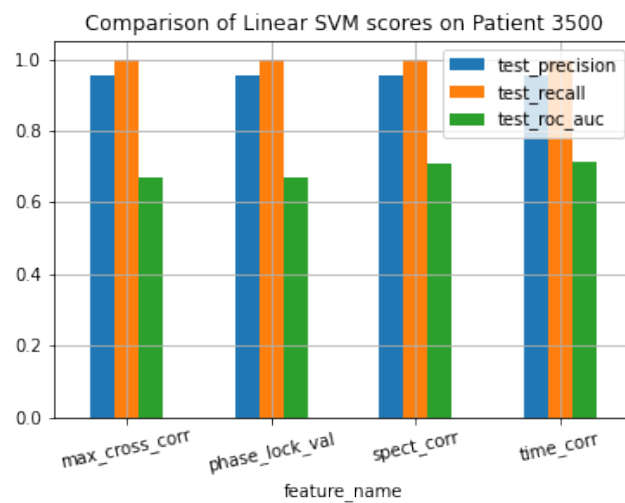
Figure 2.11. Comparison of predictive performance of Linear SVM for Patient 7200 on different feature datasets

14
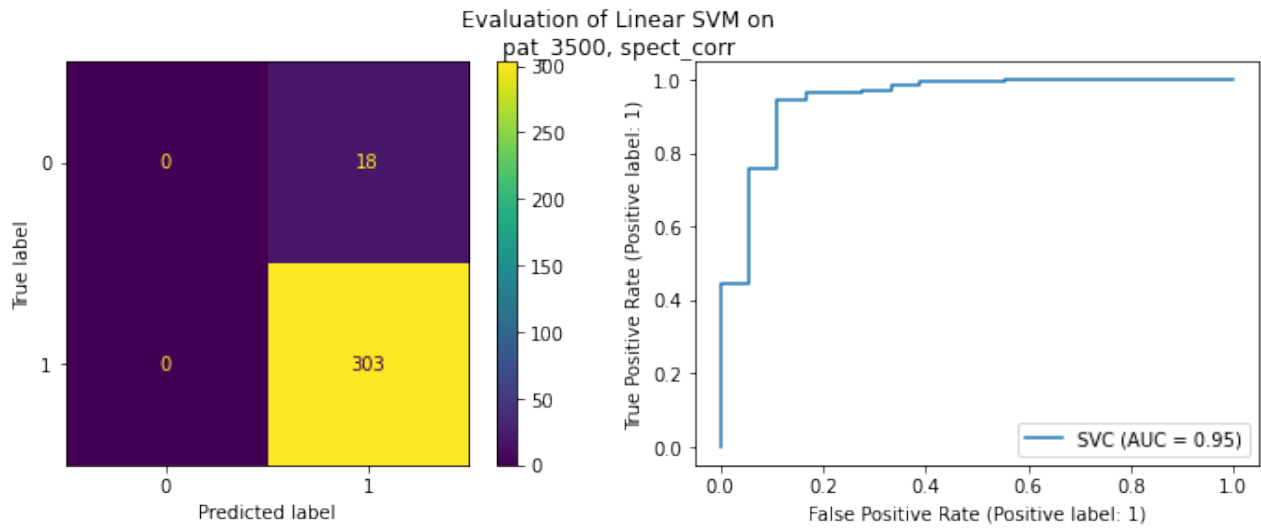
Figure 2.12. Plot of Confusion Matrix and ROC Curve for the highest scoring classifier for Patient 3500

Figure 2.13. Plot of Confusion Matrix and ROC Curve for the highest scoring classifier for Patient 3700

Figure 2.14. Plot of Confusion Matrix and ROC Curve for the highest scoring classifier for Patient 7200

3. Bayesian Seizure Likelihood Estimation

Following a line of recent works [Karoly et al., 2017, Baud et al., 2018, Baud, Proix, Rao, and Schindler, 2020], we consider a probabilistic approach to seizure prediction, expressed via an algorithm we term Bayesian Seizure Likelihood Estimation (BSLE)[1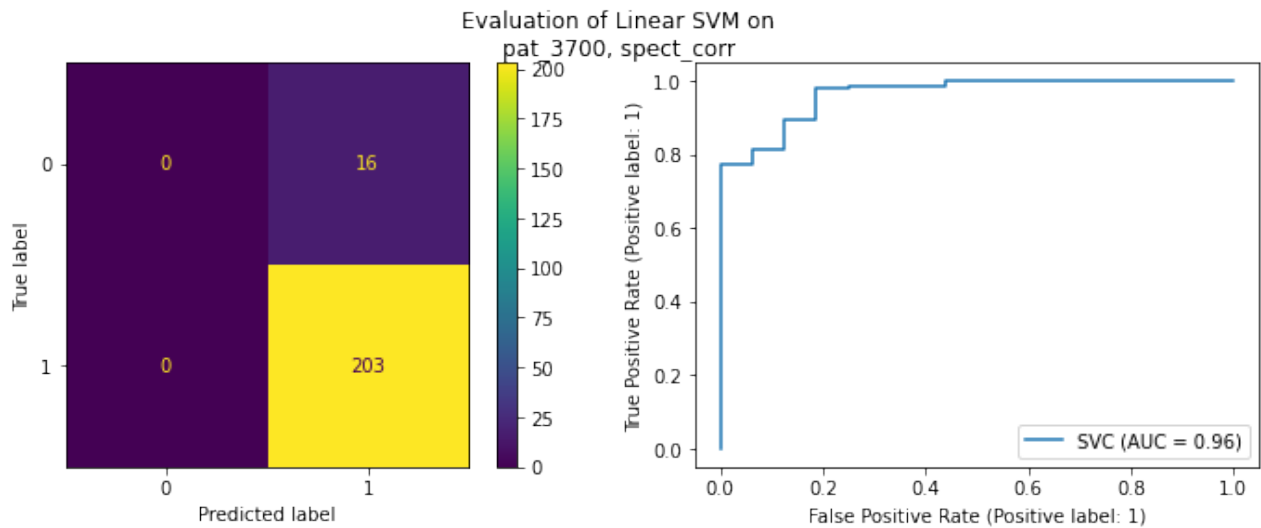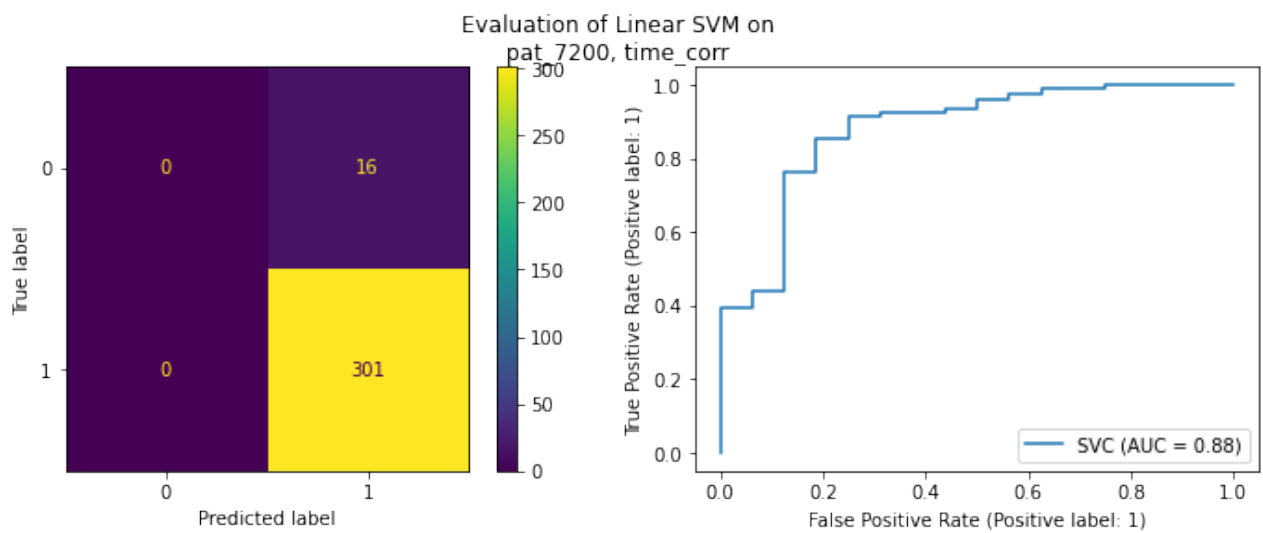]. This chapter focuses on the introduction of preliminary concepts necessary to understand the general approach, as well as the utilities chosen in our implementation.

3.1. Chapter Overview

3.1.1. General approach

Our approach utilizes pure Bayesian statistics wherever possible. We develop a probabilistic model for an observable EEG time series, and then incorporate weak supervision through the use of a domain-specific prior. Our algorithm, Bayesian Seizure Likelihood Estimation (BSLE), performs inference computations with the developed model, to assess the likelihood of an upcoming seizure. The output is tested and evaluated for model fit and predictive capabilities.

3.1.2. Problem setup

Consider the problem of modeling the relationship between EEG signals (E) from the epileptic brain, the occurrence of seizures (S), and an external stream of timestamps denoting seizure events, commonly provided by expert annotators (A), throughout time (T).

Specifically, we are given a dataset $D = \{e_t, a_t\}_{t_0}^{t_f}$ where $e_t \in \mathbb{R}^{c \times \tau}$ is the observed EEG segment with c-channels of duration τ recorded at time t, and $a_t \in \{0, 1\}$ is a clinically-approved annotation denoting whether or not a seizure event began within the time-window $[t, t + \tau]$.

We wish to construct a model for $\mathbb{P}[S, E, t]$, and then apply it with Bayes' rule to infer the likelihood of a seizure at time t determined by an EEG recording:

$$\texttt{probability}[\texttt{seizure=}S \mid \texttt{time=}t, \texttt{EEG=}E] \propto \mathbb{P}[E \mid S, t]\mathbb{P}[S \mid t] \tag{3.1}$$

It should be evident that this procedure is general in that each component on the r.h.s. can be estimated independently, and then combined via multiplication.

---

[1] code can be found at https://github.com/noamsgl/msc under the MIT license.

3.1.3. Methods for evaluating forecast skill

NS: write about probabilistic forecast skill evaluation methods

3.2. Unsupervised Seizure Likelihood Estimation

3.2.1. Gaussian processes

A Gaussian process (GP) [Rasmussen, 2006] is *a collection of random variables, any finite number of which have a joint Gaussian distribution*.

