Epileptic seizure prediction using long-term human iEEG recordings

Cédric Simar[1], Yann-Aël Le Borgne[1,2] and Gianluca Bontempi[1,2]

[1]Université Libre de Bruxelles
[2]Machine Learning Group

cedric.simar@ulb.ac.be

Abstract

Each year, thousands of people die from sudden unexpected death in epilepsy. This work introduces how the development of reliable predictive models could positively impact the daily life of epileptic patients. First, we introduce the electroencephalogram analysis and its application in epileptic seizure prediction. Second, we describe the state-of-the-art in seizure prediction. Third, we lay out the methodology and models used throughout this work and we show that a single feature, the Power Spectral Density, is independently capable of reliably forecasting the occurrence of a seizure. Finally we discuss the performances of different predictive models.

Introduction

Epilepsy is the second most common neurological disorder and affects 39 million people worldwide (D.I.I.P. Collaborators, 2016). About 80% of patients can achieve seizure control with anti-convulsive medication or resective surgery. For the remaining 20% of afflicted patients, several treatments are being developed to suppress seizures but are still in the experimental phase. The sudden and unforseen characteristic of seizure occurrences has a dramatic impact on the quality of life of patients affected by epilepsy. In addition to persistent anxiety symptoms, patients also suffer from heavy side effects related to long-term anti-epileptic medication as well as a lack of independence in daily activities, especially in regard to potentially life-threatening activities such as driving, swimming or simply taking a bath. In 2016, Kaggle hosted its second seizure prediction competition with the objective of developing a model able to reliably predict impending seizures that will eventually be embedded in a closed-loop implanted prevention system. When a warning is triggered, the imminent seizure could then be inhibited; e.g. by injecting anticonvulsant medication directly onto the seizure focus (Stein et al., 2000).

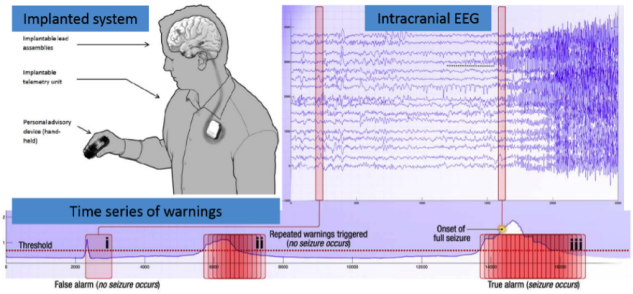

Figure 1: Closed-loop implanted seizures prevention system using intracranial EEG triggered warnings (Kaggle, 2016).

The long-term (from several months up to several years) human electroencephalogram recordings available for this competition were part of a world-first clinical trial of the implantable NeuroVista Seizure Advisory System (Cook et al., 2013). Such datasets containing a large quantity of seizures are critical to develop reliable and statistically validated models that could, one day, greatly improve the life of thousands of patients throughout the world.

Problem description

EEG Analysis

An electroencephalogram (EEG) uses multiple electrodes to measure the electrical activity of post-synaptic potentials of cortical neurons located at specific parts of the brain. The purpose of an EEG signal analysis is to use advanced signal processing techniques to extract relevant information about the brain state that are not directly visible in the time domain for the diagnosis of brain-related pathologies. The frequency domain analysis implemented in this work is a signal processing technique that divides the signal spectrum in different frequency bands, each associated with different brain states and mental activities. Typical frequency bands used in EEG analysis include delta (0.1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), low-gamma (30-70 Hz) and high-gamma (70-180 Hz) as described in (Howbert et al., 2014). Yet, the specific frequency ranges may slightly vary depending on the source and application.

Application to seizure prediction

The challenge of seizure prediction is to extract relevant features from EEG signals in order to accurately classify the patient state as either *preictal* (typically within the hour before the occurrence of a seizure) or *interictal* (between two occurrences).

