Classification of Patterns of EEG Synchronization for Seizure Prediction

Piotr Mirowski MSc*, Deepak Madhavan MD†, Yann LeCun PhD*, Ruben Kuzniecky MD‡

*Courant Institute of Mathematical Sciences, New York University, 719 Broadway, New York, NY 10003, USA

†Department of Neurological Sciences, 982045 University of Nebraska Medical Center, Omaha, NE 68198, USA

‡New York University Comprehensive Epilepsy Center, 223 East 34th St., New York, NY 10016, USA

Corresponding Author:

Piotr Mirowski, Ph.D. candidate
Courant Institute of Mathematical Sciences, New York University
719 Broadway, 12th Floor, New York, NY 10003, USA
Tel: +1 203-278-1803
Email: piotr.mirowski @ computer.org
Fax: +1 212-263-8342

§ Portions of this manuscript were presented at the 2008 American Epilepsy Society annual meeting and at the 2008 IEEE Workshop on Machine Learning for Signal Processing

Abstract

Objective: Research in seizure prediction from intracranial EEG has highlighted the usefulness of bivariate measures of brainwave synchronization. Spatio-temporal bivariate features are very high-dimensional and cannot be analyzed with conventional statistical methods. Hence, we propose state-of-the-art machine learning methods that handle high-dimensional inputs.
Methods: We computed bivariate features of EEG synchronization (cross-correlation, nonlinear interdependence, dynamical entrainment or wavelet synchrony) on the 21-patient Freiburg dataset. Features from all channel pairs and frequencies were aggregated over consecutive time points, to form patterns. Patient-specific machine learning-based classifiers (support vector machines, logistic regression or convolutional neural networks) were trained to discriminate interictal from preictal patterns of features. In this explorative study, we evaluated out-of-sample seizure prediction performance, and compared each combination of feature type and classifier.
Results: Among the evaluated methods, convolutional networks combined with wavelet coherence successfully predicted all ouf-of-sample seizures, without false alarms, on 15 patients, yielding 71% sensitivity and 0 false positives.
Conclusions: Our best machine learning technique applied to spatio-temporal patterns of EEG synchronization outperformed previous seizure prediction methods on the Freiburg dataset.
Significance: By learning spatio-temporal dynamics of EEG synchronization, pattern recognition could capture patient-specific seizure precursors. Further investigation on additional datasets should include the seizure prediction horizon.

Keywords

seizure prediction; feature extraction; classification; pattern recognition; machine learning; neural networks

1. Introduction

Recent multi-center clinical studies showed evidence of premonitory symptoms in 6.2% of 500 patients with epilepsy (Schulze-Bonhage et al., 2006). Another interview-based study found that 50% of 562 patients felt "auras" before seizures (Rajna et al., 1997). Such clinical observations give an incentive to search for premonitory changes on EEG recordings from the brain, and to implement a device that would automatically forewarn the patient. However, and despite decades of research, research in seizure prediction is still qualified as a "long and winding road" (Mormann et al., 2007).

Most current seizure prediction approaches (Arnhold et al., 1999; Iasemidis et al., 2005; Lehnertz and Litt, 2005; Lehnertz et al., 2007; Le Van Quyen et al., 2005; Litt and Echauz, 2002, Mormann et al., 2006, 2007) can be summarized into (1) extracting measurements from EEG over time and (2) classifying them into a preictal or interictal state. The ictal and postictal states are discarded from the classification, because the task is not to detect undergoing seizures, but eventually to warn the patient about future ones, so that the patient, the clinician, or an implanted device can act accordingly.

The method described in this article follows a similar methodology: (1) feature extraction, followed by (2) binary classification of patterns of features into preictal or interictal states. Section 1.1 of the Introduction overviews existing techniques for feature extraction from EEG (1), while Methods section 2.2 and Appendix A detail specific features used in the proposed method.

The breakthrough of our technique lies in the pattern recognition and machine learning-powered classification of features (2). The proposed pattern-based classification is described in Methods sections 2.3 through 2.5. As can be seen in Introduction section 1.2, the proposed method takes advantage of decade of research in image processing and vision, but is also a novelty in the field of seizure prediction. Moreover, Results section 3 shows that our method achieves superior seizure prediction results on the Freiburg EEG dataset (described in section 2.1). Finally, section 4 discusses the limitations of the proposed method.

1.1. Feature extraction from EEG

Seizure prediction methods have in common an initial building block consisting of the extraction of EEG features. All EEG features are computed over a short time window of a few seconds to a few minutes. One can distinguish between univariate measures, computed on each EEG channel separately, and bivariate (or multivariate) measures, which quantify some relationship, such as synchronization, between two or more EEG channels. Although a plethora of univariate features has been investigated for seizure prediction (Esteller et al., 2005; Harrison et al., 2005; Jerger et al., 2005; Jouny et al., 2005), none of them has succeeded in that task, as illustrated in an extensive study comparing most univariate and bivariate techniques (Mormann et al., 2005), which also confirmed the superiority of bivariate measurements for seizure prediction.

In parallel to comparative study (Mormann et al., 2005), and despite the current lack of a complete neurological understanding of the preictal brain state, researchers increasingly hypothesize that brainwave synchronization patterns might differentiate interictal, preictal and ictal states (Le Van Quyen et al., 2003). From clinical observations on the synchronization of neural activity, it has been suggested that interictal phases correspond to moderate synchronization within the brain at large frequency bands, and that there is a preictal decrease in the beta range synchronization between the epileptic focus and other brain areas, followed by a subsequent hyper-synchronization at the seizure onset. These considerations motivated our choice of bivariate EEG features.

As described in Methods sections 2.2 and Appendix A, this article evaluates four kinds of EEG synchronization (bivariate) features: one simple linear feature called Maximum Cross-Correlation (Mormann et al., 2005; Appendix A.1) and three nonlinear features. The first and popular nonlinear measure is Nonlinear Interdependence, which measures the distance, in state-space, between time-delay embedded trajectories of two EEG channels (Arnhold et al., 1999; Mormann et al., 2005) (see Appendix A.2). The second measure, also called Dynamical Entrainment, is based on the measure of chaos in the EEG. It estimates from any two observed time-series, the difference of their largest Lyapunov exponents, i.e. the exponential rates of growth of an initial perturbation (see Appendix A.3). Finally, a third type of nonlinear bivariate measures that takes advantage of the frequency content of EEG signals is phase synchronization. First, two equivalent techniques can be employed to extract the frequency-specific phase of

EEG signal: bandpass filtering followed by Hilbert transform or Wavelet transform (Le Van Quyen et al., 2001). Then, statistics on the difference of phases between two channels (such as phase-locking synchrony) are computed for specific combinations of channels and frequencies (Le Van Quyen et al., 2005).

1.2. Feature classification for seizure prediction

Once univariate or bivariate, linear or nonlinear measurements are derived from EEG, the most common approach for seizure prediction is the simple binary classification of a single variable (Lehnertz et al., 2007; Mormann et al., 2005). Their hypothesis is that there should be a preictal increase or decrease in the values of an EEG-derived feature. Statistical methods consist in an a posteriori and in-sample tuning of a binary classification threshold (e.g. pre-ictal vs. interictal) on that unique measure extracted from EEG.

The usage of a simple binary threshold has limitations detailed in the Discussion section 4.2. Essentially, it does not allow using high-dimensional features. By contrast, machine learning theory (sometimes also called statistical learning theory) easily handles high-dimensional and spatio-temporal data, as illustrated in its countless applications such as video or sound recognition.