A Gaussian process $f(x)$ is fully specified by a mean function $\mu(x)$ and a covariance function, or a kernel, $k(x, x')$, by-way-of:

$$m(x) = \mathbb{E}[f(x)] \tag{3.2}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{3.3}$$

And it is denoted:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{3.4}$$

In this work we will take the mean function to be zero.

parameter estimation (inference)

Gaussian processes are commonly used for time series modeling with machine learning. To see why this makes sense, imagine the input x is the time point, and the output $f(x)$ is the time series value at time x. Computationally, this is made feasible by evaluating the function's values at a finite number of points of interest. Using optimization techniques, the model's hyperparameters are inferred to match observed data by maximizing the likelihood function $p(f(x) \mid \vec{\theta})$ (termed maximum likelihood estimation, or MLE). The learned hyperparameters capture global evolutionary dynamics of the time series (see figure 3.4 in the Methods section).

The Matérn class of covariance functions

The Matérn class of covariance functions is given by:

$$k_{Matern}(x, x') = \frac{2^{(1-\nu)}}{\Gamma(\nu)} (\sqrt{2\nu}d)^{\nu} K_{\nu}(\sqrt{2\nu}d) \tag{3.5}$$

Where:

- $d = (x - x')^T \Phi^{-2}(x - x')$ is the distance between x and x' scaled by the *lengthscale* parameter Φ.

- ν is a smoothness parameter. In this work, it is taken to be $\frac{3}{2}$.

- K_{ν} is a modified Bessel function.

Multitask Gaussian processes

In case $f(x)$ is a vector function, multiple output functions are modeled in conjunction for the same input values, so-called multitask Gaussian process modeling. In this case, given inputs x and x', and tasks i and j, the covariance between two datapoints and two tasks is given by:

$$k([x, i], [x', j]) = k_{inputs}(x, x') \cdot k_{tasks}(i, j) \tag{3.6}$$

Where k_{inputs} is a standard kernel (e.g., Matérn) that operates on the inputs, and k_{tasks} is a lookup table containing inter-task covariance. This is akin to capturing the inter-channel synchronicity with manually engineered features (see **??**).

Gaussian mixture models [Theodoridis, 2015] are used to model the distribution of an unknown set of vectors $\{x\} \subseteq \mathbb{R}^l$ as a linear combination (i.e., a mixture) of different Gaussian distributions, that is,

$$p(x) = \sum_{k=1}^{K} p_k p(x \mid k; \zeta_k) \tag{3.7}$$

where $\{\zeta_k\}$ parametrize the individual Gaussian distributions:

$$p(x \mid k; \zeta_k) = p(x \mid k; \mu_k, \sigma_k) = \mathcal{N}(x \mid \mu_k, \sigma_k) \tag{3.8}$$

Fitting the model provides an approximation $\hat{p}(x \mid k; \zeta_k)$ of the dataset's underlying pdf, which is an estimate of the data-distribution of the GP-hyperparameters $P(E \mid D)$

3.2.2. Data

In this work we use the *Canine-Epilepsy Dataset* [Davis et al., 2011, UPenn and Mayo Clinic and Kaggle, 2014]. Three dogs with naturally occurring epilepsy are monitored continuously, for 330/451/475 days

for Dog 1, Dog 2, and Dog 3, respectively. The monitoring systems consist of 16-electrode EEG sensors implanted in each brain, digitally sampling at the rate of 400 Hz. In addition, a sequence of timestamps are provided for each dog, marking the observed seizure events (a.k.a. annotations).

> NS: insert illustration of data collection

Train-test split

> NS: write about train-test split (offline-online modes)

3.2.3. Embedding of EEG

Our unsupervised embedding method appears to show high distinguishability between interictal and ictal data. As seen in figure 3.1, in the GP parameter space the seizures are well separated from the non-seizure segments.

   In order to further demonstrate the distinguishability between the interictal and ictal states, we show in figure 3.2 that a linear-kernel SVM fit to the training set achieves 0.91 AUC on the test set in the single-channel case. The double-channel case scores higher.

3.2.4. Density estimation

We use density-estimation of the interictal segments to contain the model of normal EEG. The data likelihood was maximized for a Gaussian mixture model with 2 components, fit to the interictal segments. Figure 3.8 shows the the contour plot of the log-likelihoods scored by the GMM.

3.2.5. Anomaly detection and ictality

EEG embedding

Representing a typical 10-second EEG segment requires about 6.4×10^4 scalar values in it's raw form, thus special care is taken to overcome the curse of dimensionality. In this work, we embed the EEG data in a low-dimensional space, namely the space of GP parameters. We found the ictal embeddings to be distinguishable from the interictal embeddings even after the dimensionality reduction.

   Following the described embedding of the EEG signal in the space of GP parameters, we adopt the new notation $\mathbb{E}_t \in \mathbb{R}^d$, where the new dimension satisfies $d \ll d' \times T$.

Gaussian process parameters embedding

Inference of Gaussian process (GP) parameters is a well-documented approach to modeling time-series data [Rasmussen, 2003]. The extension to multitask GPs enables modeling of multivariate time-series, such as
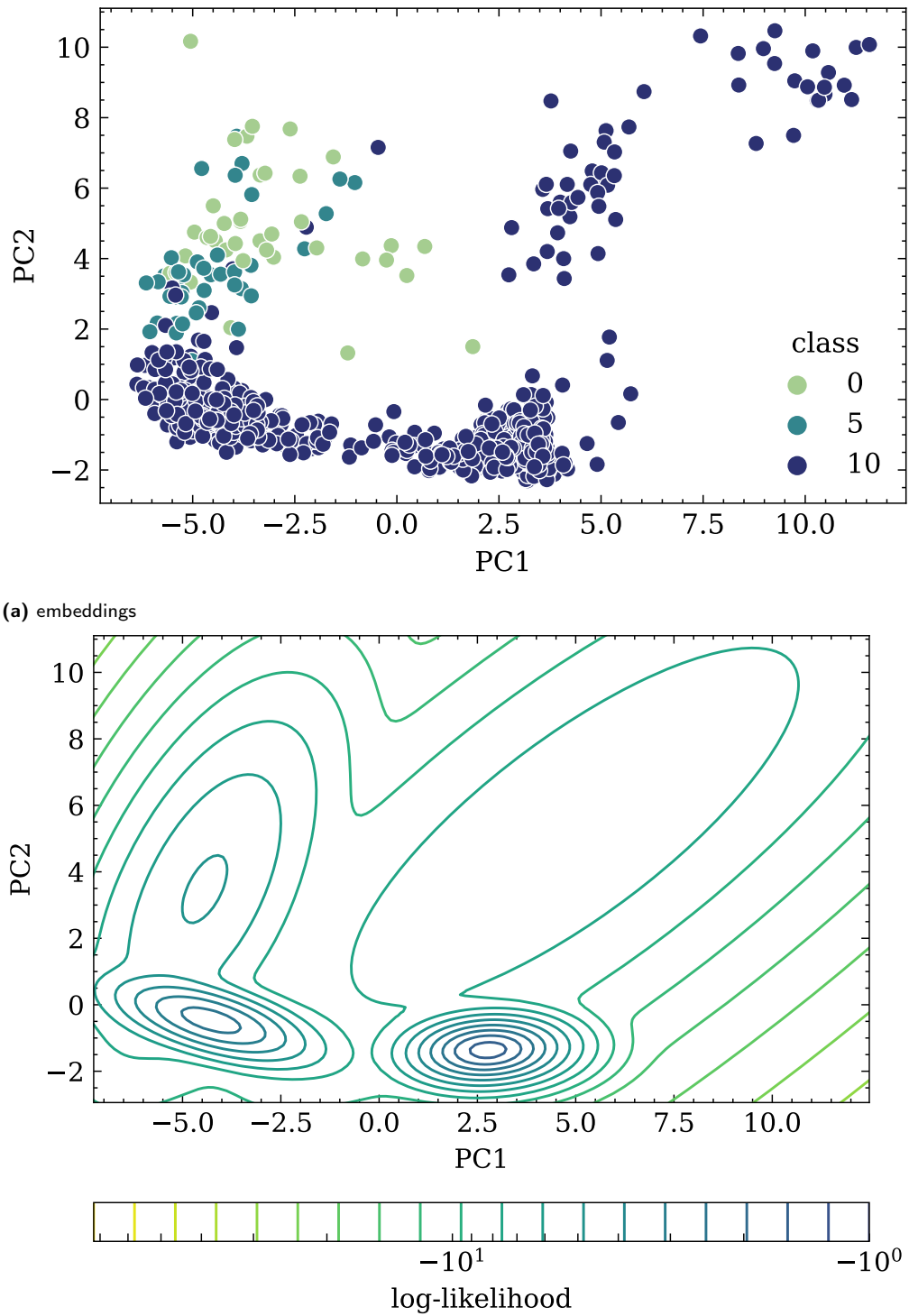
(a) embeddings

(b) log-likelihood as predicted by GMM

Figure 3.1. | **Embedding of EEG.** 2-channel segments of duration 10 (s), randomly sampled from the training set, are embedded in the GP parameter space. A GMM fit with expectation-maximization provides an approximate density estimation. Higher values are more likely to be generated by the GMM.

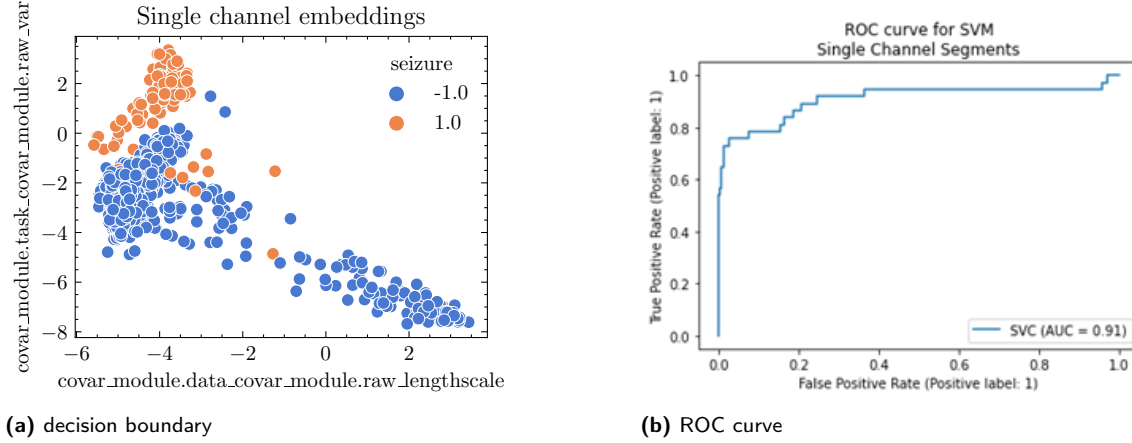NS: fix class labels in legend, remove redundant axis label

(a) decision boundary

(b) ROC curve

Figure 3.2. | **Separability in the parameter space.** A support vector machine achieves 0.91 AUC-ROC on a held-out test set for single-channel segments.

NS: remake seperability test figures

287 the case of multi-lead EEG signals.