State-of-the-art

(Mormann et al., 2007) exhaustively described the difficulty to objectively compare the results of seizure predictions studies and assess the performances of the models they propose. The main reasons include (i) the use of different types of EEG recordings (mostly surface EEG that contain much environmental artifacts or intracranial EEG that are almost artifact-free) (ii) different metrics definitions (iii) datasets of different species and recordings quality (iv) as well as a relative scarcity of preictal samples per patient. The latter issue is particularly critical because a small number of preictal samples make it impossible to generate an admissible test set and thus to perform an unbiased statistical validation of the model. Most of the models are thus fine-tuned to improve their performances on the same validation set that is subsequently used to assess their performances. Such results are likely over-optimistic and impossible to reproduce on unseen data samples.

In 2014, Kaggle hosted its first seizure prediction competition, followed in 2016 by a second competition using long-term iEEG recordings from three human patients. For the first time, the performances of different seizure prediction models could be evaluated on a same dataset containing enough preictal samples to perform an unbiased statistical validation with a separate test set. The best scores achieved in the 2014 and 2016 competitions were 0.83993 and 0.80701 AUROCC respectively.

Popular features used in winning submissions include:

- Correlation matrix and its eigenvalues (measures the similarity in the time or frequency domain between the EEG signals of two electrodes)
- Fractal Dimension (Petrosian and Higuchi)
- Hjorth Parameters (Activity, Mobility and Complexity)
- Hurst Exponent measures whether a long-term time series is trending, mean-reverting or a random walk
- Power Spectral Density
- Shannon Entropy (measures the degree of complexity of a time series, e.g. an EEG signal). Higher entropy means less predictability (Phung et al., 2014)
- Skewness (lack of a distribution symmetry) and Kurtosis (estimation of a distribution's peakedness)
- Spectral Edge Frequency at (80%, 90% and 95%) indicates the frequency under which a certain percentage of the signal is located.

Methodology

Data

The data set consists of one-hour sequences of 10-minute long data segments of raw electrical signals in the form of intracranial EEG (iEEG) sampled at a frequency of 400 Hertz from 16 electrodes placed on the outer layer of the cerebrum. The dataset contains 7,621 recordings totaling 41 GB organized in folders containing the training set (4,763 recordings) and the test set (2,858 recordings) for each of the three human patients. One-hour sequences of interictal data segments are separated from any postictal activity by a minimum period of four hours in order to avoid any signal contamination by preictal or postictal indicators. Sequences of preictal data segments covers a one-hour period prior to seizure. The training set totals 75 seizures (exactly 25 seizures per patient). Therefore, the minority (preictal) class only represents 9.4% of the total training set. This proportion is typical of medical diagnosis problems where datasets are inherently imbalanced.

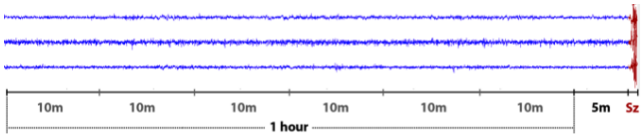

Figure 2: Raw signals from three electrodes of a one-hour sequence of data segments prior to seizure (Kaggle, 2016).

Preprocessing

Features selection

Due to the limited scope of this work, it was decided to focus the implementation on extracting one of the best single-feature and using it to compare the results of different predictive algorithms. Since previous published models were made up of multiple features, per-electrode (univariate) or cross-electrodes (multivariate) from both the time and frequency domains, evaluating the capability of a single feature to accurately predict epileptic seizures proved difficult. Nevertheless, literature has shown that features from the frequency domain, more specifically the Power Spectral Density (PSD) and the cross-electrodes correlation are independently capable of predicting seizures (Park et al., 2011; Howbert et al., 2014; Brinkmann et al., 2015). The PSD was selected to be implemented in this paper over cross-electrodes correlation because its spectrogram representation can be used by a wide range of predictive algorithms from conventional classifiers to convolutional neural networks. The PSD represents the strength of the signal varia-

tions distributed in the frequency domain. Given $x(t)$ a signal, f the signal frequency in Hertz, $\omega = 2\pi f$ the angular frequency and $j = \sqrt{-1}$. The Fourier Transform decomposes a signal (time domain) into its constituent sinusoids (frequency domain) and is defined as:

$$\mathcal{F}[x(t)] = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt$$

The Auto-Correlation Function estimate how a signal is correlated with a copy of itself shifted in time. It is used to detect periodicity in a signal and is defined as:

$$R(\tau) = x(\tau) * x(-\tau) = \int_{-\infty}^{\infty} x(t)x(t+\tau)dt$$

The Power Spectral Density is commonly defined as the Fourier Transform of the Auto-Correlation Function:

$$S_x(\omega) = \mathcal{F}[R(\tau)] = \int_{-\infty}^{\infty} R(\tau)e^{-j\omega\tau}d\tau$$