Most importantly, machine learning provides both with a methodology for learning by example from data, and for quantifying the efficiency of the learning process (Vapnik, 1995). The available data set is divided into a training set ("in-sample") and a testing set ("out-of-sample"). Training consists in iteratively adjusting the parameters of the machine in order to minimize the empirical error made on in-sample data, and a theoretical risk related to the complexity of the machine (e.g. number of adjustable parameters). The training set can be further subdivided into training and cross-validation subsets, so that training is stopped before over-fitting when the cross-validation error starts to increase.

As a paramount example of machine learning algorithms, feed-forward Neural Networks (NN) can learn a mapping between multi-dimensional inputs and corresponding targets. The architecture of a neural network is an ensemble of interconnected processing units, organized in successive layers. Learning consists in tuning the connection weights by back-propagating the gradient of classification errors through the layers of the NN (Rumelhart et al., 1986). Convolutional networks are a further specialized architecture able to extract distortion-invariant patterns such as for handwriting recognition. One such convolutional network architecture, called LeNet5, is currently used in the verification of handwriting on most bank checks in the United States (LeCun et al., 1998a) and has been more recently shown to enable autonomous robot navigation from raw images coming from two (stereoscopic) cameras (LeCun et al., 2005). This sophisticated neural network successfully learnt a large collection of highly noisy visual patterns and was capable of avoiding obstacles in unknown terrain.

Another machine learning algorithm used for multi-dimensional classification is called Support Vector Machines (SVM). SVMs first compute a metric between all training examples, called the kernel matrix, and then learn to associate the right target output to a given input, by solving a quadratic programming problem (Cortes and Vapnik, 1995; Vapnik, 1995).

Machine learning techniques have been applied, in a very limited scope, mostly to select subsets of features and corresponding EEG channels for further statistical classification, but rarely to the classification task itself. Examples of such algorithms for channel selection included Quadratic Programming (Iasemidis et al., 2005), K-means (Iasemidis et al., 2005; Le Van Quyen et al., 2005), and Genetic Optimization (D'Alessandro et al., 2003, 2005). An example of a more sophisticated machine learning procedure for seizure prediction (Petrossian et al., 2000) consisted in feeding raw EEG time series and their wavelet transform coefficients into a Recurrent Neural Network (RNN), i.e. a neural network that maintains a "memory" of previous inputs and thus learns temporal dependencies between consecutive samples. The RNN was trained to classify each EEG channel separately as being in an interictal or preictal state. That RNN however did not take advantage of bivariate measurements from EEG. Most importantly, the dataset was very short (minutes before a seizure) and the technique has not been validated on large case studies.

Our article compares three types of machine learning classifiers: logistic regression, SVMs and convolutional networks, all described in Methods sections 2.4. Instead of relying on one-dimensional features, the classifiers were trained to handle high-dimensional patterns (detailed in section 2.3) and managed to select subsets of features (channels and frequencies) during the learning process (see section 2.5).

2. Methods

Our entire seizure prediction methodology can be decomposed as following: selection of training and testing data, as well as EEG filtering (section 2.1), computation of bivariate features of EEG synchronization (section 2.2), aggregation of features into spatio-temporal, or spatio-temporal and frequency-based, patterns (section 2.3), machine learning-based optimization of a classifier that inputs patterns of bivariate features and outputs the preictal or interictal category (section 2.4) and retrospective sensitivity analysis to understand the importance of each EEG channel and frequency band within the patterns of features (section 2.5).

2.1. Data and preprocessing

We developed and evaluated our seizure prediction methodology on the publicly available EEG database at the Epilepsy Center of the University Hospital of Freiburg, Germany (https://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database/), containing invasive EEG recordings of 21 patients suffering from medically intractable focal epilepsy. Previous analysis of this dataset (Aschenbrenner-Scheibe et al., 1999; Maiwald et al., 2004; Schelter et al., 2006 a,b; Schulze-Bonhage et al., 2006) yielded at best a seizure prediction performance of 42% sensitivity and an average of 3 false positives per day. These EEG data had been acquired from intracranial grid-, strip-, and depth-electrodes at a 256 Hz sampling rate, and digitized to 16 bit by an analogue-to-digital converter. In the source dataset, a certified epileptologist had previously restricted the EEG dataset to 6 channels, from three focal electrodes (1-3) involved in early ictal activity, and three electrodes (4-6) not involved during seizure spread.

Each of the patients' EEG recordings from the Freiburg database contained between 2 and 6 seizures and at least 50 min of pre-ictal data for most seizures, as well as approximately 24 hours of EEG-recordings without seizure activity and spanning the full wake-sleep cycle. We set apart preictal samples preceding the last 1 or 2 seizures (depending on that patient's total number of seizures) and 33% of the interictal samples: these were testing (out-of-sample) data. The remaining samples were training (in-sample) data. Further 10% or 20% of training data were randomly selected for cross-validation. The training procedure (section 2.4) would be stopped either after a fixed number of iterations, or we would use cross-validation data to select the best model (and stop the training procedure prematurely). In summary, we trained the classifiers on the earlier seizures and on wake-sleep interictal data, and evaluated these same classifiers on later seizures and on different wake-sleep interictal data.

We further applied Infinite Impulse Response (IIR) elliptical filters, using code from EEGLab (Delorme and Makeig, 2004) to clean some artifacts: a 49-51Hz band-reject 12^{th}-order filter to remove power line noise, a 120Hz cutoff low-pass 1^{st}-order filter, and a 0.5Hz cutoff high-pass 5^{th}-order filter to remove the dc component. All data samples were scaled on a per patient basis, to either zero mean and unit variance (for logistic regression and convolutional networks) or between -1 and 1 (for support vector machines). At this stage, let us denote $x_i(t)$ the time series representing the i-th channel of the preprocessed EEG.

2.2. Extraction of bivariate features

A *bivariate feature* is a measure of a certain relationship between two signals. Bivariate features presented in this section and used in this study have the following common points:

(a) Bivariate features are computed on 5s windows (N=1280 samples at 256Hz) of any two EEG channels x_a and x_b.
(b) For EEG data consisting of M channels, one computes features on $M \times (M-1)/2$ pairs of channels (e.g. 15 pairs for M=6 in the Freiburg EEG dataset).

Some features are also specific to a frequency range.

We investigated in our study 6 types of bivariate features known in the literature, and which we explain in details in Appendix A. The simplest feature was cross-correlation C, a linear measure of dependence between two signals (Mormann et al., 2005) that also allows fixed delays between two spatially distant EEG signals to accommodate potential signal propagation. The second feature was nonlinear interdependence S (Arnhold et al., 1999), which measures the distance in state-space between the trajectories of two EEG channels. The third feature was dynamical entrainment *DSTL* (Iasemidis et al., 2005) i.e. the difference of short-term Lyapunov exponents, based on a common measure of the chaotic nature of a signal. Finally, the last three features that we investigated were based on phase synchrony (Le Van Quyen et al., 2001; Le Van Quyen et al., 2005). First, frequency-specific and time-dependent phase $\varphi_{a,f}(t)$ and $\varphi_{b,f}(t)$ were extracted from the two respective EEG signals $x_a(t)$ and $x_b(t)$ using Wavelet Transform. Then, three types of statistics on the difference of phases between two channels were made: phase-locking synchrony *SPLV*, entropy *H* of the phase difference and coherence *Coh*.

2.3. Aggregation of bivariate features into spatio-temporal patterns

We define in this article a *pattern* as a structured collection of features described in previous section. A pattern groups features along the spatial, time and frequency dimensions. A simplistic analogy is that a feature is like the color of a pixel at a specific location in an image. In this article, we formed 2D patterns by aggregating features from all 15 pairs of channels (across rows) and over several consecutive time frames (across columns). Specifically, we formed 1min or 5min-long patterns of 12 or 60 frames respectively. In the case of frequency-based features, we also stacked patterns, row-wise and from all frequency ranges into one pattern. The dimensionality of the feature patterns ranged from 180 (e.g. cross-correlation on 1min windows, Fig. 1), to 6300 (e.g. wavelet phase-locking synchrony on 5min windows). As mentioned in the Results section 3.1, 5min-long patterns achieved superior results to 1min-long patterns, and the article therefore reports seizure prediction results on 5min-long patterns only.