288 For each EEG segment x, the preprocessing steps include:

289   i Normalizing x by subtracting the mean and dividing by the standard deviation of the training set.

290 ii initializing a GP model with zero mean, a scaled Matérn-1.5 kernel, and a rank-1 multitask covariance
291   kernel.

292 iii Optimizing the model's parameters to obtain a maximal marginal log-likelihood (details in appx. A).

293 The optimized model's parameters θ are persisted and used henceforth to represent the original EEG
294 segment x.

## $\mathcal{Z}(E)$ - Novelty Score

NS: reconsider if novelty score is necessary

297 We derive a generative, unsupervised novelty score $\mathcal{Z}(E)$ which will be used to define the seizure likelihood
298 , $P(E \mid S) \propto \mathcal{Z}(E)$.

299 For a patient with epilepsy, we hypothesize that the anomalies in EEG data are inherently likely to reflect
300 seizures. Thus, we substitute the likelihood term in Bayes' theorem with the novelty score $z(\theta)$ to provide
301 an estimate of the likelihood of observing an EEG:

$$P(E \mid S) \equiv z(E) \tag{3.9}$$

303 We now show that our unsupervised anomaly detection method is suitable for detecting seizures in the
304 dataset. Figure 3.5 shows interictal and ictal EEG samples, along with the calculated p-values for each clip.
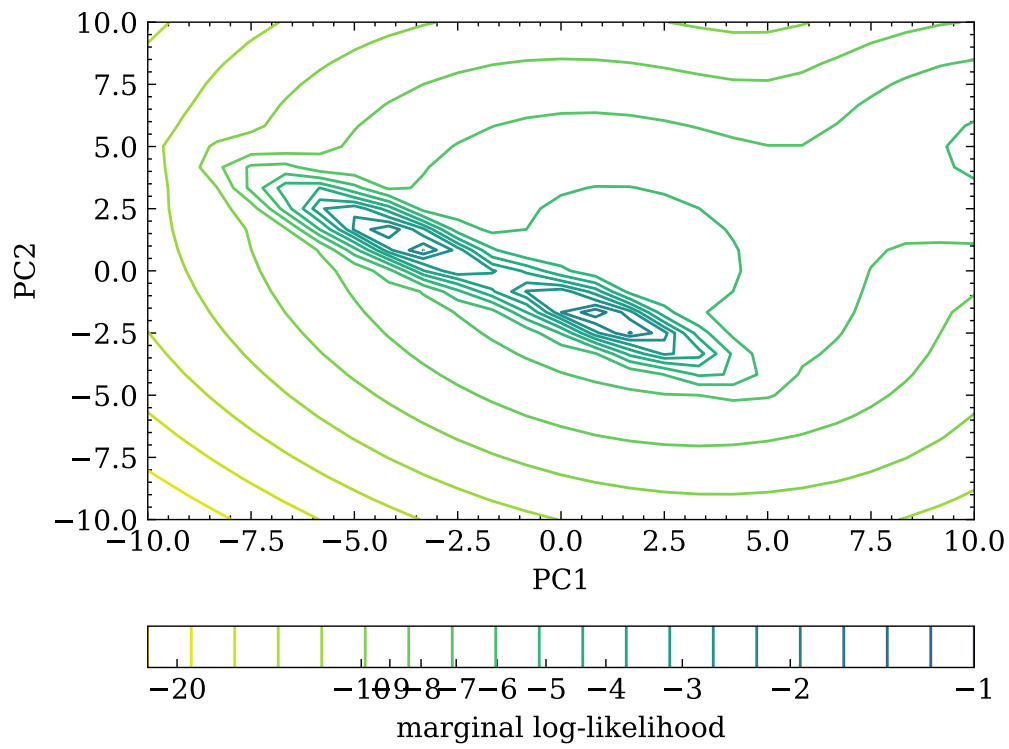
(a) Frist and second principal components
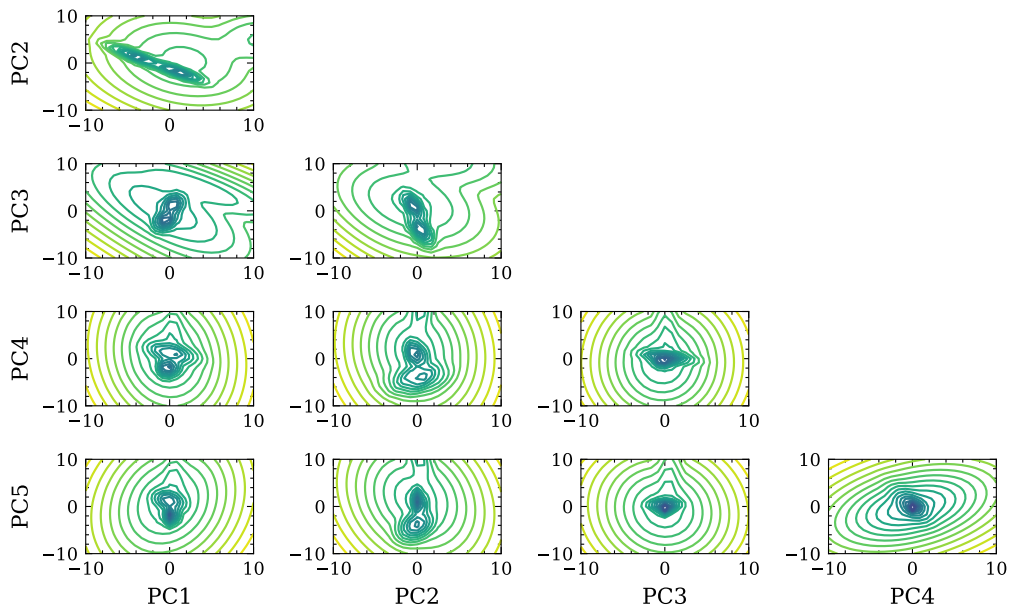
(b) Pair plots for the first five principle components

Figure 3.3. | **Gaussian Mixture Model Density Estimation.** caption
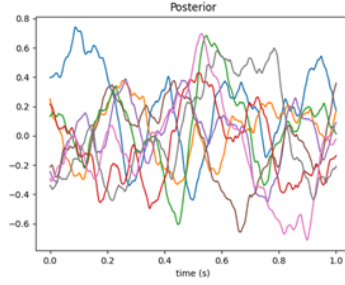
NS: write caption

Figure 3.4. | **Embedding EEG with GP hyperparameters.** Posterior draws from a Gaussian process fit to maximize the likelihood of an observed single-channel EEG segment.

NS: replace with vectorized image

NS: add original sample

305 As can be seen, the interictal clip has a p-value of 0.4355 which indicates that it is in the normal distribution.
306 On the other hand, the ictal clip has an extreme p-value of 0.9787 which indicates that it is a rare event, thus
307 in a patient with epilepsy it is likely to be a seizure.

3.3. Weakly Supervised Bayesian Seizure Likelihood Estimation

310 In the case of epilepsy, handling uncertainty in the face of evidence plays a major role, thus naturally appealing
311 to the mathematical machinery termed Bayes' theorem[2].

312 We apply Bayes' theorem to estimate the updated likelihood of a seizure after observing an EEG signal:

$$\underset{posterior}{\mathbb{P}(S \mid E)} = \frac{\mathbb{P}(E \mid S)\mathbb{P}(S)}{\mathbb{P}(E)} \propto \underset{likelihood}{\mathbb{P}(E \mid S)} \cdot \underset{prior}{\mathbb{P}(S)} \tag{3.10}$$

314 We now introduce our method to calculate the likelihood and prior, thus completing the description of
315 our inference procedure.

$$\mathbb{P}(S \mid E) = 1 - \frac{\mathbb{P}(\neg S)\mathbb{P}(E \mid \neg S)}{\mathbb{P}(E)} \tag{3.11}$$

317 due to the assumption $\mathbb{P}(S) + \mathbb{P}(\neg S) = 1$.

---

[2]much credit is due to Pierre Simon Laplace, who developed the form in common use today [McGrayne, 2011].
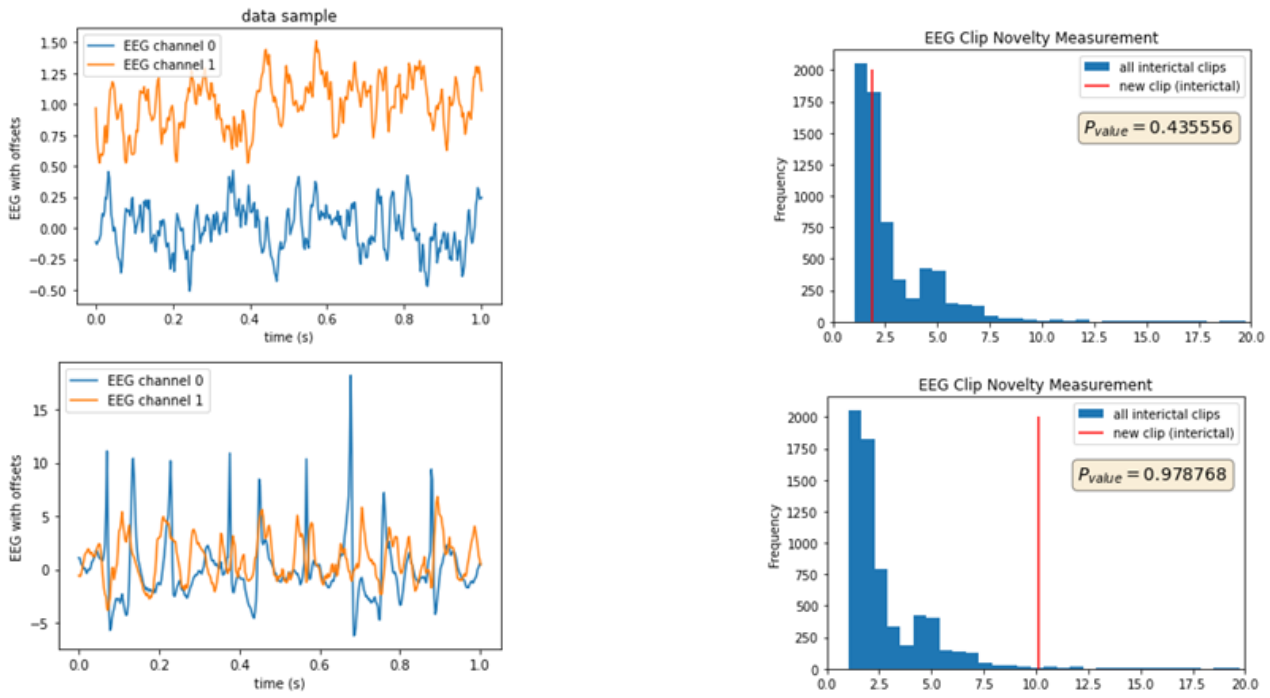
Figure 3.5. | **Anomaly detection.** An interictal double-channel segment has a p-value of 0.435 (top). An ictal double-channel segment has a p-value of 0.978 (bottom).

