Ukrainian Catholic University
Faculty of Applied Sciences
Business Analytics

Epilepsy region identification

Signal processing course

Author: Sofiya Garkot

December 2020

1. Introduction

Epilepsy is a neurological disorder characterized by the occurrences of seizures, due to abnormal neuronal activity in the brain. The disease causes lots of negative consequences. The unpredicted seizures provoke risk of falls and injuries, psychiatric disturbances, cognitive deficits and difficulties in achieving academic, social and employment goals. Prediction of the seizures for every patient can significantly improve the quality of life of sick people.

The electrical activity during seizures can be measured using EEG by monitoring electrical signals on the scalp. The aim of this study is to provide the analysis of EEG data for better understanding of the processes within a brain as well as to predict the start of the seizure onset using the characteristics of the brain's state before the seizure. The second task of choice was the prediction of the region of the epileptic seizure on EEG. Qualified doctors spend a lot of time analysing hours of EEG per patient, which can be optimized by predicting the region of epileptic seizure and result in faster seizure localization.

2. Data description

The data are collected at Children's Hospital Boston. It consists of 23 cases - 22 children (5 males, ages 3–22; and 17 females, ages 1.5–19). Each case contains between 9 and 42 continuous .edf files from a single subject. In most cases, the .edf files contain exactly one hour of digitized EEG signals[2].

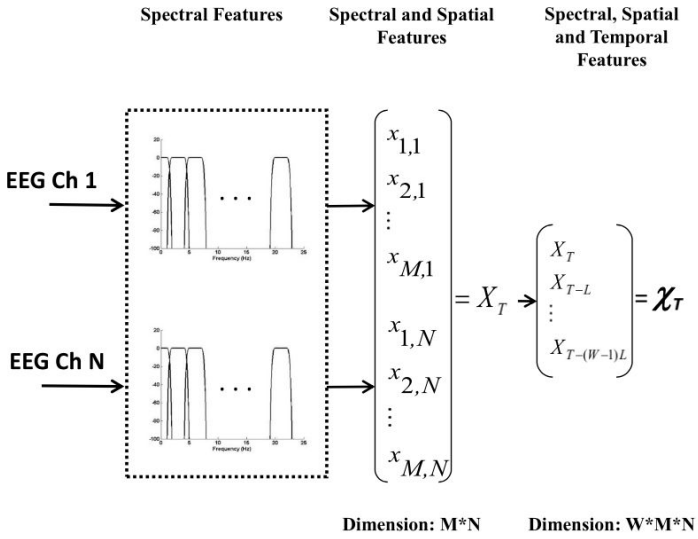
All signals were sampled at 256 samples per second. Most files contain 23 EEG signals. The International 10-20 system of EEG electrode positions and nomenclature was used for these recordings.

Variety of channels simultaneously record the information about the state of a brain: recordings of some channels change their behaviour during seizure, others do not. EEG recording includes the info not only about the 'pure' electrical activity of a brain. Variety of other artifacts are present during EEG recordings (like eye movements, muscular or heart activity). The data recorded represent a non-stationary process because in every case there are epilepsy and non-epilepsy regions.

3. Existing approaches

The first attempt to work on the same dataset was the article "Application of Machine Learning To Epileptic Seizure Detection" by Ali Shoeb and John Guttag[3]. The authors

proposed patient-specific onset detection through analysis of EEG and application of ML approaches to the task. The data from a variety of EEG channels were transformed into feature vectors characterising a patient state at one period of time. Later, the authors aimed at constructing a function f(X) that maps a feature vector X onto the labels ±1 depending on the presence of seizure. The features derived were divided into spectral and spatial groups. The size of the time period during which the features were derived equals 2 seconds. The number of channels taken into consideration equals to 18 (however the authors did not provide the information or code in order to analyse which channels were chosen and why).
Due to the duration of seizure onset (~6 seconds) the authors continuously concatenated 3 feature vectors into a single matrix and based on that representation classified the 6 seconds time period as seizure or non-seizure.
The algorithm used is the Support Vector Machine. The training was performed on the first 20 seconds of 4 seizures per patient as well as 24 hours of non-seizure EEG.
The accuracy of the proposed method reached 96% from the 173 test seizures.



The illustration of the feature derivation used in the corresponding study.

The next approach reviewed was the study 'Different permutation entropy patterns of electroencephalogram recorded during epileptiform activity' by Oleksii Avilov and Anton Popov[5]. The data used in the study was the same as in the upper. The proposed approach is based on the permutation entropy (PE) characteristic of the seizures in time domain: during ictal periods the PE becomes less because the complexity of EEG is lower. Due to that property of the characteristic in time PE was chosen as a feature for further study.

The next approach is Epileptic EEG Classification Using Synchrosqueezing Transform and Machine Learning[12]. The data[2] used in the study intersect with those used in upper approaches, which makes it comparable. The patient-based seizure detection approach is proposed using a high-resolution time-frequency (TF) representation named Synchrosqueezed Transform (SST) method. Using machine learning methods such as Decision Tree (DT), k-Nearest Neighbor (kNN), and Logistic Regression (LR), classification is conducted.