The spectral power of $x(t)$ in the frequency band $[f_1, f_2]$ is computed by integrating over the frequency range as follow:

$$P_{[f_1,f_2]} = 2\int_{f_1}^{f_2} S_x(\omega)df$$

Within a preictal state the spectral power of the delta frequency band (0.1 - 4.0 Hz) decreases and the spectral power of other frequency bands increases (Mormann et al., 2005), which corroborates the potential of the PSD feature for accurately classifying preictal and interictal states.

Dropouts management

Unlike a previous seizure prediction challenge, most of the 10-minute long recordings of the data set contain moments where the portable device failed to record the signal (*data dropouts*). Since flat segments in a recording would considerably alter the quality of the Fourier Transform, the first preprocessing step is to split the 10-minute long recording into several shorter epochs without dropout. An epoch length of 30 seconds was empirically determined from the analysis of previous winning submissions and from the resulting size of the training set. During this process, an overlap of 0.25 is used in order to artificially increase the preictal class of the training set and to later reduce the inherent unbalancing factor (the interictal class is present with an averaged 20:1 ratio). After this step, before features extraction, each files contains a number of matrices of dimension 16 electrodes x 12.000 samples corresponding to 30-second long epochs without dropout extracted from the raw iEEG recording.

Another reason to split 10-minute long recordings into shorter epochs lies in the non-stationary characteristics of preictal indicators. In a 10-minute recording labeled as preictal, the physiological signal perturbations that indicates the preictal state is relatively short and sparsely distributed. Extracting features from shorter epochs would therefore allow a model to better identify preictal indicators. The matter of how to optimally combine the model predictions based on the different epochs will be addressed in the "Models" section.

Artifacts

In addition to synaptic activity, raw EEG signals also contain a certain proportion of noise, more specifically artifacts that have to be removed before features extraction can be performed with the finest precision. EEG artifacts can be classified as either physiological or non-physiological (i.e. from recording instruments or patient environment). The main sources of physiological artifacts are eye blinks, cardiac contractions and muscle movements. A common source of non-physiological artifacts is the power line noise arising from standard AC power supply frequencies (50 or 60 Hz depending on the country). Power line noise can commonly be removed using notch filters at 50 or 60 Hz. This type of band-stop filter attenuates a narrow range of specific frequencies (e.g. 49-51 Hz) without altering the rest of the frequency spectrum. On the other hand, physiological artifacts removal requires a more complex analysis as well as the use of sophisticated algorithms such as Independent Components Analysis (Cohen, 2014).

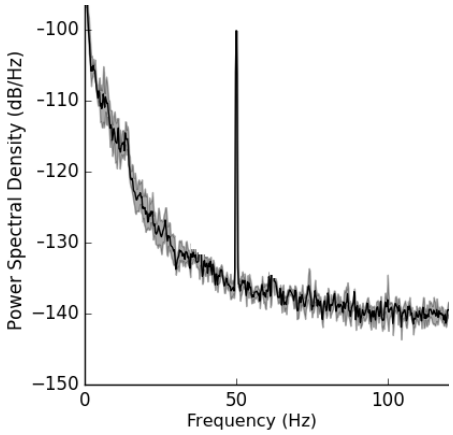

Figure 3: Power line noise at the 50 Hz frequency in an EEG recording.

Since the iEEG signals provided for the purpose of the competition were recorded with a battery powered device using common average referencing in order to attenuate the noise at the 50 Hz frequency, it was not necessary to apply a notch filter prior to extracting features.