NS: remake example likelihood estimation figures

$$\mathbb{P}(E \mid \neg S) = \mathbb{P}(\{\hat{p}(e \mid \theta^*) \leq \hat{p}(E \mid \theta^*)\}) \tag{3.12}$$

$$\mathbb{P}(E) = \mathbb{P}(\{pdf(e) \leq pdf(E)\}) \tag{3.13}$$

3.3.1. Subject-dependent circadian-profile prior

Observing the circadian seizure distribution

NS: remake circadian profile diagrams panel horizontal

The von Mises (circular Gaussian) distribution

NS: format as definition
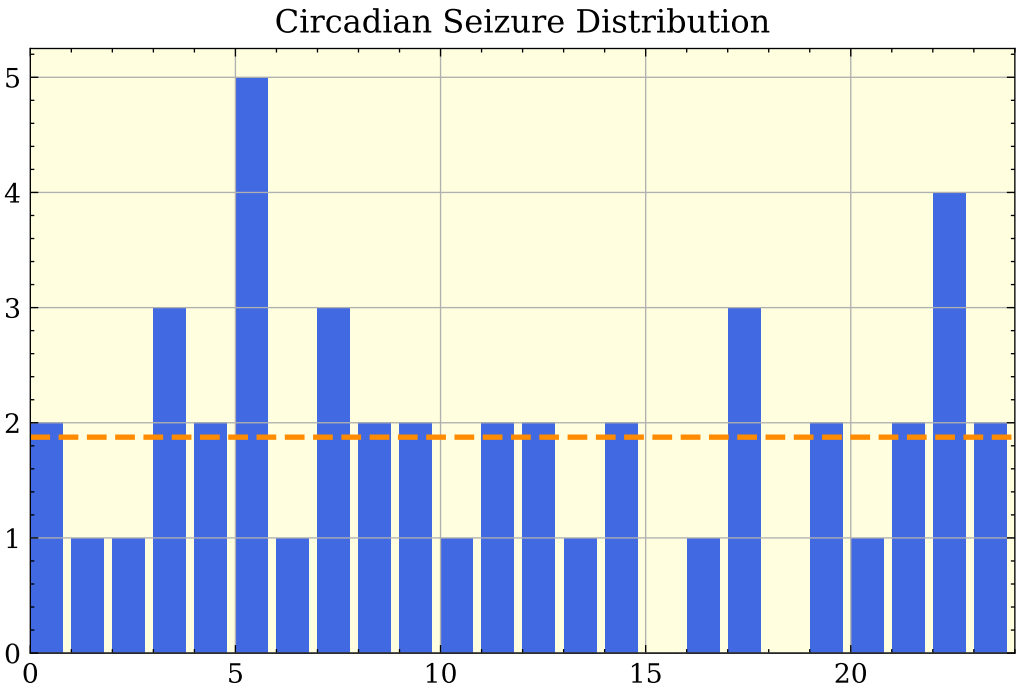
The von Mises distribution is defined as:

Figure 3.6. Circadian seizure distribution for I004_A0003_D001
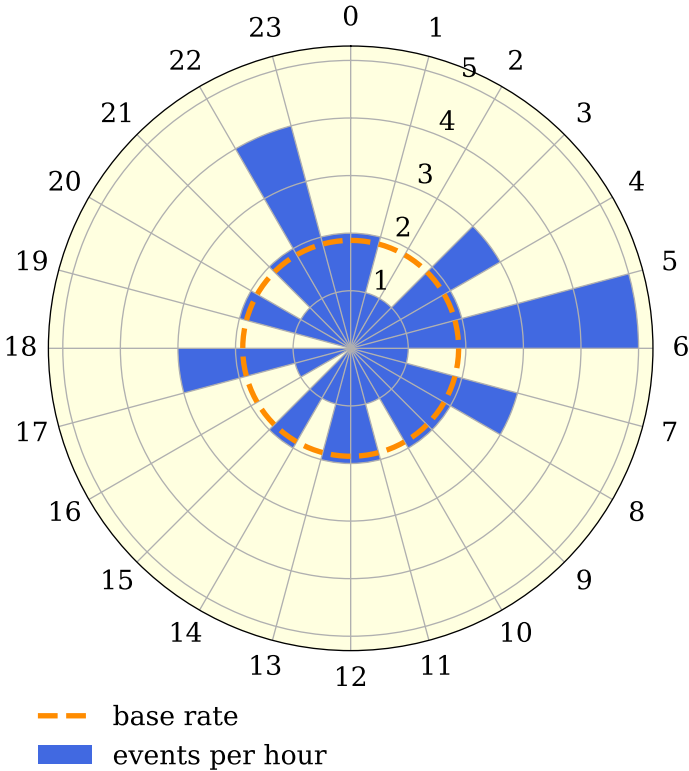

Figure 3.7. Circadian seizure distribution for I004_A0003_D001

$$f(x|\mu, \kappa; \omega) = \frac{\exp(\kappa \cos(\omega(x - \mu)))}{2\pi I_0(\kappa)} \tag{3.14}$$

Visually, the resulting distribution is similar to a bell-shaped normal distribution on a circle (see figure 3.8). In the definition above, μ determines the center of the bell, κ the spread, and ω scales the period length. In this work, we set $\omega \leftarrow \frac{2\pi}{24}$ to scale the period to 24-hours, and drop it from the notation for brevity in the following text.
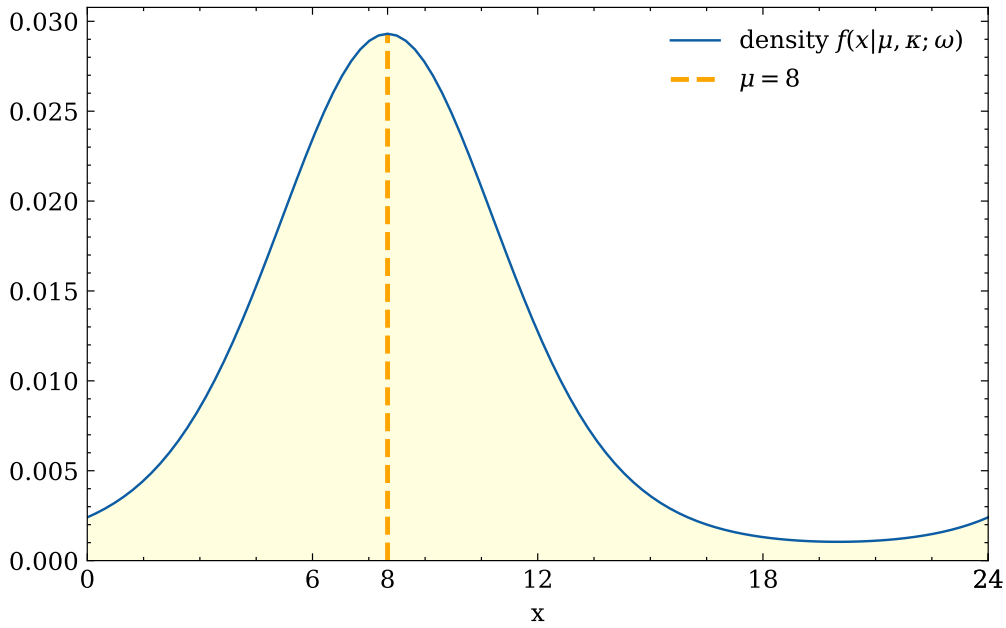
Figure 3.8. | **Von Mises distribution.** The von Mises distribution probability density function (also known as the circular normal distribution) for $\mu = 8$, $k = 1/0.6$, $\omega = \frac{2\pi}{24}$

The *Circadian distribution prior* is the same one used by Karoly et al. [2017], based on a mixture of von Mises distributions (see equation 3.14).

Our work complements the current understanding of the prior's contribution to predictive performance, and has many features that distinguish it from prior work.

NS: write list of ways this work differs from prior works

3.3.2. The Cox process prior

NS: write about Cox process (a.k.a. Inhomogeneous Poisson processes)

NS: make figure of inferring latent intensity

Figure 3.9. non-normalized v.M. prior overlaid on empirical circadian seizure distribution for I004_A0003_D001

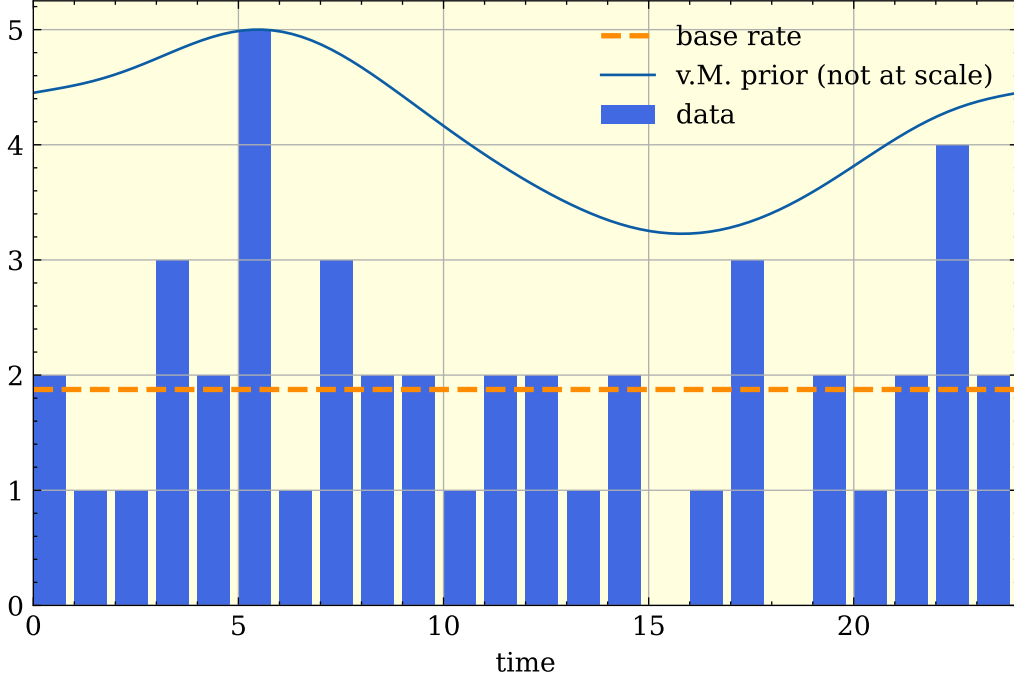### 3.3.3. Probabilistic model

We capture the notion of EEG observations and seizure events by random variables (r.v.s):

$$e_t \sim E(t) \in \Omega_E = \mathbb{R}^{c \times \tau} \tag{3.15}$$

$$s_t \sim S(t) \in \Omega_S = \{0, 1\} \tag{3.16}$$

where $\Omega_{E/S}$ is the sample space for each variable E and S, respectively. c is the number of EEG-channels, τ is the duration of EEG recorded (a.k.a. segment length). We define the event that a seizure occurred at time t as $\{S(t) = 1\}$.