Throughout the article, we denote as $\mathbf{y}_t$ a pattern at time t (i.e. a sample of bivariate features), and z_t the associated label (-1 for preictal, 1 for interictal). $\mathbf{y}_t$ can either be one long vector or a matrix indexed by time and by channel pair and frequency band.
[Insert Figure 1]

2.4. Machine learning classification of patterns of bivariate features

Bivariate patterns $\mathbf{y}_t$ described in previous sections and representing a "snapshot" of EEG synchronization around time t were input into a decision system that would classify them as preictal or interictal. The parameters of that classifier were learned on the training subset of the dataset using machine learning. Let us note z_t the label of pattern $\mathbf{y}_t$ (-1 for preictal, 1 for interictal) and $\overline{z}_t$ the output of the classifier. Although we used three different types of classifiers, with their respective machine learning algorithms, all training algorithms had in common minimizing, for every training sample $\mathbf{y}_t$, the error between output $\overline{z}_t$ and target z_t. The error between the output and the target is one term of the loss function: we explain in section 2.5 the second term (regularization). Finally, and most importantly, test data were set apart during the training phase: in other words, we validated the performance of the classifiers on out-of-sample data.

The first classifier that we tried was logistic regression, i.e. a linear classifier parameterized by weights **w** and bias b, and optimized by minimizing a loss function with stochastic gradient descent (Rumelhart et al., 1986; LeCun et al., 1998a). In a nutshell, this linear classifier performs a dot product between pattern $\mathbf{y}_t$ and weight vector **w**, and adds the bias term b (see Appendix B.1). In the loss function, an additional L1-norm penalty on the weight vector helps selecting a sparse subset of weights and enables feature selection (see sections 2.5 and 3.4).

The second classifier that we tried was built on convolutional networks (LeCun et al., 1998a). Convolutional networks are trainable, multi-layer, non-linear systems that are specifically designed to extract and classify high-dimensional patterns from images or multivariate time-series. They can be seen as multi-layer neural networks in which each layer is a bank of finite-impulse response filters followed by point-wise sigmoid squashing functions. One could make a parallel between convolutional networks and an extremely simplified model of the V1 visual cortex, because each layer processes locally inputs from the previous layer, and because this processing is replicated over the input pattern. All the layers are trained simultaneously using a version of the back-propagation learning algorithm. They can learn low-level features and high-level representations in an integrated manner. Their main advantage is that they can learn optimal time-invariant local feature detectors from input matrix $\mathbf{y}_t$ (which is indexed by time) and thus build representations that are robust to time shifts of specific feature motifs. This technique has already been applied to raw EEG data (Mirowski et al., 2007).
[Insert Figure 2]

We used a specific convolutional network architecture similar to LeNet5 (LeCun et al., 1998a) (Fig. 2) consisting in a stack of 5 layers of neural network connections (also called weight layers). Weight layers 1, 3 and 5 were convolutional layers, and layers 2 and 4 were mere subsampling layers. Each layer would compute a weighted sum over a "local" subset of inputs. Let p be the number of pairs of channels (15) times the number of frequency bands (1 or 7). Then, 12-frame patterns (i.e. 1min-long) were processed in the following way: the 1st layer would perform 5 different 5-point convolutions over the time dimension; the 3rd layer would perform 5 different 3-point convolutions over time and p-point convolutions over all channels and frequency bands; and the 5th layer was fully connected between all its inputs (i.e. the outputs the 4th layer) and the 2 output nodes (one for "preictal" and one for "interictal"). The 2nd and 4th layer merely averaged two consecutive time points (i.e. subsampled in time). 60-frame patterns (i.e. 5min-long) were processed slightly differently: the 1st layer would perform 5 different 13-point convolutions over the time dimension; and the 3rd layer would perform 5 different 9-point convolutions over time and p-point convolutions over all channels and frequency bands. As mentioned in the Results section 3.1, 5min-long patterns achieved superior seizure prediction results to 1min-long patterns, and the latter 5min-long architecture is the one for which we report seizure prediction results. As a side remark, we chose to have 2 output nodes because it enabled an asymmetric learning that penalized more false positives (false preictal alarms) than false negatives (missed preictal alarms).

Finally, we compared the two neural network architectures (logistic regression, linear, and convolutional networks, highly non-linear) with a third type of classifiers, called Support-Vector Machines (SVM) (Cortes and Vapnik, 1995). SVM are pattern matching-based classifiers that compare any input pattern $\mathbf{y_t}$ to a set of support vectors $\mathbf{y_s}$. We used in this study standard SVMs with Gaussian kernels, and optimized the Gaussian standard deviation hyper-parameter γ and regularization hyper-parameter C selected by cross-validation over a grid of values (see Appendix B.2). SVMs were implemented using the LibSVM library (Chang and Lin, 2001).

2.5. Feature selection

Training algorithms for neural network classifiers such as logistic regression and convolutional networks enable to easily add a regularization term on the weights (parameters) **w**. Typically, regularization consists in minimizing the norm of vector **w**. Specifically, we added an L1-norm (sum of absolute values) regularization term to the loss function (Eq. 2) that is minimized during gradient descent. We typically used values of 0.001 for lambda. This L1 term uniformly pulls the weights towards zero during gradient-based optimization. As a consequence, only a subset $\{w_i\}$ of these weights "survive", and the final solution **w*** contains a minimal set of weights that simultaneously minimizes the error on the training dataset. This L1-norm weight selection is also called the "LASSO" algorithm (Tibshirani, 1996). We used it as a task-specific way to select features, as opposed to a task-agnostic selection of features prior to the training algorithm. In other words, the only non-zero (or non-negligible) features are the ones that specifically discriminate between interictal and preictal patterns of that given patient.

After training the neural network, a sensitivity analysis on the inputs was performed to see which features were important for the discrimination. In the case of Logistic Regression, we simply looked at individual weights w_i. For convolutional networks, we back-propagated the gradients obtained for each testing sample onto the inputs, and then summed the squares of these gradients on inputs.

3. Results

This section shows how high-dimensional spatio-temporal patterns of bivariate features allow for better separation between interictal and preictal recordings (section 3.1). We then report results obtained with machine learning-trained classifiers: for each patient, we could find at least one combination of methods that would predict all the test seizures of the given patient without false alarm; one specific combination (convolutional networks with wavelet coherence) worked for 15 patients out of 21 and achieved 71% sensitivity and 0 false positives (sections 3.2 and 3.3). We explain how neural network-based classifiers enable a-posteriori selection of channels and frequency bands relevant for seizure prediction (section 3.4). Finally, we investigated whether there was a link between seizure prediction performance and the patient's condition (section 3.5).

3.1. Increased separability of patterns instead of individual features

First, we tried to compare the discriminative power of patterns of features as opposed to individual features. As defined in Methods section 2.3, a pattern aggregates features across successive time frames and over all pairs of channels. We believed that there was no need to prove the utility of considering information from all pairs of EEG channels, instead of taking into account only one single pair of channels or just an average value across all channels, as in (Mormann et al., 2005). An image processing analogy of the latter methods would be to try to detect faces on an image by looking at the average color of all the pixels in the image or by looking at the color of a few pixels only. By consequence, we limited our analysis to a comparison between patterns across channels vs. patterns across time and channels, and this way we assessed the benefit of adding the time dimension to patterns.