Models

Patient specific or non-specific model

When developing a predictive model with limited data available to train on, a common practice is to try increasing the size of the dataset. One way is to artificially produce new samples from the original ones as discussed in the "Dropout Management" section. Another tempting way would be to aggregate datasets from different patients in order to develop a patient-non-specific model trained with more samples from the preictal class that would therefore be able to better generalize compared to patient-specific models trained with fewer preictal samples. However, the latter intuition would lead to a significant drop in performances because the measurement baselines in specific frequency bands depend on patient-specific factors such as (i) the lack of correspondence between electrodes mapping (ii) the large divergence between human brains and (iii) patient-specific latent conditions such as chronic stress, dementia or schizophrenia. For this reason, a patient-specific model is implemented.

Convolutional Neural Network

A Convolutional Neural Network (CNN) is a biologically-inspired variant of feed-forward neural networks which neurons organization aims to emulate the animal visual cortex (Lecun et al., 1998). It is considered as a state-of-the-art technique for image classification.
The CNN architecture implemented in this work is based on a Master thesis (Korshunova, 2015) that ranked tenth in the previous Kaggle competition on seizure prediction. Unlike the previous competition's dataset, the current dataset contains a high amount of data dropouts. A number of adjustments were therefore necessary in order to use the architecture described in the thesis with the current dataset.

The features extraction for the CNN approach consists of the following steps. A band-pass filter between 0.1 and 180 Hz is applied on the iEEG recording of each electrode. Subsequently, each 30-second long signal is divided into non-overlapping 3-second epochs on which is computed the log_{10} of its Power Spectrum Density. The PSD is then divided in six frequency bands (see introduction) in each of which the mean is computed to form a binned spectrogram of dimension 6 bands $\times$ 10 epochs subsequently standardized per frequency band. The combined standardized spectrograms from each electrode form an input of dimension 96 $\times$ 10 to the CNN as illustrated in Fig. 4.

As previously stated, the signal perturbations that indicate the preictal state is sparsely distributed within the hour before a seizure occurs. A perturbation can be located anywhere in the 30-second epoch or might not even appear at all. The convolution kernels configuration should therefore enable the network to learn features from different time segments separately and combine them to form a prediction.

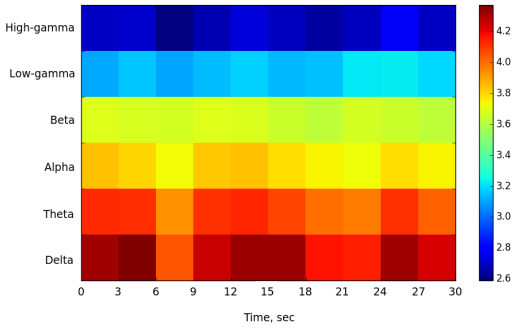

Figure 4: Binned spectrogram of one electrode computed from the PSD in the CNN approach.

The network architecture illustrated at Fig. 5 is structured as follow. The first layer (C1) performs a convolution with 16 kernels of dimension 96 x 1 corresponding to 1 minute of the combined standardized spectrograms from each electrode. The result of the first convolution is 16 feature maps of dimension 1 x 10 respectively. The second layer (C2) performs a convolution with 32 kernels of dimension 16 x 2. The result of the second convolution is 32 feature maps of dimension 1 x 9 respectively. The third layer (GP3) is a Global Temporal Pooling layer that computes statistics (mean, maximum, minimum, variance, geometric mean and L2 norm) across the whole time axis of the 32 feature maps from the second convolution. The fourth layer (F4) contains 128 neurons and fully connects the 192 outputs of the GP3 to a logistic regression layer (RegLog). The architecture totals 27.312 trainable variables: 1.536 weights and 16 biases for C1, 1.024 weights and 32 biases for C2 and 24.576 weights and 128 biases for F4.

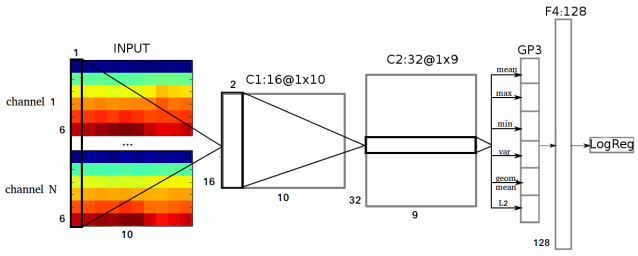

Figure 5: Convolutional Neural Network architecture as implemented (Korshunova, 2015).