Our work introduces a multilevel probabilistic model for dataset annotations (see figure 3.10). The model is inspired by the hypothesis that clinical annotations are very precise but not highly sensitive. That is, we believe that a proportion of seizures don't become recorded annotations, and explicitly take this belief into account in our model. These can often be explained by the annotator's reliance on video-footage and a possible lack of clinical symptom manifestations.

NS

add citation

The model describes the process by which seizure annotations are generated. The W, k parameters control the shape of the individual's circadian profile. $\lambda(t)$ is the latent intensity function for the seizure-occurrence Cox process. After sampling a seizure-history $S(t)$, each seizure is dropped with a probability of p, and becomes an observable annotation with probability $1 - p$. This reflects annotators' missing sections of the EEG recordings, as well as sub-clinical seizures.

The diagram shows a plate model with the following equations:

$$\phi_i \sim \mathcal{D}_{\phi_i} \qquad \text{// subject-specific params}$$
$$\lambda_{\phi_i}(t) \leftarrow \text{prior}_{\phi_i} : \mathbb{R}^+ \rightarrow \mathbb{R}^+ \qquad \text{// intensity function}$$
$$s_t \sim \text{Cox}(\lambda_{\phi_i}) \qquad \text{// stochastic counting process}$$
$$e_t \mid s_t \sim \mathcal{D}_{E|S} \qquad \text{// EEG conditioned on seizure var}$$
$$o_t \sim \text{Bernoulli}(p) \qquad \text{// was the seizure observed?}$$
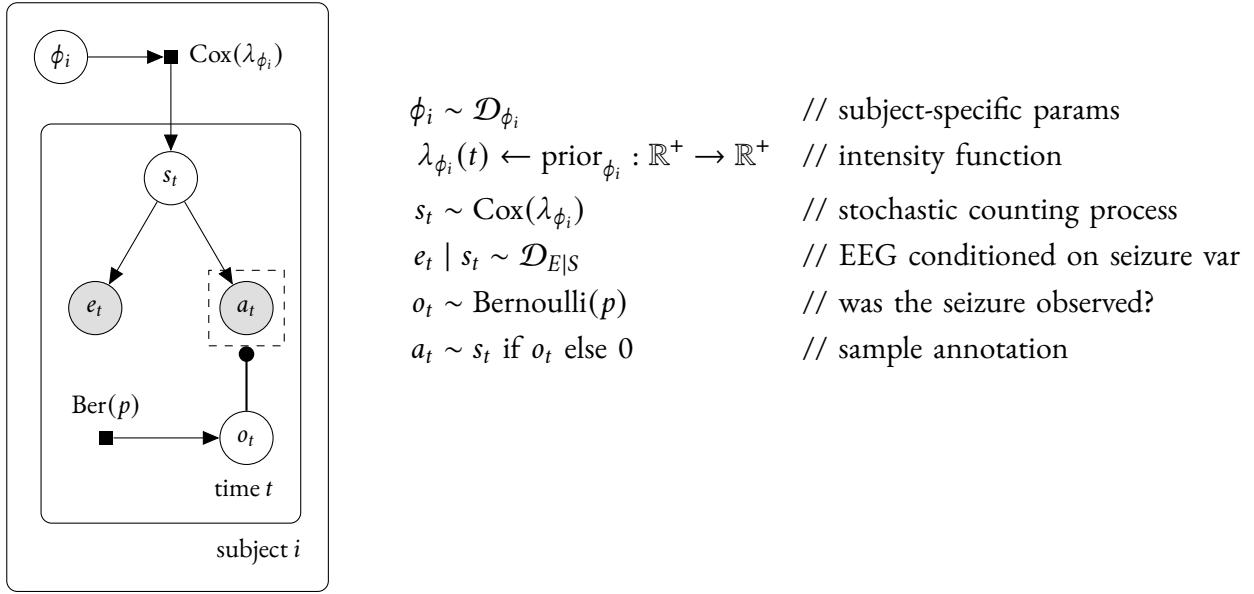$$a_t \sim s_t \text{ if } o_t \text{ else } 0 \qquad \text{// sample annotation}$$

Figure 3.10. Probabilistically Modeling Seizure Occurrence, EEG signal and Annotations

More formally (see algorithm 1), the probability of obtaining an annotation at time t is a zero-inflated seizure occurrence model. Sequentially, the seizure-occurrence model is a Cox process with a stochastic intensity function $\lambda(t)$. In turn, the intensity function $\lambda(t)$ is a mixture model of 24 von-Mises distributions, one for each hour of the day. Finally, the weights W and common spread k parameters assume noninformative priors.

Algorithm 1 Seizure Annotation Model

$k \sim \text{Exp}(1)$
$\alpha[0, ..., 23] \leftarrow 1$
$W[0, ..., 23] \sim \text{Dir}(\alpha)$
$\lambda(t) \leftarrow \sum_{i=0}^{23} W[i] \cdot f_{v.M.}(t; i, k)$
$s_t \sim \text{Cox}(\lambda)$
$a_t = \text{ZeroInflated}(s_t, p)$

Figure 3.11. A generative model for seizure events and annotations with a von-Mises-mixture intensity prior.
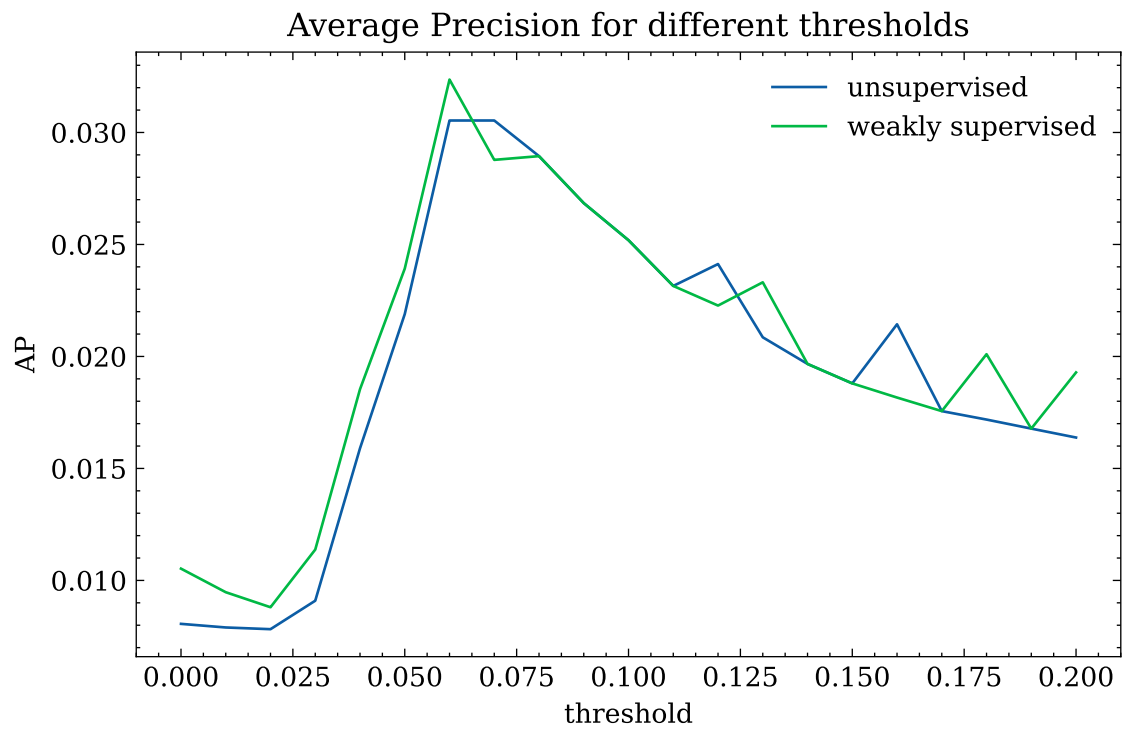
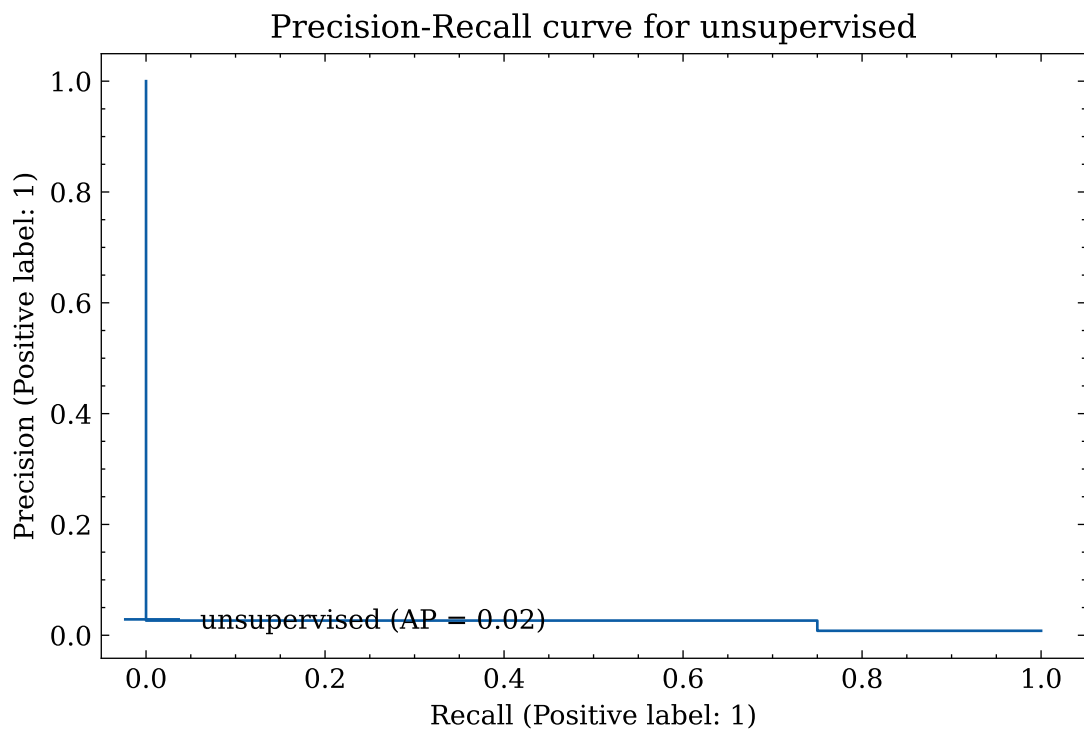**Figure 3.12.** Average Precision scores (higher is better)