We performed a Principal Component Analysis (PCA) of patterns of bivariate features with different lengths of aggregation across time. Namely, we investigated purely spatial patterns (1 single time-frame, where features had been computed on a 5s window), short spatio-temporal patterns (12 time-frames covering 1min) and long spatio-temporal patterns (60 time-frames covering 5min). To account for the variability between patients, we performed this PCA individually for each patient and for each type of feature (cross-correlation, nonlinear interdependence, difference of Lyapunov exponents, phase-locking value, wavelet coherence and entropy of phase difference). We visually inspected the projections of all the interictal, preictal and ictal/postictal patterns along their first two principal components. These top PCA components corresponded to the directions of highest variability of the feature values. We observed that the 2D projections of preictal and interictal 1-frame patterns overlapped considerably, more than the projections of 12-frame or 60-frame patterns. An illustration of this phenomenon is given on Figure 3, which shows the PCA projection of patterns of phase-locking values for patient 1: whereas it is difficult to see a boundary between the interictal and preictal clusters of 1-frame patterns (without time aggregation), the boundary becomes more apparent for time-aggregated 12-frame patterns, and even more apparent for 60-frame patterns.
[Insert Figure 3]

This intuitive observation about spatio-temporal patterns was later empirically confirmed, since seizure prediction performance was superior for 5min-long patterns than for 1min-long patterns. More precisely, 1min-long patterns of features could predict seizures without false positives only in 8 patients out of 21 (i.e. that there was at least one such combination of 1min-long pattern feature and classifier for each patient), and the best combination (1min-long pattern of wavelet coherence with SVM classifier) predicted seizures perfectly only in 4 patients out of 21. For this reason, the next section reports only results obtained with 5min-long patterns.

3.2. Classification results

As introduced earlier, we investigated in this seizure prediction study different combinations of one type of feature patterns (cross-correlation *C*, nonlinear interdependence *S*, difference of Lyapunov exponents *DSTL*, phase-locking synchrony *SPLV*, entropy of phase difference *H* and distribution or wavelet coherence *Coh*) and one type of classifier (Logistic Regression *log reg*, convolutional networks *conv-net* or *SVM*). For each patient, there were 18 possible combinations of 6 types of features and 3 types of classifiers; however, because the *DSTL* feature did not yield good results with SVM classifiers, we discontinued evaluating the *DSTL* feature with the two other classifiers, and for this reason report results for only 16 combinations in Tables 1 and 4.

Because the goal of seizure prediction is the epileptic patient's quality of life, we report the following classification performance results in terms of false alarms per hour and sensitivity, i.e. number of seizures where at least one preictal sample is classified as such.

For each patient, at least one of our combined methods could predict all the test seizures, on average 60 min before the onset and with no false alarm. On the other hand, not all combinations of feature and classifier yielded perfect prediction: to the contrary, many combinations of feature and classifier failed the seizure prediction task either because there were more than 0.25 false positives per hour (i.e. more than 3 false positives per day) or because the seizure was not predicted. The main limitation of our patient-specific multiple-method approach lies in the lack of a criterion for choosing the best combination of methods for each patient, other than cross-validating each method on long EEG recordings.

The best results were obtained using patterns of wavelet coherence *Coh* features classified using convolutional networks (zero false positive and all test seizures predicted on 15 patients out of 21, i.e. 71% sensitivity), then patterns of phase-locking synchrony *SPLV* using a similar classifier (13 patients out of 21, i.e. 62% sensitivity). Both

Coh patterns classified using logistic regression *log-reg*, as well as patterns of phase difference entropy *H* classified using *conv-net* predicted all test seizures without false positive on 11 patients (52% sensitivity). Finally, *SPLV* classified using *log-reg* and nonlinear interdependence *S* classified using *conv-net* worked without false alarm on 10 patients (48% sensitivity). Table 1 summarizes the above sensitivity results. Results on our best classifier and features outperform previously published 42% sensitivity and 3 false positives per day on the Freiburg dataset.

Irrespective of the EEG feature, convolutional networks achieved a zero-false alarm seizure prediction on 20 patients out of 21, compared to 11 only using SVM (however, good results were obtained for patient 5, contrary to convolutional networks). Surprisingly, the linear classification boundary of logistic regression enabled perfect seizure prediction on 14 patients.

Tables 1-4 recapitulate how many patients had "perfect prediction" of their test seizures, i.e. zero-false alarm during interictal phases and at least one alarm during pre-ictal phases, given a combination of feature and classifier (see Table 1), as well as given each type of feature pattern (see Table 2) or classifier (see Table 3). Table 4, organized by patient, feature type and classifier, displays the frequency of false alarm per hour, and how many minutes ahead were the one or two test seizures predicted. Figure 4 shows the times of preictal alarms for each patient, achieved using the best patient-specific method.
[Insert Figure 4]

It has to be noted that both for convolutional networks and logistic regression, 100% of training samples (patterns of bivariate features) were correctly classified. The only exceptions were patients 17, 19 and 21, where we allowed a larger penalty for false positives than for false negatives. On these three patients we obtained only some false negatives and no false positive on the training dataset, while managing to predict all train seizures.

We did not evaluate the classification results obtained by a combination of all 6 types of features because of two reasons. First, combining a large number of features would yield very high-dimensional inputs. Secondly, the computational cost of the features could make it impractical to compute many types of features at once in a runtime setting (see section 4.3).

3.3. Verification of EEG for artifacts

Analysis of Table 4 reveals that for a given patient and a given test seizure, most feature-classifier combinations share the same time of first preictal alarm. The simple justification is that most of these time-aligned first preictal alarms also correspond to the beginning of the preictal recording. Going back to the original raw EEG, and with the help of a trained epileptologist, we performed additional sanity checks. First, we verified that there were no recording artifacts that would have helped differentiate interictal from preictal EEG, and second, we verified that EEG segments corresponding to the pattern at the time of the first preictal alarm were not artifacts either. Through visual inspection, we compared several EEG segments: at the time of the first preictal alarm, right before the seizure and a few randomly chosen 5min segments of normal interictal EEG.

We noticed that there seemed to be high frequency artifacts on preictal recordings for patients 4 and 7, and that no such artifacts were visible on interictal recordings. However, for all other patients, short artifacts were indiscriminately present on both preictal and interictal segments. Moreover, we observed what appeared to be sub-clinical events or even seizures on the preictal EEG of patients 3, 4, 6, and 16: we hypothesize that these sub-clinical events might have been (correctly) classified by our system as preictal alarms.