The CNN architecture was implemented in Python 2.7 using the open-source Tensorflow 0.12 library for numerical computation using data flow graphs. The kernel values are initialized following a truncated normal distribution with a standard deviation of $\sqrt{\frac{2}{n_l}}$ where n_l represents the number of inputs in layer l as recommended in (He et al., 2015) in order to break symmetry, ensure consistent gradients back-

propagation and avoid neurons saturation. An activation function is a differentiable and non-linear function applied to the output of each neuron in order to create a non-linear decision boundary from a linear combination of inputs and weights that formed these outputs. The rectified linear unit (ReLU) (Nair and Hinton, 2010) was used as an activation function for all the layers of the CNN except the GP3. In order to reduce network overfitting, L2 regularization was applied to F4 and dropout regularization (Hinton et al., 2012) was applied to GP3 and F4.

Conventional approach

Most of the published papers on seizure prediction uses the Power Spectral Density as a part of their model made up of multiple features. The PSD feature extraction for this approach consists of the following steps. The PSD is estimated on the whole 30-second epoch using Welch's method (Welch, 1967) that computes an estimate of the PSD by partitioning the signal in overlapping windows, forming a modified periodogram based on a DFT on each window using specific frequencies and averaging the values of the periodogram. The selected windows size and overlap are 512 samples and 25% respectively. For each of the 16 electrodes, the corresponding PSD is subsequently divided in six different frequency bands in each of which the mean is computed to form a binned spectrogram of dimension 6 bands $\times$ 1 epoch. The 16 spectrograms are normalized by their total power before a log_{10} is eventually applied. The combined spectrograms from each electrode are flattened into a one-dimensional array of 96 values to form an input to the classifier as illustrated in Fig. 6.

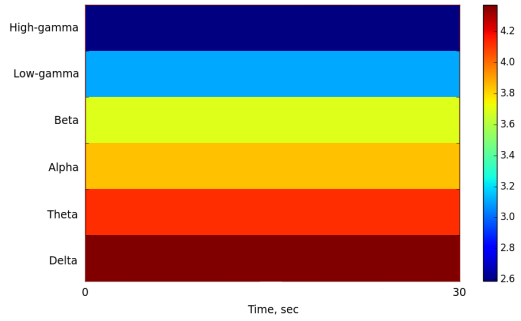

Figure 6: Binned spectrogram of one electrode computed from the PSD in the conventional approach.

This approach mainly focuses on learning features (i.e. the increase or decrease of spectral power in different frequency bands) from a single time segment and thus does not contain a temporal dimension, unlike the CNN approach.

The performance of the following classifiers are evaluated in the "Benchmark" section using the features from the conventional approach: Gradient Boosting (Friedman, 2001), Logistic Regression (Cramer, 2002), Random Forest (Breiman, 2001) and Support Vector Machine (Cortes and Vapnik, 1995).

XGBoost (Chen and Guestrin, 2016) is a tree boosting system based on (Friedman, 2001) that has recently gained popularity after having achieved state-of-the-art performances on various classification problems. The boosting principle is to form an ensemble of individually weak classifiers and linearly combines their output to form one strong classifier (Schapire, 1990). XGBoost iteratively builds an ensemble of weak decision tree classifiers, sums the predictions from each tree to form the ensemble prediction and subsequently train a separate decision tree to minimize the difference between the ensemble prediction and the objective function.

How to combine predictions

As stated hereabove, preictal indicators are sparsely distributed within the hour before a seizure occurs. Hence, in a 10-minute long signal labeled as preictal there might be one, several or none of such preictal indicators. A certain proportion of 30-second epochs labeled as preictal thus contain such indicators and are correctly classified. However, the other epochs included in the 10-minute segment are wrongly classified as interictal since they do not contain any preictal indicators. Therefore, in order to correctly classify the 10-minute signal as preictal, the final prediction can be computed using the maximum (and not the mean) of the combined predicted probabilities from every 30-second epochs contained in the 10-minute segment. The standard deviation of the combined predicted probabilities is also reported to perform well.