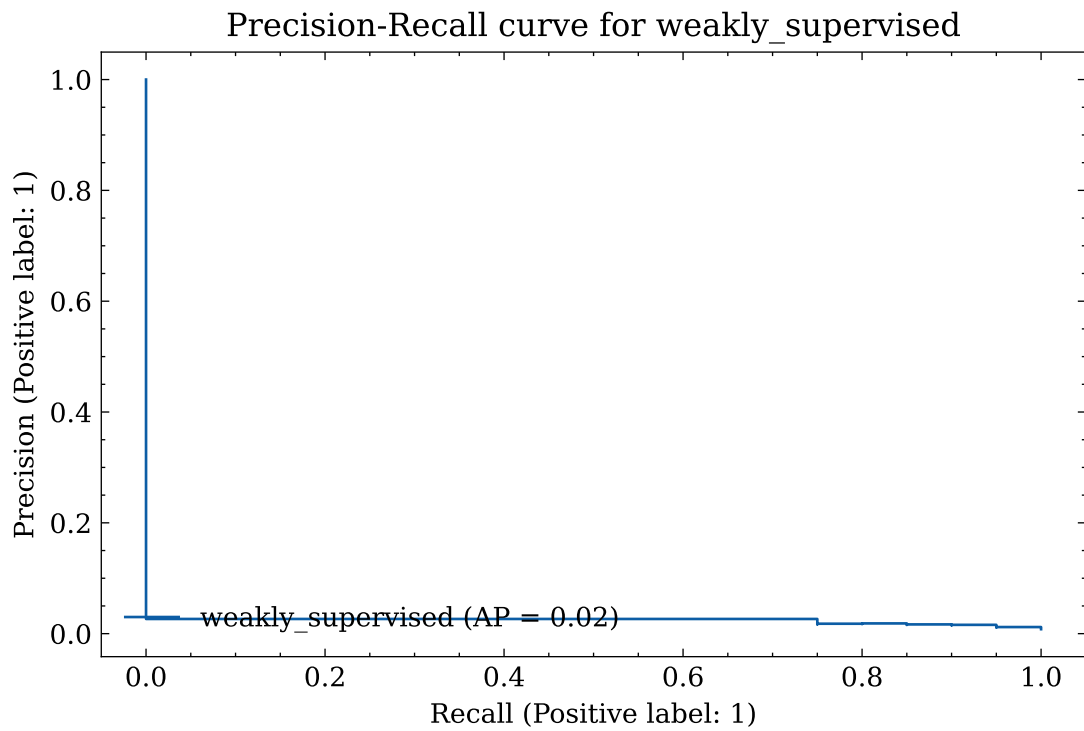**Figure 3.13.** AP for unsupervised

Figure 3.14. AP for weakly supervised (higher is better)
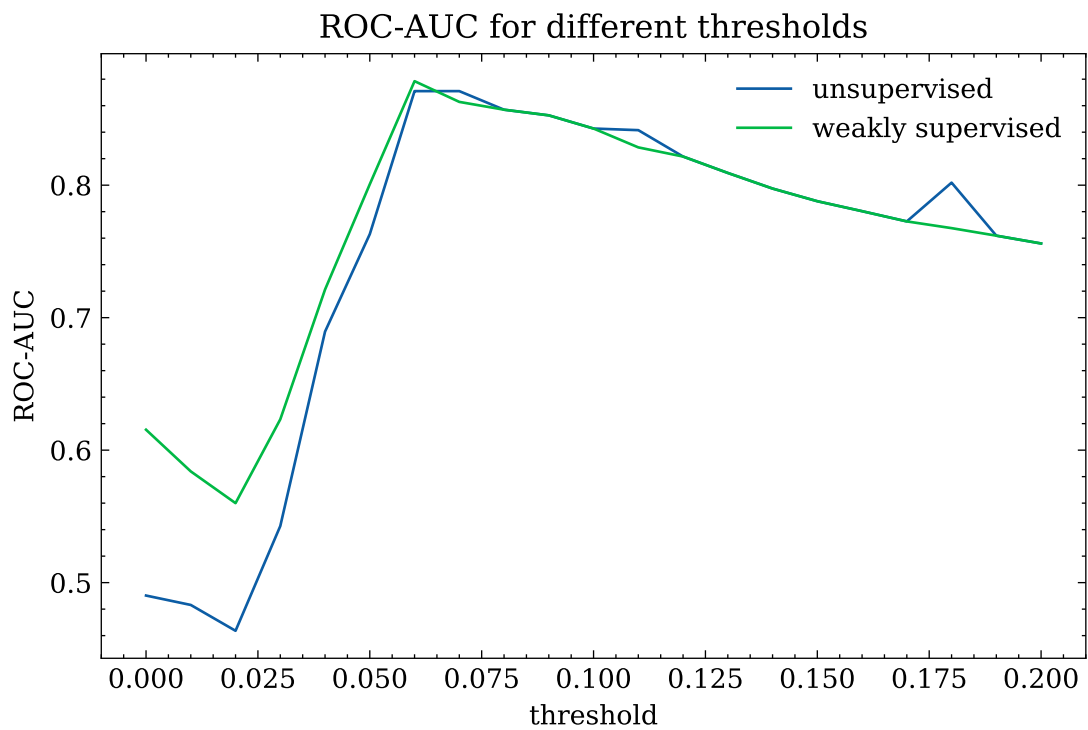

Figure 3.15. ROC-AUC scores (higher is better)
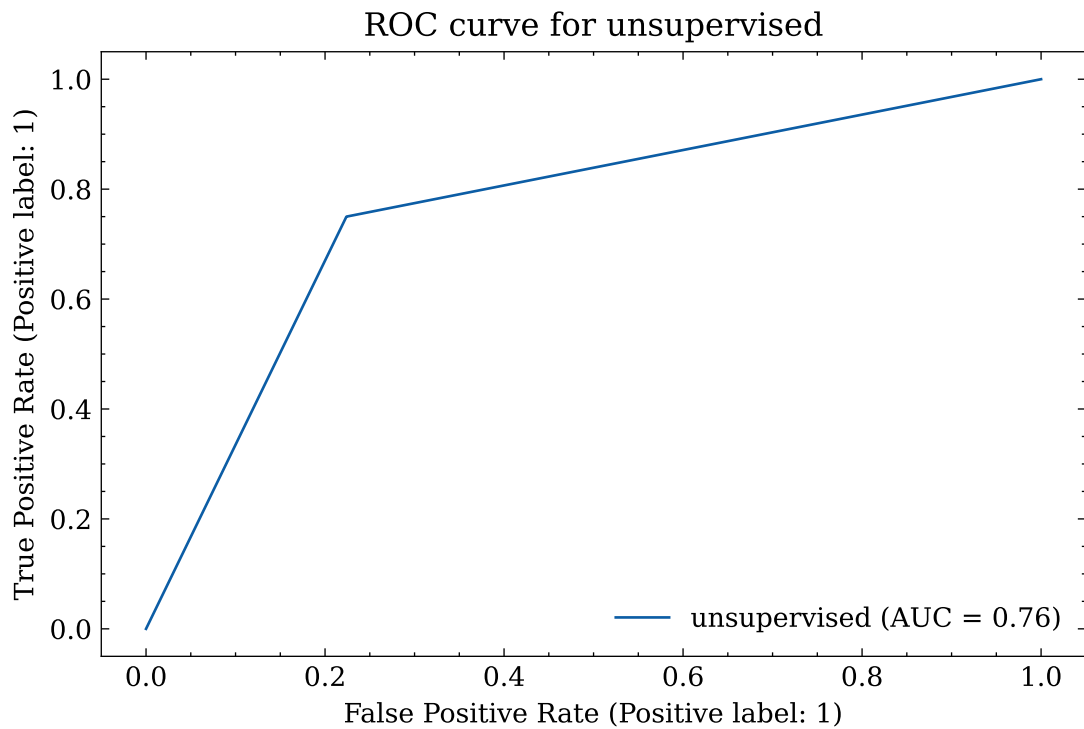
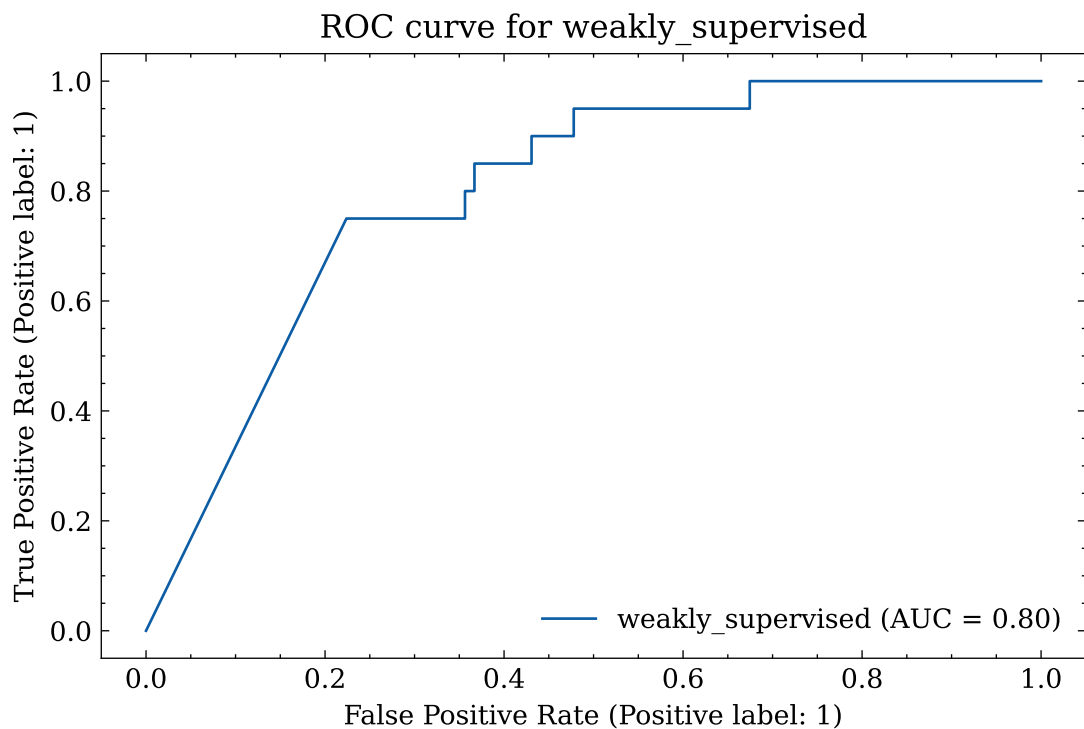**Figure 3.16.** ROC for unsupervised