3.4. Feature selection results

The additional functionality of our seizure prediction algorithm is the feature selection mechanism detailed in Methods section 2.5. This feature selection could help narrowing down the set of input bivariate features. When learning the parameters of the logistic regression or convolutional network classifiers (but not the support vector machine), weight parameters are driven to zero thanks to L1-norm regularization, and the few remaining non-zero parameters are those that enable successful classification on the training, cross-validation and testing datasets. We performed a sensitivity analysis on individual classifier inputs and identified which couples of EEG channels were discriminative between preictal and interictal patterns. We observed that out of the 15 pairs of channels, generally only 3 or 4 pairs were actually necessary for seizure prediction when using non-frequency-based features (cross-

correlation C and nonlinear interdependence S). Similarly, only a subset of frequency bands was discriminatory for seizure prediction classification when using wavelet-analysis based measures of synchrony (phase-locking $SPLV$, coherence Coh or entropy H). Interestingly, that subset always contained high frequency synchronization features (see Figure 5).
[Insert Figure 5]

3.5. Prediction results vs. patient condition

Finally, we investigated whether the epileptic patient's condition can impact the seizure prediction task, and compared the number of combinations of feature and classifier that achieved perfect seizure prediction performance, versus several characteristics of the patients. These characteristics, summarized for the Freiburg dataset in table 2 of (Maiwald et al., 2004), included the Engel classification of epilepsy surgery outcome (I through IV), the types of epilepsy (simple partial, complex partial, or generalized tonic-clonic) and the localization of the epileptogenic focus (hippocampal or neo-cortical). Like in the rest of our study, we defined perfect seizure prediction as having no false positives and all test seizures predicted for a given patient. We did not observe any significant correlation between the patient condition and the number of successful feature-classifier combinations for that same patient. For instance, only 3 combinations of feature and classifier worked flawlessly for patient 6, who was seizure-free after surgery, whereas most combinations of feature and classifier worked perfectly for patients 2 and 12, whose condition did not improve much or even worsened after surgery. Patients 3 and 10 presented the opposite case. Therefore, we cannot draft at that stage of our investigations any hypothesis, neither about the applicability of our seizure prediction method to specific cases of epilepsy, or about how well it predicts the surgery outcome. It seems that albeit being patient-specific, our method is not condition-specific, and should be applied individually to predict seizures in various types of localized epilepsies.

4. Discussion

As detailed in the Results section, this article introduced a new approach to seizure prediction. We presented machine learning techniques that outperform previous seizure prediction methods, as our best method achieved 71% sensitivity and 0 false positives on the Freiburg dataset. Such results were enabled by our pattern recognition approach applied to spatio-temporal patterns of EEG synchronization features. The following section discusses the uniqueness and advantages of pattern recognition approaches to seizure prediction, running-time considerations; we also explain the need for further validation on other datasets, and for an alternative to our current binary classification approach.

4.1. Choice of linear or nonlinear features

An important task for seizure prediction is the choice of type of EEG features. Generally, among bivariate (or multivariate) features, one can make two distinct assumptions about the nature of the model underlying the observed EEG; indeed, EEG can either be viewed as a realization of a noise-driven linear process, or as an observation of a non-linear, possibly chaotic, dynamical system (Stam, 2005). The linear or nonlinear hypotheses imply different sets of mathematical tools and measurements to quantify EEG.

On one hand, linear methods for EEG analysis assume that over short durations of time, the EEG time series are generated by a system of linear equations with superimposed observation noise. Although this hypothesis is restrictive, maximum cross-correlation (Mormann et al., 2005), was shown to achieve quite a good discrimination performance between interictal and preictal stages.

The other assumption about the EEG signal is its nonlinearity. Although deterministic by nature, systems of nonlinear differential equations can generate highly complex or even unpredictable ("chaotic") time series. The trajectory or "attractor" of the generated sequence of numbers can be extremely sensitive to initial conditions: any perturbation in those conditions can grow at an exponential rate along the attractor. Nonlinear, chaotic, dynamical systems have become a plausible model for many complex biological observations, including EEG waveforms (Stam, 2005). Even if not all the variables of a chaotic system are observed, one can theoretically reconstruct the original chaotic attractor, thanks to time-delay embedding of the time series of the limited subset of observed variables, assuming the right embedding dimension and time delay (Takens, 1981). Similarly, although one cannot know all the variables behind the chaotic dynamical system of the neuronal networks of the brain, one can try to reconstruct, in the state-space, attractors from time-delay embedded observed EEG.

As described in Results section 3.2, this study seems to discard the difference of Lyapunov exponents, and tends to favor nonlinear interdependence and wavelet-analysis-based statistics of synchrony. From the analysis of seizure prediction results on 21 patients, there was however no specific EEG feature that would work for every patient. Moreover, the superiority of nonlinear features over linear features could not be demonstrated in other comparative studies (Mormann et al., 2005).

4.2. Comparison with existing threshold-based seizure prediction methods

Most current seizure prediction techniques resort to a simple binary threshold on a unique EEG feature. Such an approach has two major limitations. First, in order to ensure the predictability, and in absence of testing data, binary thresholds require validation using the Seizure Time Surrogates method (Andrzejak et al., 2005). Besides, simple statistical classification not only uses simplistic linear decision boundaries, but also requires reducing the number of variables. A typical shortcoming of an ill-designed binary classification algorithm is illustrated in (Jerger et al., 2005). Hilbert-based phase-locking synchrony is computed for all frequencies without prior band-pass filtering, and cross-correlation is computed for zero delay only. Bivariate measurements from several channels are collapsed to single values. Finally, the final decision boundary is a simple line in a 2D space covered by the two bivariate measurements. Unsurprisingly, the seizure prediction performance of (Jerger et al., 2005) is very weak. We believe that the explanation for such unsatisfying results is that relevant seizure-discriminative information has been lost as the dimensionality of the features has been reduced to two.

Let us now make a crude analogy between the feature derived from one or two EEG signals around time *t*, and the value of a "pixel" in a "movie" at time *t*. Most current seizure prediction methods look at "individual pixels" of the EEG-based feature "image" instead of looking at the "full picture", i.e. the relationship between the "pixels" within that "image"; moreover they forego the dynamics of that "movie", i.e. do not try to capture how features change over time. By contrast, our method learns to recognize patterns of EEG features.

4.3. Running-time considerations

The patent-pending system described in this article (Mirowski et al., patent application filed in 2009) does not require extensive computational resources. Although our seizure prediction method is still under evaluation and refinement, we consider in this section whether it could be implemented as real-time dedicated software on an embedded computer connected to the patient's intracranial EEG acquisition system.

The whole software process, from raw numerical EEG to the seizure prediction alarm can be decomposed in 3 stages: EEG preprocessing, feature computation and pattern classification. The first stage (EEG preprocessing) is implemented by 4 standard Infinite Impulse Response (IIR) filters that have negligible runtime even in real-time signal processing. The third stage (pattern classification) is done only every minute or every 5 minutes (depending on the pattern size) and corresponds to a few matrix-vector multiplications and simple floating-point numerical operations (addition, multiplication, exponential, logarithm), involving vectors with a few thousand dimensions. The most computationally expensive part is the training (parameter fitting) of the classifier, but it is done offline and thus does not affect the runtime. The second stage (feature computation from EEG) is also relatively fast: it takes in the order of seconds to process a 5 minute-long window of 6-channel EEG and extract features such as wavelet analysis-based synchrony (*SPLV*, *Coh* or *H*), nonlinear interdependence *S* or cross-correlation *C*. However, since the 5min patterns are not overlapping, stage 2 is only repeated every minute or 5 minutes (like stage 3). It has to be noted that this running time analysis was done on a software prototype that could be further optimized for speed.

The software for computing features from EEG was implemented in Matlab™ and can be run under its free open-source counterpart, Octave™. Support vector machine classification was performed using LibSVM™ (Chang and Lin, 2001) and its Matlab/Octave interface. Convolutional networks and logistic regression were implemented in Lush™, an open-source programming environment (Bottou and LeCun, 2002) with extensive machine learning libraries.

4.4. Overcoming high number of EEG channels through feature selection

In addition to real-time capabilities during runtime, the training phase of the classifier has an additional benefit. Our seizure prediction method enables further feature selection through sensitivity analysis, namely the discovery of subsets of channels (and if relevant, frequencies of analysis), that have a strong discriminative power for the preictal versus interictal classification task.