Time-frequency representation of pre-seizure (or inter-seizure) and seizure EEG segments are obtained using the SST approach. Non-overlapping, 1second duration EEG segments are obtained from the seizure EEG signals. Using the feature vectors the upper described classifiers were trained in order to predict the region of seizure onset.
The evaluation metric is accuracy. Its value has reached 94.47% and 95.15% for corresponding HOJ-TF and GLCM based approaches.

The other study conducted on this dataset is “Epileptic Seizure Detection Based on EEG Signals and CNN” by Mengni Zhou et al[4]. In this study, a convolutional neural network (CNN) was trained based on raw EEG signals instead of manual feature extraction. The data was used to distinguish ictal, preictal, and interictal segments for epileptic seizure detection.

The algorithm is well described using this visualization:

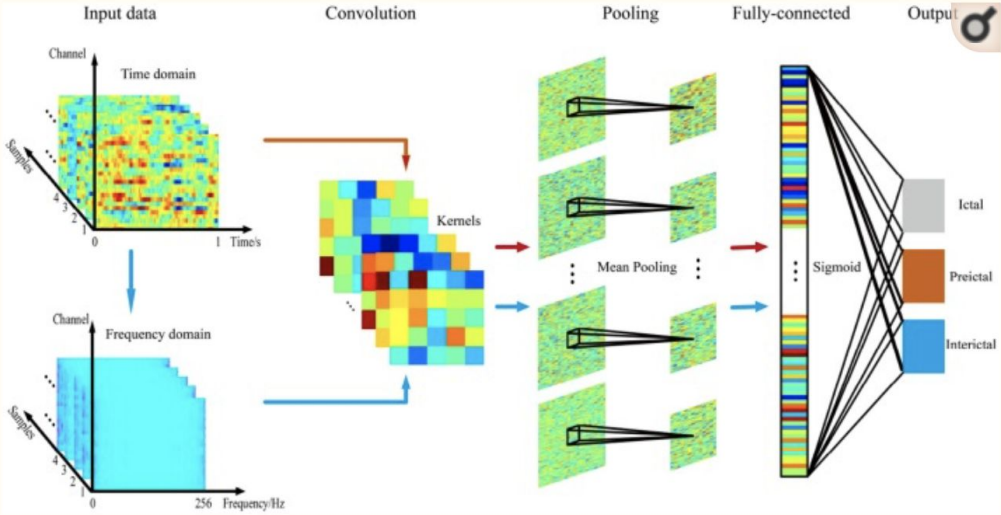


Figure 2

Illustration of the CNN.

The illustration of the algorithm used in the study.

The results were measured using the following measures: accuracy, sensitivity and specificity. For comparison reasons the accuracy of such an approach ranges from 0.678 to 0.986 varying from patient to patient. The average accuracy was 95.6%.

Another approach by Tzallas et al.[13] is based on Seizure Detection in EEGs Using Time–Frequency Analysis. The authors generate the features in the time-frequency domain using windows (through time segments as well as certain frequency bands). The calculated features represent the energy during chosen time periods. Then those features are fed into feedforward ANN with the depth of NN equaling 5 layers. In addition the authors tested a variety of other classifiers (naïve Bayes classifier, decision trees, k-nearest-neighbors (k-NNs) and logistic regression).
The authors concentrated on 3 different classification tasks:
1) seizure vs. non-seizure prediction;
2) seizure, normal, and non-seizure classification;
3) five classes distinguishing based on patient state and movements during the data collection for the study.

The most similar to mine is the first classification task. The results vary from approach to approach, starting from 84.8%–89% to 99.8%.

The last three overviewed approaches are discussed in the study by Baldassano[1]. The author describes three approaches that reached the highest accuracy on the organized kaggle challenge.

The first and best-performing algorithm was developed by Michael Hills. Three types of features were generated:

1) Pairwise cross-correlation between channel signals as well as the sorted eigenvalues of the cross-correlation matrix
2) The frequency magnitudes of each channel in range from 1 to 47 Hz.
3) Pairwise cross-correlation between normalized channel power spectra in the range of 1 to 47 Hz and the sorted eigenvalues of the cross-correlation matrix