Figure 3.17. ROC for weakly supervised

NS: check which threshold made the roc curve fig and write it
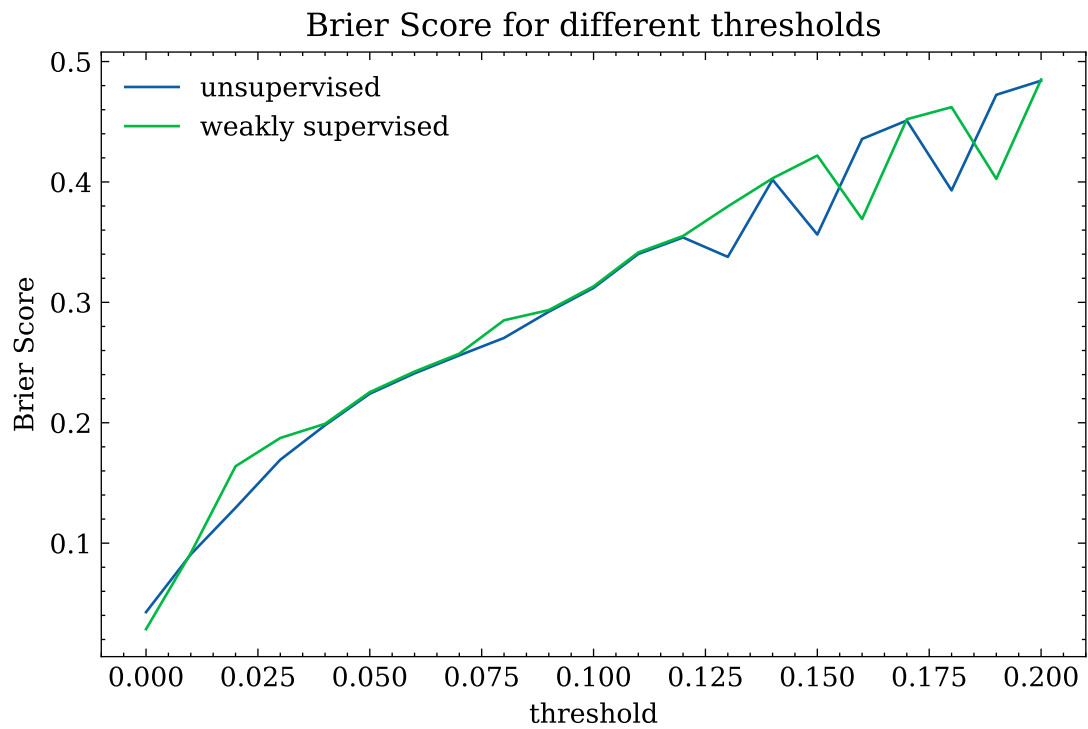
Figure 3.18. Brier scores (lower is better)
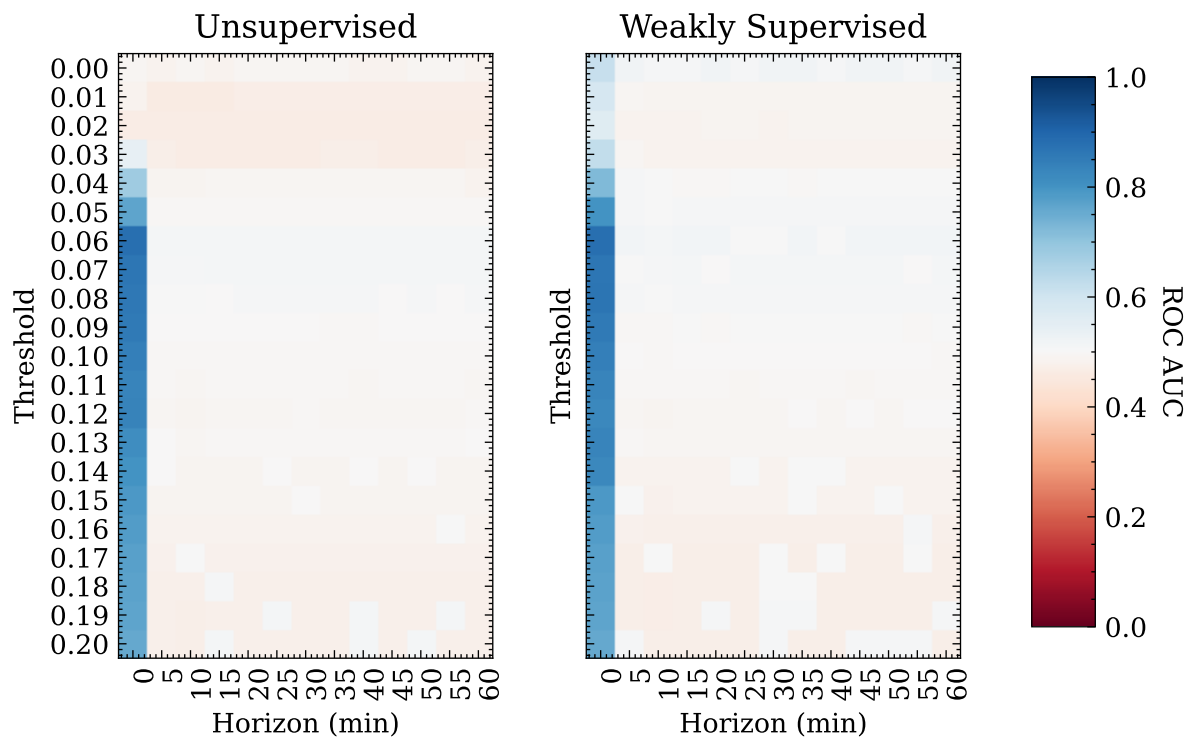
Figure 3.19. AUC-ROC Heatmap for minutes time scale
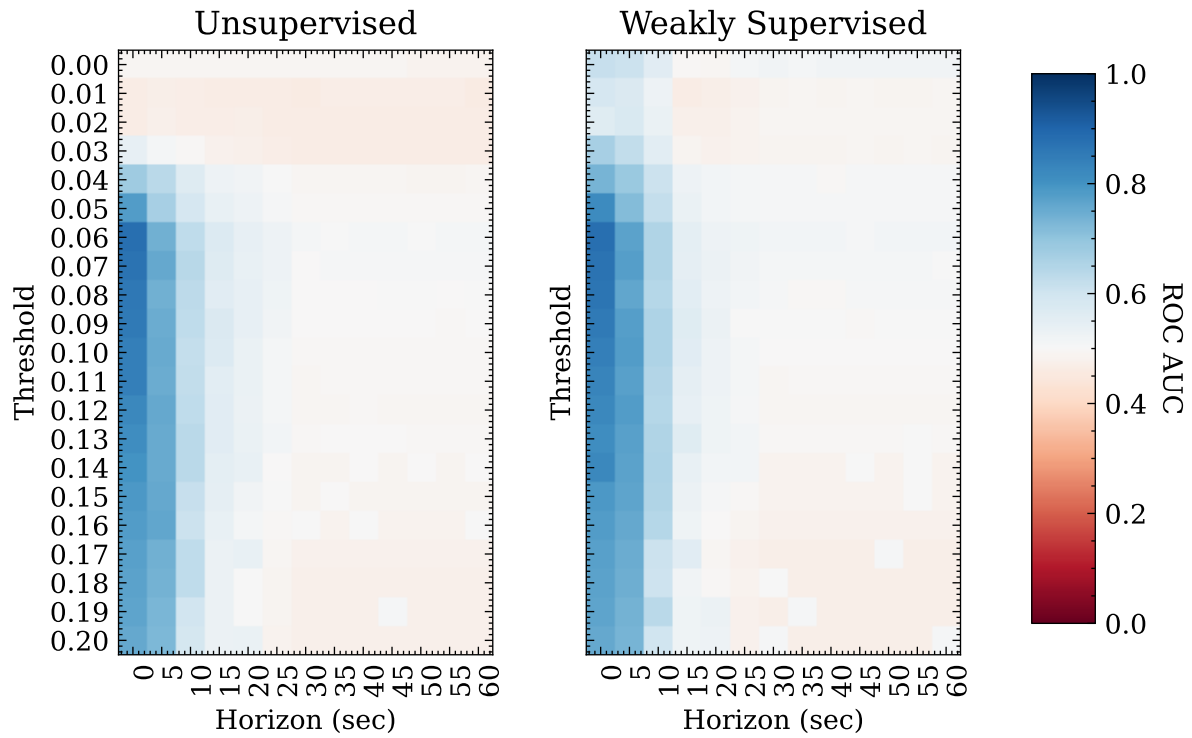
Figure 3.20. AUC-ROC Heatmap for seconds time scale

3.3.4. Empirical Evaluation

Prior

Seizure Likelihood Estimation

Model validation

Linear probing

NS: rewrite Linear probing section

The *support vector machine* (SVM) is a popular method for solving problems in classification, regression and novelty detection [Bishop, 2006]. Fundamentally, it is a two-class linear classifier, i.e., it relies on linear models of the form

$$y(x) = w^T \phi(x) + b \tag{3.17}$$

where $\phi(x)$ denotes a fixed feature-space transformation, and b is the bias parameter. The training data set comprises N input vectors $x_1, ..., x_N$, with corresponding target values $t_1, ..., t_N$, where $t_n \in \{-1, 1\}$, and new data points x are classified according to the sign of $y(x)$.

374   By fitting the SVM to a training set and scoring the predictive accuracy on a hold-out test set, we can
375   quantify the linear separability of the dataset. This is used as a proxy for representation quality.

4. Conclusion

Historically, automatic seizure prediction based on EEG data has been researched since the 1970's, and continues to be an active field of research today [Mormann et al., 2007, Kuhlmann, Lehnertz, Richardson, Schelter, and Zaveri, 2018]. In the course of time, an attempt to characterize a *pre-ictal state* - a short-term (think minutes) pattern in neural activity, which upon manifestation is deterministically followed by a seizure - was dominantly pursued.

The notable methods reported in [Mirowski et al., 2009] influenced the field of seizure risk gauging. For example, 5 years later, the same correlation features were found in each of the three top-scoring submissions to the Kaggle seizure detection competition [**?**].

To this date, most published methods require labeled data, and suffer from high false alarm rates. Since they are fully-supervised, the cost dependence on expert-level domain knowledge makes these algorithms unsuitable for real-world forecasting devices, since unbiased expert labeling solutions are still not available . Furthermore, even clinical-grade annotations, considered to be the "gold standard" in seizure documentation, suffer from observation bias.

To overcome this problem, we propose a weakly-supervised Bayesian framework for likelihood estimation of epileptic seizures. We introduce a multilevel probabilistic model of seizure occurrences and clinical annotations (dataset labels). The model takes into account the possibility of missed seizures unrecorded by the annotator. We implemented an algorithm to compute the inference, namely Bayesian Seizure Likelihood Estimator (BSLE). Lastly, we validate the model and examine it's applicability to a practical seizure warning system.