This capability could help the system cope with a high number of EEG channels. Indeed, the number of bivariate features grows quadratically with the number of channels *M*, and this quadratic dependence on the number of EEG channels becomes problematic when EEG recordings contain many channels, e.g. one or two 64-channel grids with additional strip electrodes. This limitation might slow down both the machine learning (training) and even the runtime (testing) phases. Through sensitivity analysis, one could narrow down the subset of EEG channels necessary for a good seizure prediction performance. One could envision the following approach: first, long and slow training and evaluation phases using all the EEG channels, followed by channel selection with respect to their discriminative power, and a second, faster, training phase, with, as end product, a seizure prediction classifier running on a restricted number of EEG channels. The main advantage of this approach is that the channel selection is done a posteriori with respect to the seizure prediction performance, and not a priori as in previous studies (D'Alessandro et al., 2003; Le Van Quyen et al., 2005). In our method, the classifier decides by itself which subset of channels is the most appropriate.

4.5. Statistical validity

One of the recommended validation methods for seizure prediction algorithms is Seizure Time Surrogates (STS) (Andrzejak et al., 2005). As stated in the introduction, STS is a necessary validation step required by most current statistical seizure prediction methods, which use all available data to find the boundary thresholds (in-sample optimization using the ROC curve) without proper out-of-sample testing. STS consists in repeatedly scrambling the preictal and interictal labels and checking that the subsequent fake decision boundaries are statistically different from the true decision boundary.

Such surrogate methods are however virtually unknown in the abundant machine learning literature and its countless applications, because the validation of machine learning algorithms relies instead on the Statistical Learning Theory (Vapnik, 1995). The latter consists in regularizing the parameters of the classifier (as described in section 2.5), and in separating the dataset into a training and cross-validation set for parameter optimization, and a testing set that is unseen during the optimization phase (as described in details in section 2.1).

On one hand, the use of a carefully designed separate and unseen testing set verifies that the classifier works well in the general case, within the limits of the testing dataset. Given the long time required to train a machine learning classifier, such an approach is less computationally expensive than surrogate methods.

On the other hand, the regularization permits to choose, among the infinity of configurations of parameter values (e.g. the "synaptic" connection weights of a convolutional network or the matrix of logistic regression), the "simplest" one, generally satisfying a criterion such as choosing the feasible parameter vector with the smallest norm. The regularized classifier does not overfit the training dataset (e.g. it does not learn the training set patterns "by heart") but has instead good generalization properties, i.e. a low theoretical error on unseen testing set patterns. Moreover, regularization enables to cope with datasets where the number of inputs is greater than the number of training instances. This is for instance the case with machine-learning based classification of biological data, where very few micro-array measurements (each micro-array being a single instance in the learning dataset) contain tens of thousands of genes or protein expression levels.

Nevertheless, let us devise the following combinatorial verification of the results. Since our study focused on non-overlapping 5min-long patterns, and since our patient-specific predictors would ignore the time stamp of each pattern, we consider a random predictor that gives independent predictions every 5 minutes on one patient's data, and emits a preictal alarm with probability p. Each patient's recording consists of at least 24h of interictal data (out of which, at least 8h are set apart for testing), which contain, respectively, at least n_i=288 or n_i=96 patterns, and m preictal recordings of at most 2h each (out of which, one or two are set apart for testing), with at most n_p=24 patterns per preictal recording. Using binomial distributions, we can compute the probability: $A(p) = f(0; n_i, p)$ of not emitting any alarm during the interictal phase, as well as the probability of emitting at least one alarm before each seizure: $B(p) = 1 - f(0; n_p, p)$. The probability of predicting each seizure of a patient, without false alarm, is a function of the predictor's p: $C(p) = A(p)B(p)^m$.

After maximization with respect to the random predictor "firing rate" p, the optimal random predictor could predict, without false alarm during the 8h of out-of-sample interictal recording, one test seizure with over 8% probability and two test seizures with over 2% probability. In our study, we evaluated 16 different combinations of features and classifiers. If one tried 16 different random predictors for a given patient, and using again binomial distributions, the expected number of successful predictions would be computed as 1.3 for one test seizure, and 0.4 for two test seizures. Considering that the random predictor also needs to correctly classify patterns from the training and cross-validation dataset, in other words to correctly predict the entire patient's dataset (this was the case of the successful classifiers reported in Table 1), then, by a similar argument, this expected number of successful predictions goes down from 0.05 for a 2-seizures dataset to 10^{-4} for a 6-seizures dataset.

Although the above combinatorial analysis only gives an upper bound on the number of "successful" random predictors for a given patient, it motivates a critical look at the results reported in Table 4. Specifically, seizure prediction results obtained for certain patients where only 1 or 2 classifiers (out of 16) succeeded in predicting without false alarm should be considered with reserve (such is the case for patients 13, 17, 19 and 21).

4.6. Limitations of binary classification for seizure prediction

A second limitation of our method lies in our binary classification approach. When attempting seizure prediction, binary classification is both a simplification and an additional challenge for training the classifier. In our case, 2-hour-long preictal periods imply a 2-hour prediction horizon, which naturally drives the sensitivity up. At the same time, the classifier is forced to consider patterns as remote as 2 hours prior to a seizure as "preictal", whereas there might be no difference between such a pattern and an interictal pattern.

For this reason, we suggest, as further refinements of our method, to replace the binary classification by regression. For instance, one could regress a function of the inverse time to the seizure, taking a value of 0 away from a seizure then continuously increasing up to a value of 1 just before the seizure. Such an approach would naturally integrate a seizure prediction horizon and could be considered a variation of the Seizure Prediction Characteristic (Winterhalder et al, 2004) formulated into a machine learning problem.

4.7. Importance of long, continuous EEG recordings

As suggested in the above discussion about testing datasets, one could see a third potential limitation of the EEG Freiburg database: indeed, while it provides, for each patient, with at least 24 hours of interictal and a few hours of preictal, ictal and postictal recording, it does not cover the whole duration of the patient monitoring, and there are sometimes gaps of several days between the preictal segments and the interictal segments (e.g. this is the case for patient 12). One could therefore argue that what has been picked by our EEG classification algorithm was not a preictal vs. interictal signal, but a large time-scale physiological, medical or acquisition artifact. However, there are also patients where preictal and interictal segments are interleaved. An example is patient 8, where one continuous EEG recording spans a long interictal segment and then a preictal segment, including the transition from interictal to preictal. As illustrated on Figure 6, our algorithm succeeded in raising several preictal alarms before the test seizure, without emitting any false alarms.

Unfortunately, no information about the patient's circadian variations, level of medication, or state of vigilance is available in the 21-patient Freiburg dataset; it is therefore necessary for our method to be further validated on different datasets. While our algorithm passed certain sanity checks (e.g. patient 8 in the Freiburg dataset), we reiterate the guideline (Lehnertz et al., 2007) for seizure prediction studies, which stipulates that datasets need to contain long, continuous and uninterrupted EEG recordings so that one can prove that a seizure prediction algorithm works round the clock.

Acknowledgements

This research has been funded by FACES (Finding A Cure for Epilepsy and Seizures). The authors wish to thank

Dr. Nandor Ludvig and Dr. Catherine Schevon for useful discussion and helpful comments.

Appendix A. Bivariate features computed on the EEG

A.1. Maximal cross-correlation

Cross-correlation (C) values $C_{i,j}(\tau)$ between pairs (x_i,x_j) of EEG channels $x_i(t)$ and $x_j(t)$ are computed at delays τ ranging from -0.5s to 0.5s, in order to account for the propagation and processing time of brainwaves, and only the maximal value of such cross-correlation values is retained (Mormann et al., 2005), as in:

(A1) $$C_{a,b} = \max_{\tau}\left\{\left|\frac{C_{a,b}(\tau)}{\sqrt{C_a(0)\cdot C_b(0)}}\right|\right\} \text{ where } C_{a,b}(\tau) = \begin{cases} \frac{1}{N-\tau}\sum_{t=1}^{N-\tau} x_a(t+\tau)x_b(\tau) & \tau \geq 0 \\ C_{b,a}(-\tau) & \tau < 0 \end{cases}$$

and N is the number of time points within the analysis window (N=1024 in this study).