Results

Models evaluation methodology

In order to perform an unbiased estimation of the model prediction accuracy and its ability to generalize, the model has to be evaluated using data that are completely separated from the model's learning process. Therefore, as a rule of thumb, the dataset is originally split into a training set (used in the learning phase) and a test set (used to evaluate the selected model) with a proportion of 75% and 25% respectively. During the learning phase, a predictive model adapts its parameters in order to minimize a cost function. The more a model is trained with the same training data, the more it tends to fit the training data to reduce the training error. If a model fits very well the training data but is not able to generalize to accurately classify unseen data, it is *overfitting* the training set. In order to avoid overfitting, the model's performances should be monitored during the training phase and the training should be stopped when the model's generalization worsens (*early stopping* (Yuan et al., 2007)). To this end, the original training set is split into a smaller training set and a validation set which is used to monitor the model accuracy and fine tune its parameters during the training phase. The choice of the best model is therefore based on its per-

formance on the validation set, which biases its accuracy estimate. The selected model is trained again using the whole training set before an unbiased statistical validation can be performed using the test set.

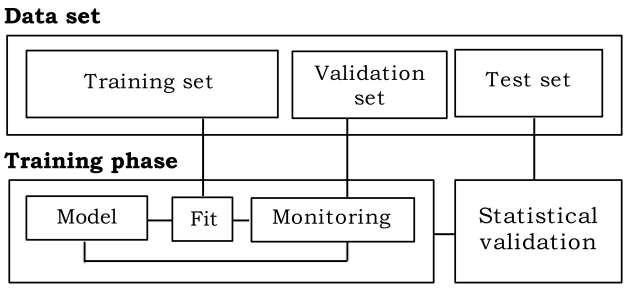

Figure 7: Simple validation diagram.

The accuracy estimate from the conventional validation technique described hereabove is computed on the fixed validation set and, therefore, does not allow to determine how sensitive the accuracy estimate is to specific training samples. In order to compute a more robust estimate, we use a k-fold cross validation that partitions the training set in k different folds. Iteratively, k-1 folds are used to train a model and the remaining fold is used as a validation set to estimate the model accuracy. After k iterations, k accuracy estimates were computed using all k folds exactly once. With this process, the whole training set is used both for training and validation. The k estimates are averaged to form a more robust accuracy estimate of the model.

In the scope of this work, each predictive model's accuracy is hereunder estimated using a 10-fold cross validation. The size and long-term characteristics of the dataset provided as well as the number of seizures included make it possible to perform a statistical validation.

Evaluation metrics

In the domain of epileptic seizures prediction, as in many classification problems in the medical field, the interictal class is overly dominant and the consequences of a misclassification (e.g. a false negative) can be life-threatening. Under these constraints, the mean accuracy is not the most adequate measure to evaluate a model's performances.

The metric used during the competition is the Area Under the Receiver Operating Characteristic Curve (AUROCC). The Receiver Operating Characteristic (ROC) curve plots the True Positive (TP) rate (also called Sensitivity or Recall) on the Y axis against the False Positive (FP) rate (1-Specificity) on the X axis. A random classifier has an AUROCC of 0.5.

Another interesting metric to consider when evaluating a model trained with an imbalanced dataset is the Precision / Recall (PR) curve that plots the Precision $TP/(TP + FalseNegatives)$ on the Y axis against the Recall on the X axis. The PR curve represents the fraction of predictions that are FP.

Both ROC and PR curves drawn hereunder are computed using the average of the individual curves from each fold of the cross validation.

Benchmark

A benchmark of the following classifiers is presented hereunder: Convolutional Neural Network (CNN), XGBoost, Logistic Regression with L2 penalty (LR), Random Forest (RF) and Support Vector Machine (SVM). The optimal parameters for each classifier were selected empirically (CNN) or with the cross-validated grid-search of the Scikit-Learn library (the others). The AUROCC Kaggle metric corresponds to the mean between public and private leaderboard scores.