4.1. Research questions

This thesis addresses the following research questions:

1. To what degree are various classifiers able to discriminate between interictal and ictal EEG segments, when using synchronicity-measures as features? (cf. chapter 2)

2. How can we model seizure likelihood estimation as a Bayesian inference problem, and how well does this model forecast seizures over different time-horizons? (cf. chapter 3)

4.2. Evaluation plan

Proper examination of seizure timing algorithms must account for all types of classification errors and be considerate of the imbalanced nature of the data. It is also essential to be able to benchmark the new

405   algorithm against existing methods. Perhaps above all, it is worthy to evaluate the algorithm's applicability to
406   real-world scenarios. Therefore, we will report a set of standard evaluation metrics for both the deterministic
407   and the probabilistic settings, and attempt to replicate the setting proposed in [Karoly et al., 2017] as much
408   as possible.

409   In evaluating the anomaly detection method proposed in this work, we will follow the guidelines provided
410   by the Kaggle Seizure Detection Challenge [UPenn and Mayo Clinic and Kaggle, 2014]. Namely, the model
411   will receive as input the training data from the Dog 1 dataset, and output class probabilities (0 for interictal,
412   1 for ictal). We will report ROC-curves and ROC-AUC on the held-out test set.

413   NS: report Brier Score

414   NS: report Snyder 2008

5. Discussion

Seizures are states of abnormal brain function, which are associated with unwanted symptoms such as involuntary muscle movements (e.g., limb jerking, convulsing), partial or complete loss of consciousness and memory relapses. Epilepsy is a condition in which unprovoked seizures occur recurrently. People suffering from epilepsy may also suffer from side effects such as stress, fear, and sleep disorder. Prolonged seizures are especially dangerous and can lead to sudden death. Moreover, the uncertainty regarding seizure occurrence is regarded as one of the most disturbing factors of the disease.

5.1. Model validation

NS: write about model validation

NS: rewrite discussion. Points to include: (1) future work on hierarchical patient modeling

The problem of automatic seizure detection is challenging, inducing many attempts over the years. In this work we attempted a probabilistic approach, relying on Bayes' rule to estimate the likelihood of a seizure given an EEG segment. We assumed that seizures are rare, which led us to a novelty-score-based likelihood. Following recent findings, we also assumed that seizures are approximately cyclical, taking this into account in our prior.

Although the method we proposed works well, there are some drawbacks. First, the channel spatio-temporal location is disregarded, and the pairing (in the double-channel case) was made arbitrarily. Further work should utilize better channel selection and modeling the topographic qualities of the channels for potentially improved results.

Second, the model of normal EEG was fit using the segments from the Canine-epilepsy-dataset which were chosen for the Kaggle challenge. The class distribution differs from the true class distribution. To combat this discrepancy, we dropped the ictal segments and used only the interictal segments for the model of normality. Because ictal EEG is extremely less common than interictal EEG, we assume that using the interictal segments is a sufficiently close approximation of the natural distribution. Further studies should recalibrate the method based on the raw recordings in order to provide an even better estimate.

Thirdly, from the clinical perspective, the evaluation method is limited since the data originates from canines with naturally occurring epilepsy, instead of humans. Dogs with naturally occurring epilepsy show similar semiology to epilepsy in humans, but more work is required to validate this method on human EEG.

In summary, we present an anomaly detection method based on Gaussian processes embeddings, and evaluate it on a seizure detection task. The method is significant because it is unsupervised, thus eliminated the need for costly annotators.

446   We hope that future attempts at seizure likelihood estimation will utilize our Bayesian by incorporating

447   new likelihood functions and priors.

Bibliography

Melinda S Kelley, Margaret P Jacobs, and Daniel H Lowenstein. The ninds epilepsy research benchmarks. *Epilepsia*, 50(3):579, 2009.

Sonya B Dumanis, Jaqueline A French, Christophe Bernard, Gregory A Worrell, and Brandy E Fureman. Seizure forecasting from idea to reality. outcomes of the my seizure gauge epilepsy innovation institute workshop. *Eneuro*, 4(6), 2017.

Piotr Mirowski, Deepak Madhavan, Yann LeCun, and Ruben Kuzniecky. Classification of patterns of eeg synchronization for seizure prediction. *Clinical neurophysiology*, 120(11):1927–1940, 2009.

Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, and Sarah Mack. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

Ingrid E Scheffer, Samuel Berkovic, Giuseppe Capovilla, Mary B Connolly, Jacqueline French, Laura Guilhoto, Edouard Hirsch, Satish Jain, Gary W Mathern, Solomon L Moshé, et al. Ilae classification of the epilepsies: position paper of the ilae commission for classification and terminology. *Epilepsia*, 58(4): 512–521, 2017.

Palak Handa, Monika Mathur, and Nidhi Goel. Open and free eeg datasets for epilepsy diagnosis. *arXiv preprint arXiv:2108.01030*, 2021.

Matthias Ihle, Hinnerk Feldwisch-Drentrup, César A Teixeira, Adrien Witon, Björn Schelter, Jens Timmer, and Andreas Schulze-Bonhage. Epilepsiae–a european epilepsy database. *Computer methods and programs in biomedicine*, 106(3):127–138, 2012.

J-B Schiratti, Jean-Eudes Le Douget, Michel Le van Quyen, Slim Essid, and Alexandre Gramfort. An ensemble learning approach to detect epileptic seizures from long intracranial eeg recordings. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 856–860. IEEE, 2018.

Klaus Lehnertz, Christian Geier, Thorsten Rings, and Kirsten Stahn. Capturing time-varying brain dynamics. *EPJ Nonlinear Biomedical Physics*, 5:2, 2017.

Michael G Rosenblum, Arkady S Pikovsky, and Jürgen Kurths. From phase to lag synchronization in coupled chaotic oscillators. *Physical Review Letters*, 78(22):4193, 1997.

Jean-Philippe Lachaux, Eugenio Rodriguez, Jacques Martinerie, and Francisco J Varela. Measuring phase synchrony in brain signals. *Human brain mapping*, 8(4):194–208, 1999.

[477] Florian Mormann, Ralph G Andrzejak, Christian E Elger, and Klaus Lehnertz. Seizure prediction: the long and winding road. *Brain*, 130(2):314–333, 2007.

[479] Piotr W Mirowski, Yann LeCun, Deepak Madhavan, and Ruben Kuzniecky. Comparing svm and convolutional networks for epileptic seizure prediction from intracranial eeg. In *2008 IEEE workshop on machine learning for signal processing*, pages 244–249. IEEE, 2008.

[482] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[487] Philippa J Karoly, Hoameng Ung, David B Grayden, Levin Kuhlmann, Kent Leyde, Mark J Cook, and Dean R Freestone. The circadian profile of epilepsy improves seizure forecasting. *Brain*, 140(8):2169–2182, 2017.

[490] Maxime O Baud, Jonathan K Kleen, Emily A Mirro, Jason C Andrechak, David King-Stephens, Edward F Chang, and Vikram R Rao. Multi-day rhythms modulate seizure risk in epilepsy. *Nature communications*, 9(1):1–10, 2018.

[493] Maxime O Baud, Timothée Proix, Vikram R Rao, and Kaspar Schindler. Chance and risk in epilepsy. *Current opinion in neurology*, 33(2):163–172, 2020.

[495] CE Rasmussen. Cki williams gaussian processes for machine learning, 2006.

[496] Sergios Theodoridis. *Machine learning: a Bayesian and optimization perspective*. Academic press, 2015.

[497] Kathryn A Davis, Beverly K Sturges, Charles H Vite, Vanessa Ruedebusch, Gregory Worrell, Andrew B Gardner, Kent Leyde, W Douglas Sheffield, and Brian Litt. A novel implanted device to wirelessly record and analyze continuous intracranial canine eeg. *Epilepsy research*, 96(1-2):116–122, 2011.

[500] UPenn and Mayo Clinic and Kaggle. seizure detection & prediction challenges, 2014. https://www.kaggle.com/c/seizure-detection and https://www.kaggle.com/c/seizure-prediction.

[503] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[505] Sharon Bertsch McGrayne. The theory that would not die. In *The Theory That Would Not Die*. Yale University Press, 2011.

[507] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[509] Levin Kuhlmann, Klaus Lehnertz, Mark P Richardson, Björn Schelter, and Hitten P Zaveri. Seizure prediction—ready for a new era. *Nature Reviews Neurology*, 14(10):618–630, 2018.

[511] Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. Gpy-
[512] torch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural*
[513] *Information Processing Systems*, 2018.

[514] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 3 2019. URL https://github.
[515] com/PyTorchLightning/pytorch-lightning.

Appendices

A. GP embedding: training details

For dimensionality reduction of an observed EEG segment, we performed exact inference of the GP parameters maximizing the observation likelihood, using the GPyTorch and PyTorch Lightning frameworks [Gardner et al., 2018, Falcon and The PyTorch Lightning team, 2019].

More formally, in fitting the Gaussian processes to the EEG samples we carried out exact Type-II Maximum Likelihood Estimation for each sample. This means optimizing the model's hyperparameters (mean module, covariance module, etc.) w.r.t maximization of the *marginal log likelihood* (MLL) of the given data $\mathbf{E}, \mathbf{t}$:

$$\theta \leftarrow \operatorname*{argmax}_{\theta} p_f(\mathbf{E} \mid \mathbf{t}) = \int p(\mathbf{E} \mid f(\mathbf{t})) p(f(\mathbf{t}) \mid \mathbf{t}) df \qquad (A.1)$$

where $f \sim \mathcal{GP}(\mu, K)$ is the modeled signal before adding the homescedastic Gaussian noise, and θ is the set of parameters to be optimized. See code for implementation details.

GP & inference (training) configuration details		
	Parameter	Value
GP params	mean module	zero mean
	covariance module	scaled Matérn-1.5 kernel
	task covariance rank	1
	number of tasks	2
	likelihood (noise model)	Gaussian (homoscedastic)
Training params	optimizer	Adam
	learning rate	0.01
	max. number of epochs	1000
	patience (early stopping)	8

Table A.1. The GP parameters and training parameters used in our experiments.

תקציר

התקציר שלי

חיזוי הסתברותי של התקפי אפילפסיה

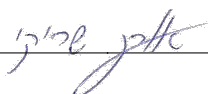מחקר לשם מילוי חלקי של הדרישות לקבלת תואר
"מגיסטר למדעים"

מגיש

נועם סיגל

העבודה נעשתה בהדרכת

ד״ר אורן שריקי
וד״ר דוד טולפין

המחלקה למדעי המחשב

אישור המנחים: _____

אישור להשליׁ:

_____

אב ה׳תשפ״ב

באר שבע

חיזוי הסתברותי של התקפי אפילפסיה

מחקר לשם מילוי חלקי של הדרישות לקבלת תואר
"מגיסטר למדעים"

מגיש

נועם סיגל

הוגש למחלקה למדעי המחשב באוניברסיטת בן גוריון בנגב

אב ה׳תשפ״ב

באר שבע