A.2. Nonlinear interdependence

Nonlinear interdependence (S) is a bivariate feature that measures the Euclidian distance, in reconstructed state-space, between trajectories described by two EEG channels $x_a(t)$ and $x_b(t)$ (Arnhold et al, 1999).
First, each EEG channel $x(t)$ is time delay-embedded into a local trajectory $\mathbf{x}(t)$ (Stam, 2005), using delay τ=6 (approximately 23ms) and embedding dimension d=10, as suggested in (Arnhold et al., 1999; Mormann et al., 2005):

(A2) $$\mathbf{x}(t) = \{x(t-(d-1)\tau),\ldots,x(t-\tau),x(t)\}.$$

After time-delay embedding of EEG waveforms into respective sequences of vectors $\mathbf{x_a}(t)$ and $\mathbf{x_b}(t)$, one computes a non-symmetric statistic $S(x_i|x_j)$:

(A3) $$S(x_a|x_b) = \frac{1}{N}\sum_{t=1}^{N}\frac{R(t,x_a)}{R(t,x_a|x_b)},$$

where the distance of $\mathbf{x_a}(t)$ to its K nearest neighbors in state space is defined as (A3) and the distance of $\mathbf{x_a}(t)$ to the K nearest neighbors of $\mathbf{x_b}(t)$ in state space is defined as (A4):

(A3) $$R(t,x_a) = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{x}_a(t)-\mathbf{x}_a(t_k^a)\right\|_2^2$$

(A4) $$R(t,x_a|x_b) = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{x}_a(t)-\mathbf{x}_a(t_k^b)\right\|_2^2,$$

where:

(A5) $\{t_1^a, t_2^a, \ldots, t_K^a\}$ are the time indices of the K nearest neighbors of $\mathbf{x_a}(t)$ and

(A6) $\{t_1^b, t_2^b, \ldots, t_K^b\}$ are the time indices of the K nearest neighbors of $\mathbf{x_b}(t)$.

In this research, K=5. The nonlinear interdependence feature is a symmetric measure:

(A7) $$S_{a,b} = \frac{S(x_a|x_b)+S(x_b|x_a)}{2}.$$

A.3. Difference of short-term Lyapunov exponents

The difference of short-term Lyapunov exponents (DSTL), also called dynamical entrainment, is based on chaos theory (Takens, 1981). First, one estimates the largest short-time Lyapunov coefficients STL_{max} on each EEG channel $x(t)$, by using moving windows on time-delay embedded time-series $\mathbf{x}(t)$. STL_{max} is a measure of the average exponential rates of growth of perturbations $\boldsymbol{\delta}\mathbf{x}(t)$ (Winterhalder et al., 2003; Iasemidis et al., 1999):

(A8) $$STL_{\max}(\mathbf{x}) = \frac{1}{N\Delta t}\sum_{t=1}^{N}\log_2\left|\frac{\delta\mathbf{x}(t+\Delta t)}{\delta\mathbf{x}(t)}\right|,$$

where Δt is the time after which the perturbation growth is measured. Positive values of the largest Lyapunov exponent are an indication of a chaotic system, and this exponent increases with the unpredictability. In this research, where EEG is sampled at 256Hz, time delay is τ=6 samples or 20ms, embedding dimension is d=7 and evolution time Δt=12 samples or 47ms, as suggested in (Iasemidis et al., 1999, 2005). The bivariate feature is the difference of STL_{max} values between any two channels:

(A9) $$DSTL_{a,b} = |STL_{max}(x_a) - STL_{max}(x_b)|.$$

A.4. Wavelet-based measures of synchrony

Three additional frequency-specific features are investigated in this study, based on wavelet analysis measures of synchrony (Le Van Quyen et al., 2001, 2005). First, frequency-specific and time-dependent phase $\varphi_{i,f}(t)$ and $\varphi_{j,f}(t)$ are extracted from the two respective EEG signals $x_i(t)$ and $x_j(t)$ using wavelet transform. Then, three types of statistics on these differences of phase are computed: phase-locking synchrony *SPLV* (Eq. A10), entropy *H* of the phase difference (Eq. A11) and coherence *Coh*. For instance, phase-locking synchrony *SPLV* at frequency *f* is:

(A10) $$SPLV_{a,b}(f) = \left|\frac{1}{N}\sum_{t=1}^{N} e^{i[\phi_{a,f}(t)-\phi_{b,f}(t)]}\right|$$

(A11) $$H_{a,b}(f) = \frac{\ln(M) - \sum_{m=1}^{M} p_m \ln(p_m)}{\ln(M)},$$

where $p_m = \Pr[(\varphi_{a,f}(t) - \varphi_{a,f}(t)) \in \Phi_m]$ is the probability that the phase difference falls in bin m and M is the total number of bins.

Synchrony is computed and averaged in 7 different frequency bands corresponding to EEG rhythms: delta (below 4Hz), theta (4-7Hz), alpha (7-13Hz), low beta (13-15Hz), high beta (14-30Hz), low gamma (30-45Hz) and high gamma (65-120Hz), given that the EEG recordings used in this study is sampled at 256Hz. Using 7 different frequency bands increased the dimensionality of 60-frame, 15-pair synchronization patterns from 900 to 6300 elements.

Appendix B. Bivariate features computed on the EEG

B.1. Logistic regression

Logistic regression is a fundamental algorithm for training linear classifiers. The classifier is parameterized by weights **w** and bias *b* (Eq. B1), and optimized by minimizing loss function (Eg. B2). In a nutshell, this classifier performs a dot product between pattern $\mathbf{y}_t$ and weight vector **w**, and adds the bias term *b*. The positive or negative sign of the result (Eg. B1) decides whether pattern $\mathbf{y}_t$ is interictal or preictal. By consequence, this algorithm can be qualified as a linear classifier: indeed, each feature $y_{t,i}$ of the pattern is associated its own weight w_i and the dependency is linear. Weights **w** and bias *b* are adjusted during the learning phase, through stochastic gradient descent (Rumelhart et al., 1986; LeCun et al., 1998a).

(B1) $$\bar{z}_t = \mathrm{sign}(\mathbf{w}^T\mathbf{y}_t + b)$$

(B2) $$L(\mathbf{y}_t, z_t, \mathbf{w}, b) = 2\log(1 + e^{-z_t(\mathbf{w}^T\mathbf{y}_t + b)}) + \lambda|\mathbf{w}|$$

B.2. Support Vector Machines with Gaussian kernels

Support-Vector Machines (SVM) (Cortes and Vapnik, 1995) are pattern matching-based classifiers that compare any input pattern $\mathbf{y_t}$ to a set of support vectors $\mathbf{y_s}$. Support vectors are a subset of the training dataset and are chosen during the training phase. The function used to compare two patterns $\mathbf{y_t}$ and $\mathbf{y_s}$ is called the kernel function $K(\mathbf{y_t}, \mathbf{y_s})$ (Eq. B3). The decision function (Eq. B4) is a weighted combination of the kernel functions. We used in this study SVMs with Gaussian kernels (Eq. B3). The set S of support vectors $\mathbf{y_s}$, the Lagrange coefficients α and bias b were optimized using Quadratic Programming. Gaussian standard deviation parameter γ and regularization parameter were selected by cross-validation over a grid of values. The whole classifier and training algorithm was implemented using the LibSVM library (Chang and Lin, 2001).