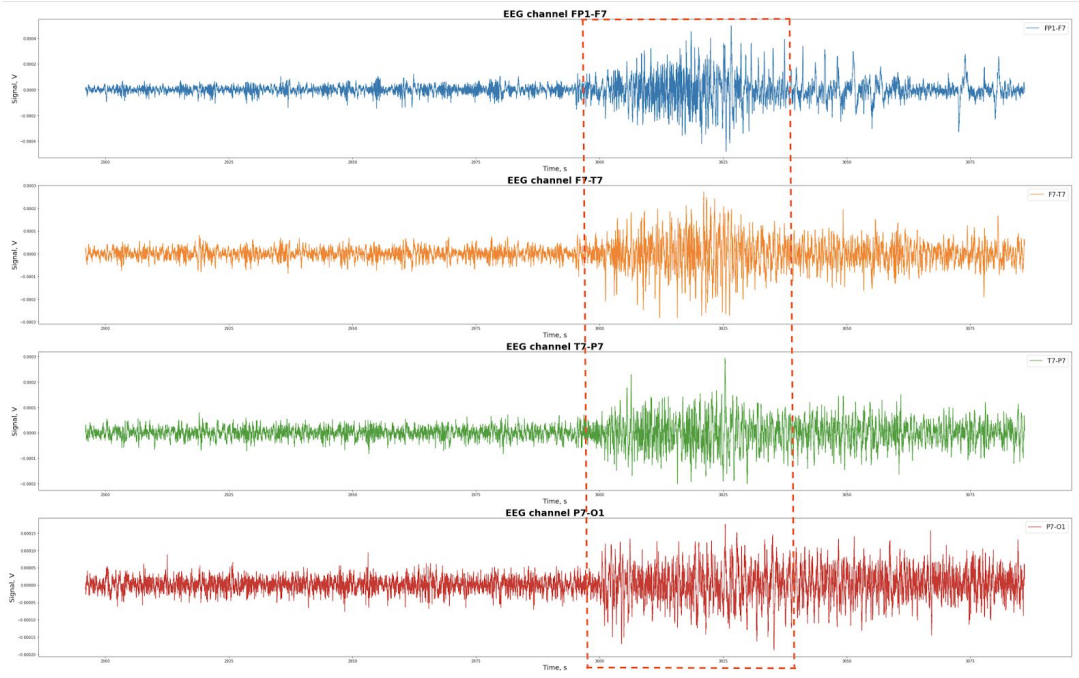
The random forest classifier of 3000 trees is trained on a complete feature set. The evaluation metric was the area under the ROC curve (AUC) for two prediction tasks -- prediction of the region of seizure as well as prediction of the start of the seizure. I compare the methods based on the first one, because it is more similar to the task I solved in my project. For this method the AUC of prediction of seizure onset equals 0.981.

The second-best algorithm was developed by Eben Olson and Damian Mingle. The combination of a variety of filters covering the range 5–200Hz was used and 3 best-performing combinations were retained. The covariance matrices calculated after filtration was used as the feature set. The classification is made using an ensemble of 100 multi-layered neural networks. The AUC for this method equals 0.976.

The third place algorithm was developed by Ishan Talukdar, Nathan Moore, and Alexander Sood. The derived features are channel-specific (maximum amplitude, mean amplitude, absolute deviation, and variance of every channel as well as variety of features used from FFT - maximum power, mean power, variance, and frequency at which the maximum power occurs). Classification is made by averaging the outputs of 1000 decision trees using the Extremely Randomized Trees algorithm.
This method reached the AUC of 0.984.

4. Methodology

The data were used to solve two kinds of tasks: the first one lies in prediction of the start of a seizure depending on signal characteristics before the event. The other aspect of the project was to predict a labelled vector that will map onto the signal detecting the time period during which a seizure was observed.

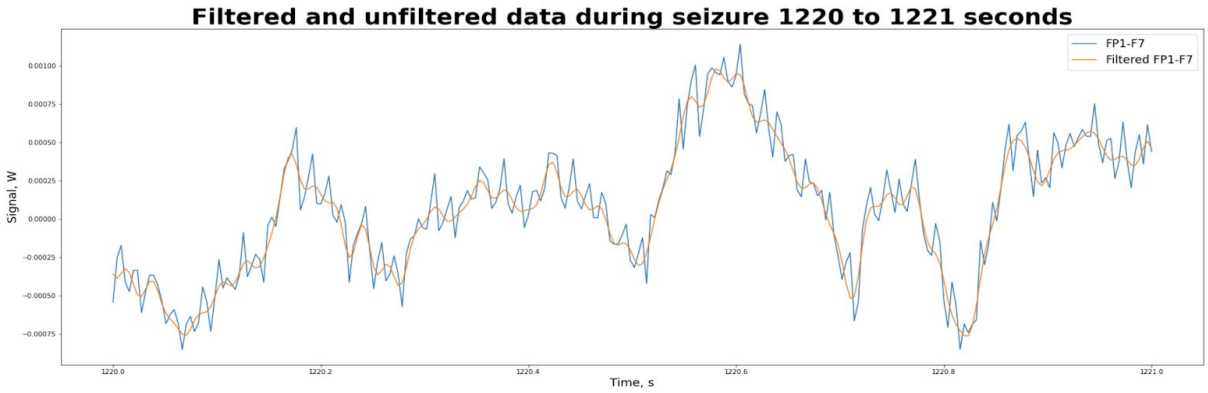


The EEG data plotted from a variety of channels. The red frame identifies the seizure.

4.1 Preprocessing

4.1.1 Filtration

The pass band of choice was in range from 0.5 Hz to 40 Hz[7].



The 1-second signal observation from channel FP1-F7. Blue → unfiltered signal, orange → filtered signal.