	CNN	XGBoost	LR	RF	SVM
CV Metrics					
Sensitivity	0.56	0.57	0.55	0.31	0.56
Specificity	0.55	0.72	0.72	0.86	0.69
AUROCC	0.62	0.76	0.76	0.75	0.73
Kaggle Metric					
AUROCC	0.62	0.71	0.69	0.64	0.67

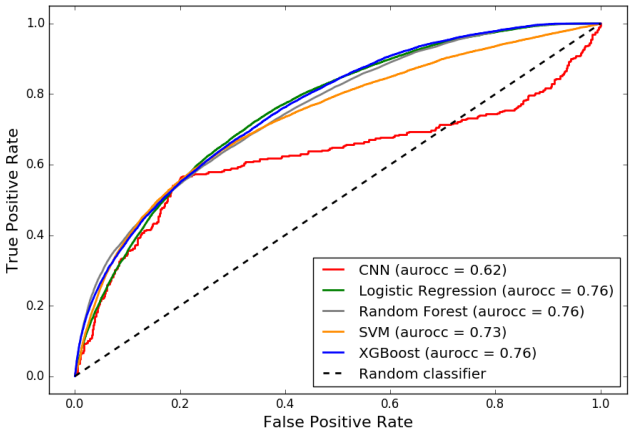

Figure 8: ROC curves and AUROCC of the five classifiers

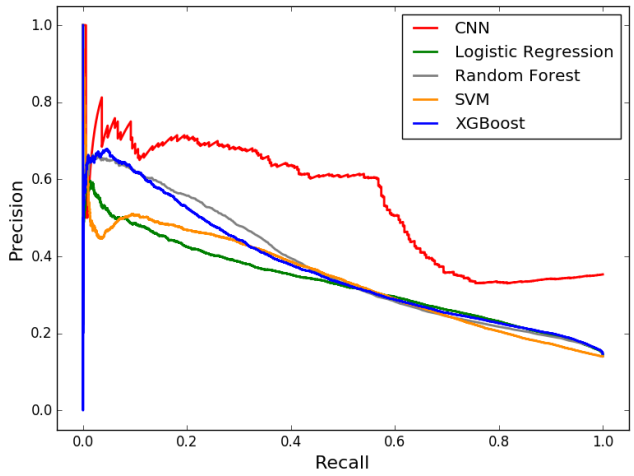

Figure 9: PR curves of the five classifiers

Discussion

Cross-validation pitfall

With regard to the k-fold cross-validation, a first intuition was to apply a stratified k-fold cross validation that partitions the training set in k different folds containing approximately the same proportion of each class as in the complete training set. However, this approach leads to significantly over-optimistic metrics (CV AUROCC around 0.85) that did not reflect the underlying performances of the model to accurately classify new samples from the test set (Kaggle AUROCC around 0.68). The hypothesis shared by several competitors was that 10-minute segments from the same hour are closely related. Thus, if such segments are present in both the training and validation sets, the model tends to classify a sample according to its similarity with a previously learned sample rather than the presence of preictal indicators. In order to avoid this leak in the validation set, we use a group k-fold cross validation that partitions the training set in k different folds so that all six segments from the same hour belongs to the same fold. After using the corrected cross-validation, the resulting AUROCC estimates matched more closely the Kaggle metric as reported in the benchmark.

Results comparison

Regarding the conventional approach, within the winning submission ensemble of models, two XGBoost classifiers used the combined spectrograms flattened into a one-dimensional array of 96 values as a single source of features and both reported an AUROCC of 0.77. The difference between their results and the result presented in this work (0.71 AUROCC) should be explained by the fact that they doubled the number of their preictal samples available for training by using a leak that happened during the competition. Unfortunately, this work could not reproduce their results since the old test set containing the leak is no longer publicly available.

Regarding the CNN approach, the result of 0.62 AUROCC is consistent with the results of other competitors who ventured into a Deep Learning approach using the PSD features. However, it is unclear if the lack of performances compared to the conventional approach is inherent to the model implementation itself or the adaptations required to transpose the original model to this particular competition. Unlike more conventional classifiers, the CNN model implemented in this work either failed to converge to a global minimum when using group k-fold cross-validation or failed to generalize due to overfitting when using stratifed k-fold cross validation.

Improvements

Firstly, in order to refine the model presented in this work and reach state-of-the-art performances, it is critical to include more classifiers trained with different features to form an ensemble which combines its individual classifier predictions in order to produce a better prediction that the best of its individual models. The winning submission, for example, includes three ensembles that combine four, four and three classifiers respectively.