4.1.2 Independent Component Analysis

ICA was performed in order to detect various artifacts and remove them.
The previous assumption is that our signal (x) (that is recorded by 23 channels) can be represented as a sum of 12 independent components (s) with the corresponding weights (A), such that:

x = As

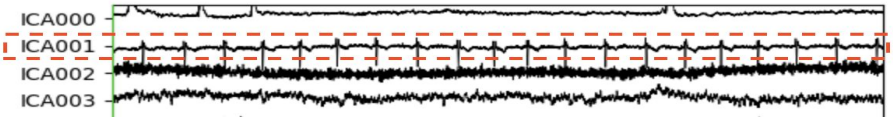
x → 23 x 921600 matrix
A → 23 x 12 weights mixing matrix
s → 12 x 921600 matrix

The independent components are latent variables, meaning that they cannot be directly observed. Also the mixing matrix is assumed to be unknown. All we observe is the random vector **x**, and we must estimate both **A** and **s** using it [10, 11].

According to a variety of studies[9, 10, 11], the optimal choice of number of components if 0.9 multiplied by the number of channels, that in my case would be 24. However due to large computational resources taken for such computations I used only 12 components.

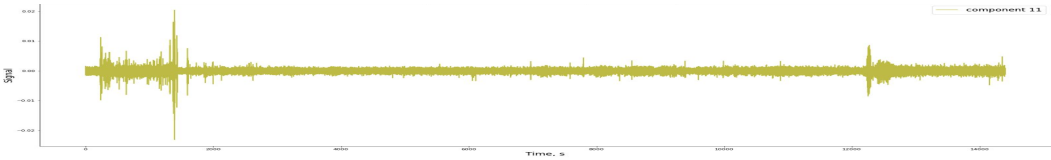
After the components were derived, they were manually listed and analyzed for the likelihood with the usual artifacts: heart beats, muscular movements.

Previously I assumpted that the components would look the following way:



Example of driven independent components.

However several of them look the following way.



Independent component 11 derived from the patient 'chb09'.

Those components are not obvious for classifying as heart beats, that is why I decided to test each of the components for stationarity and remove those that represent stationary processes. The logic behind is that if a component represents a stationary process then it is possibly not bearing the information about the electrical activity of the brain during seizure, thus cannot influence their appearance.

The stationarity test of choice was the Dickrey-Fuller test, but it required too much time for computation for a single component, that is why the further analysis of non-stationary components was not performed.

4.1.3 Permutation entropy

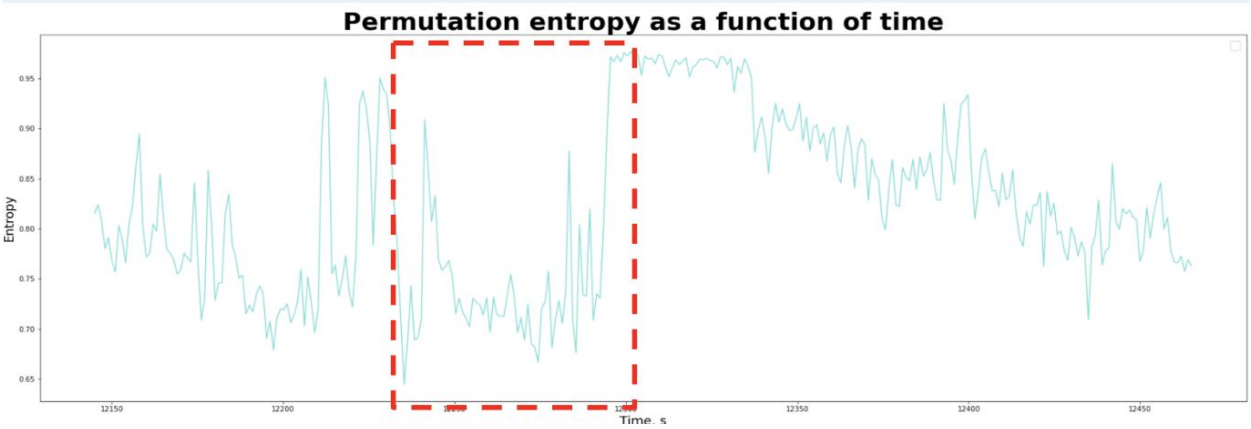
Permutation entropy (PE) is a measure of disorder (randomness) of information contained in comparing the consecutive values of the signal, and it uses the relative frequencies of various patterns encountered in signal samples[5]. This value is the measure of the amount of information contained in comparing m consecutive signal samples over some time interval.

$$PermEn_x(m,l) = -\sum_{j=1}^{m!} p(\pi_j)\log p(\pi_j).$$

A study by Nicolaou[6] has shown that the 'randomness' of the signal decreases during the seizure period.

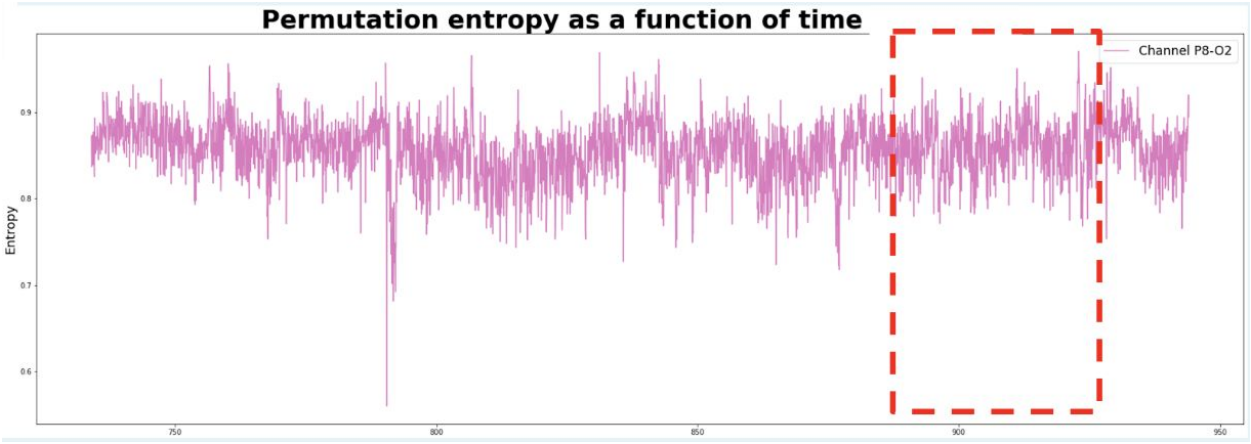
The value of PE depends on lag and order. The lag is the amount of displacement. The order is the number of points between two consecutive points used for measurement of PE. The entropy was measured in a signal for a period of 1 second. Then the vector characterising the signal was derived and plotted.

Firstly PE analysis was applied to raw signal. On average in all channels the PE significantly decreased during the seizure period.



Permutation entropy averaged through the channels. Red - the region of seizure.

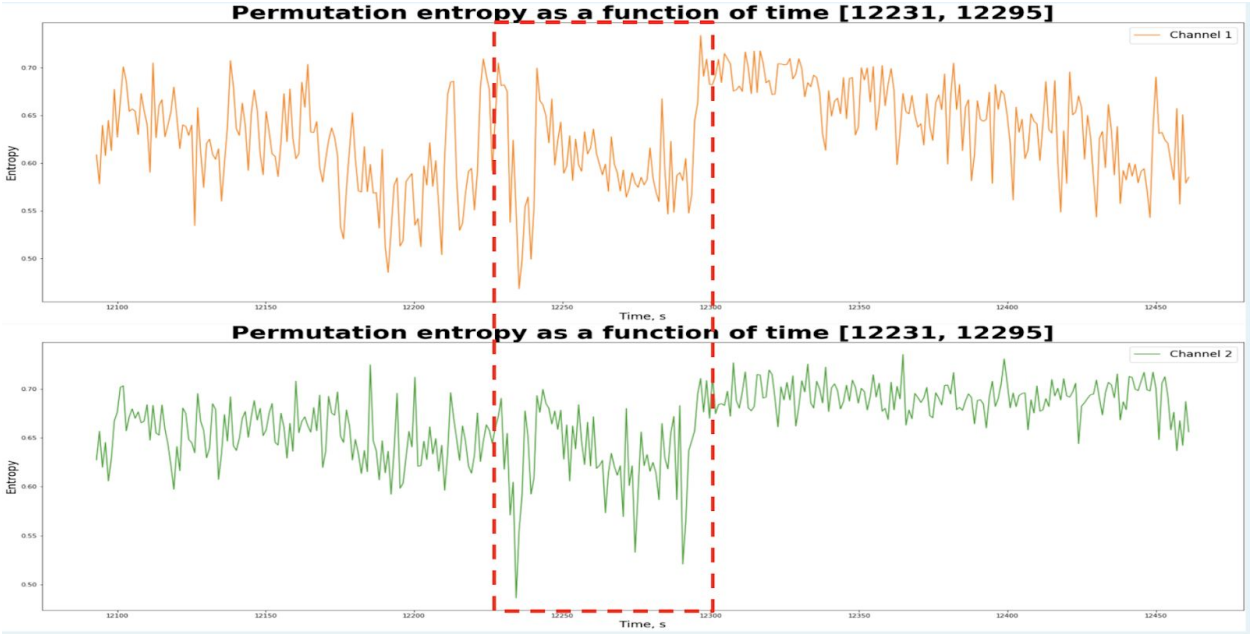
Looking at PE for some channels separately I did not observe the significant PE shift during seizure onset.



Permutation entropy of channel P8-O2 (case of patient chb03) as a function of time. Red - the region of seizure.

The second step was to apply PE analysis to independent components and observe how they change before and during seizure onset.

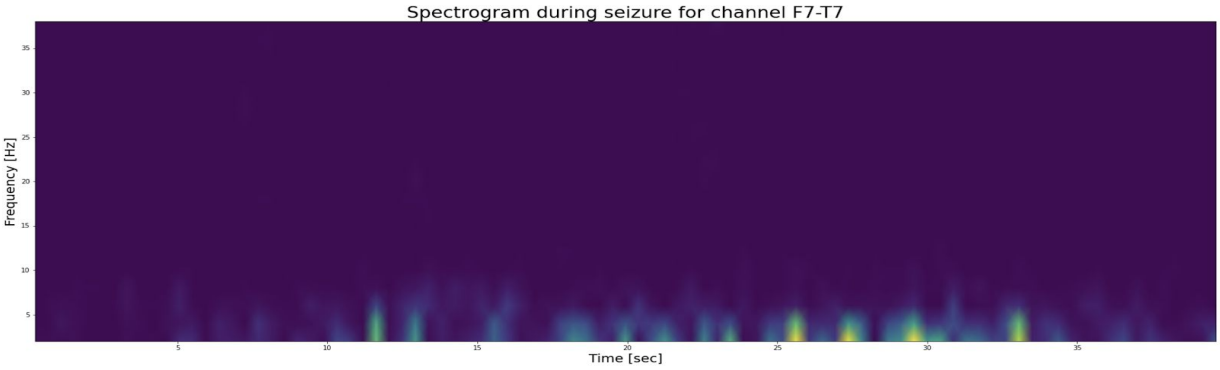
Most of the independent components have shown decrease in PE before the onset and increase after, which can be used for region of seizure identification.