Lastly, since the signal perturbations that indicate the preictal state is sparsely distributed in the 10-minute segment, from 20 to 50% of 30-second epochs constituting this segment do not contain any preictal indicators but are nevertheless labeled as preictal. It would be interesting to see if a classifier would draw a sharper decision boundary if these specific epochs would be withdrawn from the training set, and, *in fine*, quantify the evolution of its performances on the test set.

Conclusion

This work introduced the EEG analysis and its application in state-of-the-art epileptic seizure prediction. After having thoroughly laid out the methodology and underlined the need for an unbiased statistical validation, we showed that a single feature, the Power Spectral Density, reaching 0.71 AUROCC in a Kaggle competition using a XGBoost classifier, is independently capable of reliably forecasting the occurrence of a seizure. Finally we compared the performances of different predictive algorithms and proposed some improvements for future work.

Acknowledgements

I gratefully thank Gianluca Bontempi for giving me the chance to work on such an inspiring topic and Yann-Aël Le Borgne who guided and supported me throughout the development of this work.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Brinkmann, B. H., Patterson, E. E., Vite, C., Vasoli, V. M., Crepeau, D., Stead, M., Howbert, J. J., Cherkassky, V., Wagenaar, J. B., Litt, B., and Worrell, G. A. (2015). Forecasting seizures using intracranial eeg measures and svm in naturally occurring canine epilepsy. *PLoS ONE*, 10:12.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *arXiv:1603.02754*, 3:13.

Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. The MIT Press.

Cook, M. J., O'Brien, T. J., Berkovic, S. F., Murphy, M., Morokoff, A., Fabinyi, G., D'Souza, W., Yerra, R., Archer, J., Litewka, L., Hosking, S., Lightfoot, P., Ruedebusch, V., Sheffield, D., Snyder, D., Leyde, K., and Himes, D. (2013). Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. *The Lancet Neurology*, 12:563–571.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Cramer, J. (2002). The origins of logistic regression. *Tinbergen Institute Working Paper*, 4:119:16.

D.I.I.P. Collaborators (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 19902015: a systematic analysis for the global burden of disease study 2015. *The Lancet Neurology*, 388:1545–1601.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Compute Vision Foundation*, 1:1026–1034.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580*, 1:18.

Howbert, J. J., Patterson, E. E., Stead, S. M., Brinkmann, B., Vasoli, V., Crepeau, D., Vite, C. H., Sturges, B., Ruedebusch, V., Mavoori, J., Leyde, K., Sheffield, W. D., Litt, B., and Worrell, G. A. (2014). Forecasting seizures in dogs with naturally occurring epilepsy. *PLoS ONE*, 9:8.

Korshunova, I. (2015). Epileptic seizure prediction using deep learning. Master's thesis, Universiteit Gent, Gent, Belgium.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324.

Mormann, F., Andrzejak, R. G., Elge, C. E., and Lehnertz, K. (2007). Seizure prediction: the long and winding road. *Brain*, 130:314–333.

Mormann, F., Kreuz, T., Rieke, C., Andrzejak, R. G., Kraskov, A., David, P., Elger, C. E., and Lehnertz, K. (2005). On the predictability of epileptic seizures. *Clinical Neurophysiology*, 116:569–587.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27 th International Conference on Machine Learning*, 1:807–814.

Park, Y., Luo, L., Parhi, K. K., and Netoff, T. (2011). Seizure prediction with spectral power of eeg using cost-sensitive support vector machines. *Epilepsia*, 52:1761–1770.

Phung, D., Tran, D., Ma, W., Nguyen, P., and Pham, T. (2014). Using shannon entropy as eeg signal feature for fast person identification. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 7:413–418.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5:197–227.

Stein, A., Eder, H., Blum, D., Drachev, A., and Fisher, R. (2000). An automated drug delivery system for focal epilepsy. *Epilepsy Res*, 39:103–114.

Welch, P. D. (1967). The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio and Electroacoust.*, 15:70–73.

Yuan, Y., Rosasco, L., and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315.