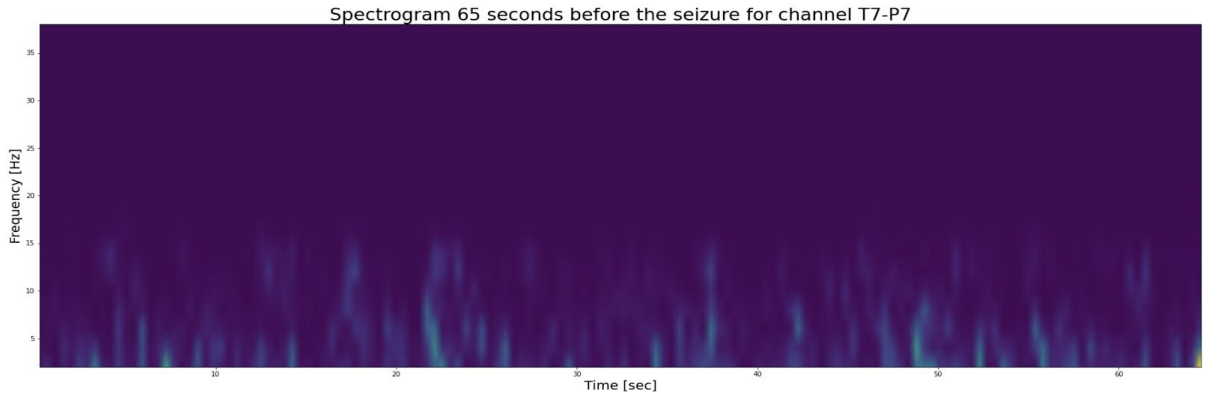
Permutation entropy plot for component 1 and 2 of a signal a patient chb09.

4.1.4 Short-Time Fourier Transform

As the additional feature that was used in the previously described study[13], I decided to use STFT for feature engineering purposes. After visualizing them I observed in a variety of patients no significant dependence between energy on different frequencies and seizure regions.



Spectrogram during the seizure on the channel F7-T7.



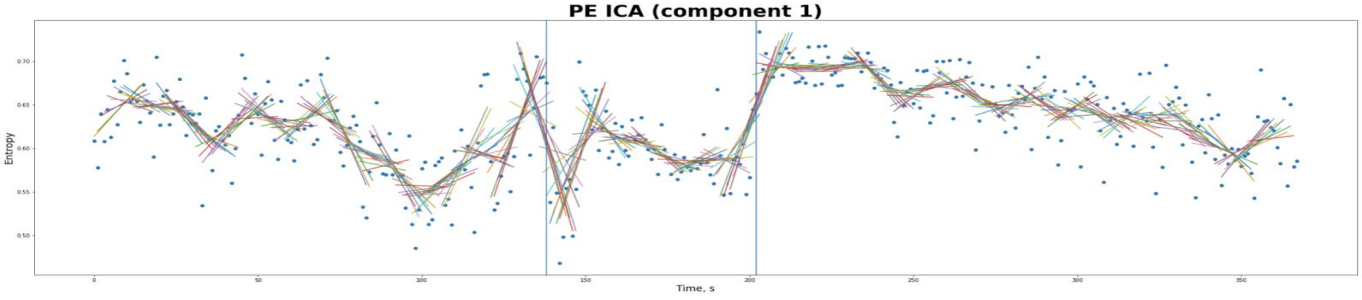
Spectrogram before the seizure of the channel T7-P7.

That was the reason why spectral energy was not chosen as a feature for training.

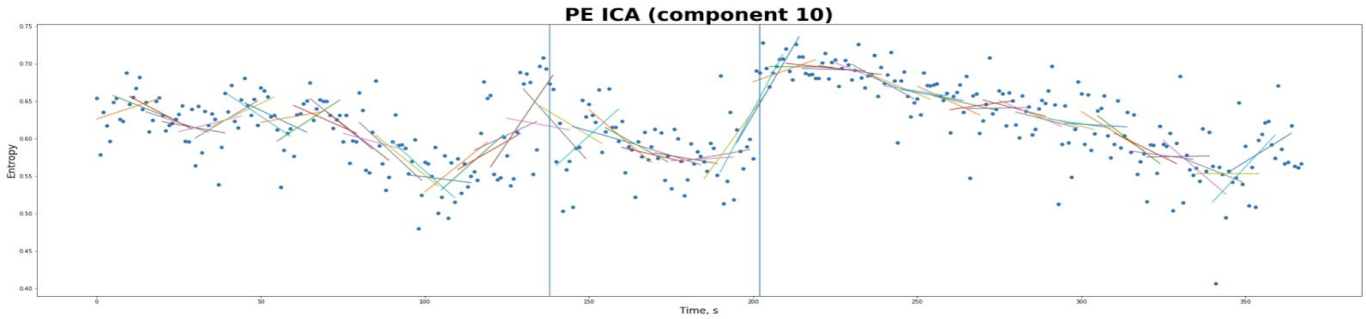
4.2 Time of the seizure prediction

Based on derived characteristics of permutation entropy in time I, firstly, wanted to measure what is a significant and non-significant decrease in PE and whether it correlates with time of the onset. For that reason I performed the following algorithm: I moved through the PE with the window of M seconds and shift of N seconds, took all of the points -- PEs according to the points -- and tried to fit linear regression (LR) to that time segment (time counts in range of number of seconds were chosen as independent variable for the following reason: the slope should be equally deriven for every time period, not depending on time).

The dependency on window size and shift is illustrated below:



Fitting the LR for IC number 1 in case of patient chb09 with window length = 15, shift - 1 second, k = 0.008. The region of seizure is highlighted blue.



Fitting the LR for IC number 10 in case of patient chb09 with window = 20 seconds, shift = 5 seconds, k = 0.008. The region of seizure is highlighted blue.

The absolute values of slopes of LR were then measured and compared to the boundary K. The boundary K value is a number that is used for classifying a fitter LR: if the slope is more than K then the time segment is classified as the one with a big shift. The PE taken was the averaged PE through all ICs.

For shorter calculation performance there were 2 steps: 1 - subset of patients definition and 2 - in every patient a segment around a seizure was chosen for the prediction. The subset of patients consisted of 26 cases of the patients: chb09, chb01, chb09, chb15 and chb20. The time segment in case of every seizure was chosen randomly: a time shift to the left of the seizure start time was the random number in range from 200 to 50 seconds, and to the right after the seizure end time -- from 100 to 200 seconds. Later for a PE in time all those significant slopes were collected and the one with the least (thus negative and the biggest absolute) value was chosen as the probable seizure start time.

The optimization was run in order to find out the optimal parameters M and K. The optimizer of choice was the Tree of Parzen Estimator. The optimized values were:
K (k boundary) = 0.5600061001302202,
N (shift) = 4,
M (window size) = 8.
The mean absolute error equals to 3 seconds.

4.3 Region of seizure prediction

The other task stated was to predict a mask vector of zeros and ones that will map onto the signal predicting the time period during which a seizure was observed. The 1s would correspond to the segments of signal when a seizure happened and 0s to non-seizure periods.

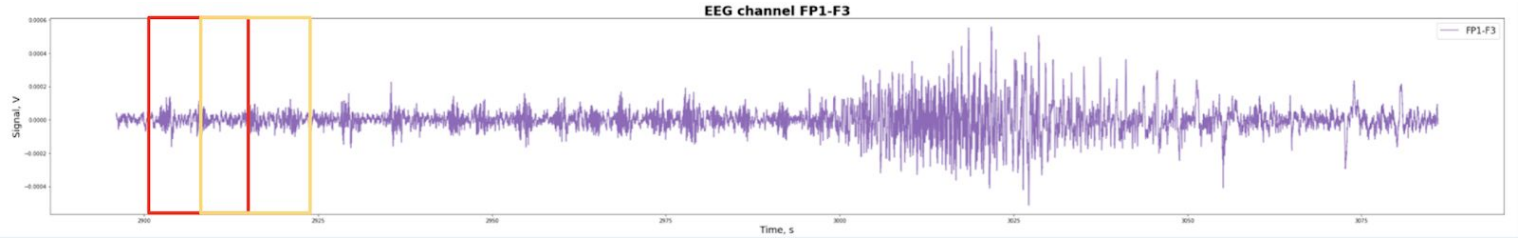
For this classification task the following features were proposed:

1) The cross correlation between the independent components
2) The cross correlation between the channels (as it performed good in the kaggle challenge)[1]
3) The averaged signal through all channels
4) The independent components
5) The permutation entropy of independent components

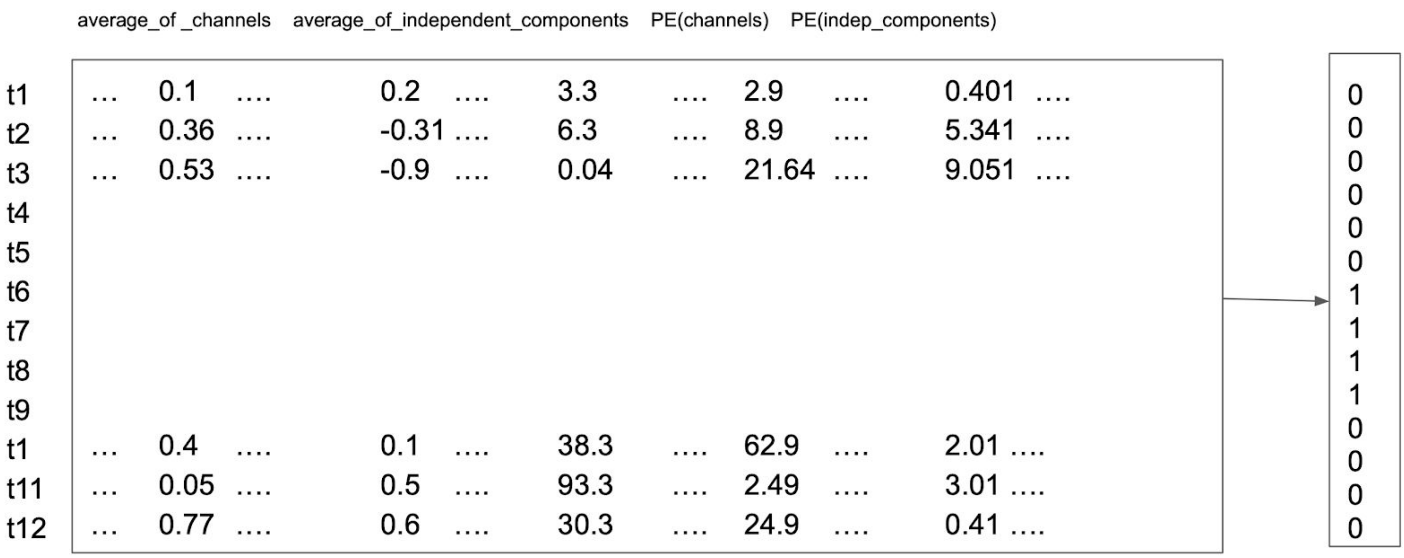
The cross correlation calculation was performed only at the beginning for the demonstration purposes due to limited computational and time resources.

In order to identify the epilepsy regions precisely the features were generated in time domain using the time window of 5 seconds and moving that window in time with shift equals to one second.

The feature derivation is visualized the following way:



The data are visualized the following way:



Example of how data is organized and the prediction task is performed. 'T1' notation can be interpreted as time period № 1.

Using generated features I performed the logistic regression using 26 cases of seizures. The evaluation metric was accuracy. The prediction was made not on the whole dataset of patients, but only for several cases due to resource limitation in memory: the calculation of PE, ICA and following stationarity tests took too much memory, such that the computations were stopped in the middle of training.

The reached accuracy equals 93.45%.

All of the related code can be found on github: https://github.com/sofiagarkot/epilepsy-prediction-signals .

5. Results and further steps

The results obtained during the project represent relative high accuracy in detection of seizures. The possible bias could be hidden in downsampling from the entire dataset. The further investigation requires optimization of needed memory resources as well as the derivation of other features. The testing of other classification techniques is also useful for comparison.

6. References

1. Baldassano SN, Brinkmann BH, Ung H, Blevins T, Conrad EC, Leyde K, Cook MJ, Khambhati AN, Wagenaar JB, Worrell GA, Litt B. Crowdsourcing seizure detection: algorithm development and validation on human implanted device recordings. Brain. 2017 Jun 1;140(6):1680-1691. doi: 10.1093/brain/awx098. PMID: 28459961; PMCID: PMC6075622.
2. "Public Data set: CHB-MIT EEG dataset," https://physionet.org/content/chbmit/1.0.0/, accessed: 2020-05-16.
3. Ali Shoeb, John Guttag. Application of Machine Learning to Epileptic Seizure Onset Detection. 27th International Conference on Machine Learning (ICML), June 21-24, 2010, Haifa, Israel.
4. M. Zhou, C. Tian, R. Cao, B. Wang, Y. Niu, T. Hu, H. Guo, and J. Xiang, "Epileptic seizure detection based on EEG signals and CNN," Frontiers in neuroinformatics, vol. 12, p. 95, 2018.
5. O. Avilov, O. Popov. Different permutation entropy patterns of electroencephalogram recorded during epileptiform activity.http://elc.kpi.ua/old/article/view/142299/151010
6. Nicolaou N., Georgiou J. (2012), "Detection of epileptic electroencephalogram based on Permutation Entropy and Support Vector Machines". Expert Systems with Applications. Vol. 39 . Pp. 202–209.
7. Federico Zilio et. al. Are intrinsic neural timescales related to sensory processing? Evidence from abnormal behavioural states.
8. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw. 2000;13(4-5):411–30.
9. Ulykbek Kairov, Determining the optimal number of independent components for reproducible transcriptomic data analysis
10. Aapo Hyvärinen , Independent Component Analysis. http://cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/node2.html
11. Tools overviewed before analysing: https://mne.tools/stable/auto_tutorials/preprocessing/plot_10_preprocessing_overview.html#sphx-glr-auto-tutorials-preprocessing-plot-10-preprocessing-overview-py
12. Aydin Akan, Ozlem Karabiber Cura. "Epileptic EEG Classification Using Synchrosqueezing Transform and Machine Learning", *International Journal of Neural Systems.*
13. Tzallas, Alexandros & Tsipouras, Markos & Fotiadis, Dimitrios. (2009). Epileptic Seizure Detection in EEGs Using Time–Frequency Analysis. IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society. 13. 703-10. 10.1109/TITB.2009.2017939.