(B3) $$K(\mathbf{y}_t, \mathbf{y}_s) = \exp(-(\mathbf{y}_t - \mathbf{y}_s)^2 / \gamma)$$

(B4) $$\bar{z}_t = \text{sign}(\sum_{s \in S} \alpha_s K(\mathbf{y}_t, \mathbf{y}_s) + b)$$

References

D'Alessandro M, Esteller R, Vachtsevanos G, Hinson A, Echauz J, Litt B. Epileptic Seizure Prediction Using Hybrid Feature Selection Overn Mulitple EEG Electrode Contacts: A Report of Four Patients. *IEEE Trans Biomed Eng*. 2003:50(5):603-615.

D'Alessandro M, Vachtsevanos G, Esteller R, Echauz J, Cranstoun S, Worrell G, Parish L, Litt B. A multi-feature and multi-channel univariate selection process for seizure prediction. *Clin Neurophy*. 2005:116:505-516.

Andrzejak RG, Mormann F, Kreuz T, Rieke C, Kraskov A, Elger CE, Lehnertz K. Testing the null hypothesis of the non-existence of the pre-seizure state. *Phys Rev E*. 2003:67.

Arnhold J, Grassberger P, Lehnertz K, Elger CE. A robust method for detecting interdependence: applications to intracranially recorded EEG. *Physica D*. 1999:134:419-430.

Aschenbrenner-Scheibe R, Maiwald T, Winterhalder M, Voss HU, Timmer J. How well can epileptic seizures be predicted? An evaluation of a nonlinear method. *Brain*. 2003:126:2616-2626.

Bottou L, LeCun Y. Lush : Lisp Universal Shell programming language. http://lush.sourceforge.net/index.html. 2002.

Chang CC, Lin CJ. LIBSVM: a library for support vector machines. http://www.csie.nyu.edu.tw/cjlin/libsvm. 2001.

Cortes C, Vapnik V. Support-Vector Networks. *Mach Learn*. 1995:20(3):273-297.

Delorme A, Makeig S. EEGLab: an open-source toolbox for analysis of single-trial EEG dynamics including ICA. *J Neurosci Method*. 2004:134(1):9-21.

Esteller R, Echauz J, D'Alessandro M, Worell G, Cranstoun S, Vachtsevanos G, Litt B. Continuous energy variation during the seizure cycle: towards an online accumulated energy. *Clin Neurophy*. 2005:116:517-526.

Iasemidis LD, Principe JC, Sackellares JC. Measurement and Quantification of Spatio-Temporal Dynamics of Human Epileptic Seizures. In: M. Akay, editors. *Nonlinear Signal Processing in Medicine*. IEEE Press, 1999.

Iasemidis LD, Shiau DS, Pardalos PM, Chaovalitwongse W, Narayanana K, Prasada A, Tsakalis K, Carney PR, Sackellares JC. Long-term prospective online real-time seizure prediction. *Clin Neurophy*. 2005:116:532-544.

Harrison MA, Frei MG, Osorio I. Accumulated energy revisited. *Clin Neurophy*. 2005:116:527-531.

Jerger KK, Weinstein SL, Sauer T, Schiff SJ. Multivariate linear discrimination of seizures. *Clin Neurophy*. 2005:116:545-551.

Jouny C, Franaszczuk P, Bergey G. Signal complexity and synchrony of epileptic seizures: is there an identifiable preictal period. *Clin Neurophy*. 2005:116:552-558.

LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-Based Learning Applied to Document Recognition. *Proc IEEE*. 1998:86(11):2278-2324.

LeCun Y, Bottou L, Orr GB, Muller KR. Efficient Backprop. In: Orr GB, Müller K-R, editors. *Neural Networks: Tricks of the Trade*. Springer, 1998.

LeCun Y, Muller U, Ben J, Cosatto E, Flepp B. Off-Road Obstacle Avoidance through End-to-End Learning. In: *Advances in Neural Information Processing Systems NIPS'97*. Cambridge, MA: Morgan Kaufmann, MIT Press, 2005.

Lehnertz K, Litt B. The first international collaborative workshop on seizure prediction: summary and data description. *Clin Neurophy*. 2005:116(3):493-505.

Lehnertz K, Mormann F, Osterhage H, Müller A, Prusseit J, Chernihovskyi A, Staniek M, Krug D, Bialonski S, Elger CE. State-of-the-art seizure prediction. *J Clin Neurophy*. 2007:24(2).

Le Van Quyen M, Foucher J, Lachaux J-P, Rodriguez E, Lutz A, Martinerie J, Varela FJ. Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony. *J Neurosci Method*. 2001:11:83-98.

Le Van Quyen M, Navarro V, Martinerie J, Baulac M, Varela FJ. Toward a Neurodynamical Understanding of Ictogenesis. *Epilepsia*. 2003:44(12):30-43.

Le Van Quyen M, Soss J, Navarro V, Robertson R, Chavez M, Baulac M, Martinerie J. Preictal state identification by synchronization changes in long-term intracranial recordings. *Clin Neurophy*. 2005:116:559-568.

Litt B, Echauz J. Prediction of epileptic seizures. *Lancet Neurol*. 2002:1(1):22-30.

Maiwald T, Winterhalder M, Aschenbrenner-Scheibe R, Voss HU, Schulze-Bonhage A. Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic. *Physica D*. 2004:194:357-368.

Mirowski P, Madhavan D, LeCun Y. Time-Delay Neural Networks and Independent Component Analysis for EEG-Based Prediction of Epileptic Seizures Propagation. In: *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, BC, Canada*. AAAI Press, 2007:1982-1983.

Mormann F, Andrzejak RG, Elger CE, Lehnertz K. Seizure prediction: the long and winding road. *Brain*. 2007:130:314-333.

Mormann F, Elger CE, Lehnertz K. Seizure anticipation: from algorithms to clinical practice. *Current Opinion in Neurology*. 2006:19:187-193.

Mormann F, Kreuz T, Rieke C, Andrzejak RG, Kraskov A. On the predictability of epileptic seizures. *Clin Neurophy*. 2005:116:569-587.

Petrosian A, Prokhorov D, Homan R, Dasheiff R, Wunsh II D. Recurrent neural network based prediction of epileptic seizures in intra- and extracranial EEG. *Neurocomputing*. 2000:30:201-218.

Rajna P, Clemens B, Csibri E. Hungarian multicentre epidemiologic study of the warning and initial symptoms (prodrome, aura) of epileptic seizures. *Seizure*. 1997:6:361–68.

Rumelhart DE, Hinton GE, Williams RJ, *Learning internal representations by error backpropagation*, Cambridge, MA: MIT Press, 1986.

Schelter B, Winterhalder M, Maiwald T, Brandt A, Schad A. Do False Predictions of Seizures Depend on the State of Vigilance? A Report from Two Seizure-Prediction Methods and Proposed Remedies. *Epilepsia*. 2006:47:2058-2070.

Schelter B, Winterhalder M, Maiwald T, Brandt A, Schad A. Testing statistical significance of multivariate time series analysis techniques for epileptic seizure prediction. *Chaos*. 2006: 16: 013108.

Schulze-Bonhage A, Kurth C, Carius A, Steinhoff BJ, Mayer T. Seizure anticipation by patients with focal and generalized epilepsy: a multicentre assessment of premonitory symptoms. *Epilepsy Res*. 2006:70:83-88.

Stam CJ. Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *Clin Neurophy*. 2005:116:2256-2301.

Takens F. Detecting strange attractors in turbulence. *Lecture Notes in Mathematics*. 1981:898:366-381.

Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B*. 1996:58(1):267-288.

Vapnik V, *The Nature of Statistical Learning Theory*. New York, NY: Springer Verlag, 1995.

Winterhalder M, Maiwald T, Voss HU, Aschenbrenner-Scheibe R, Timmer J. The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods. *Epilepsy Beh*. 2003:4(3):318-325.

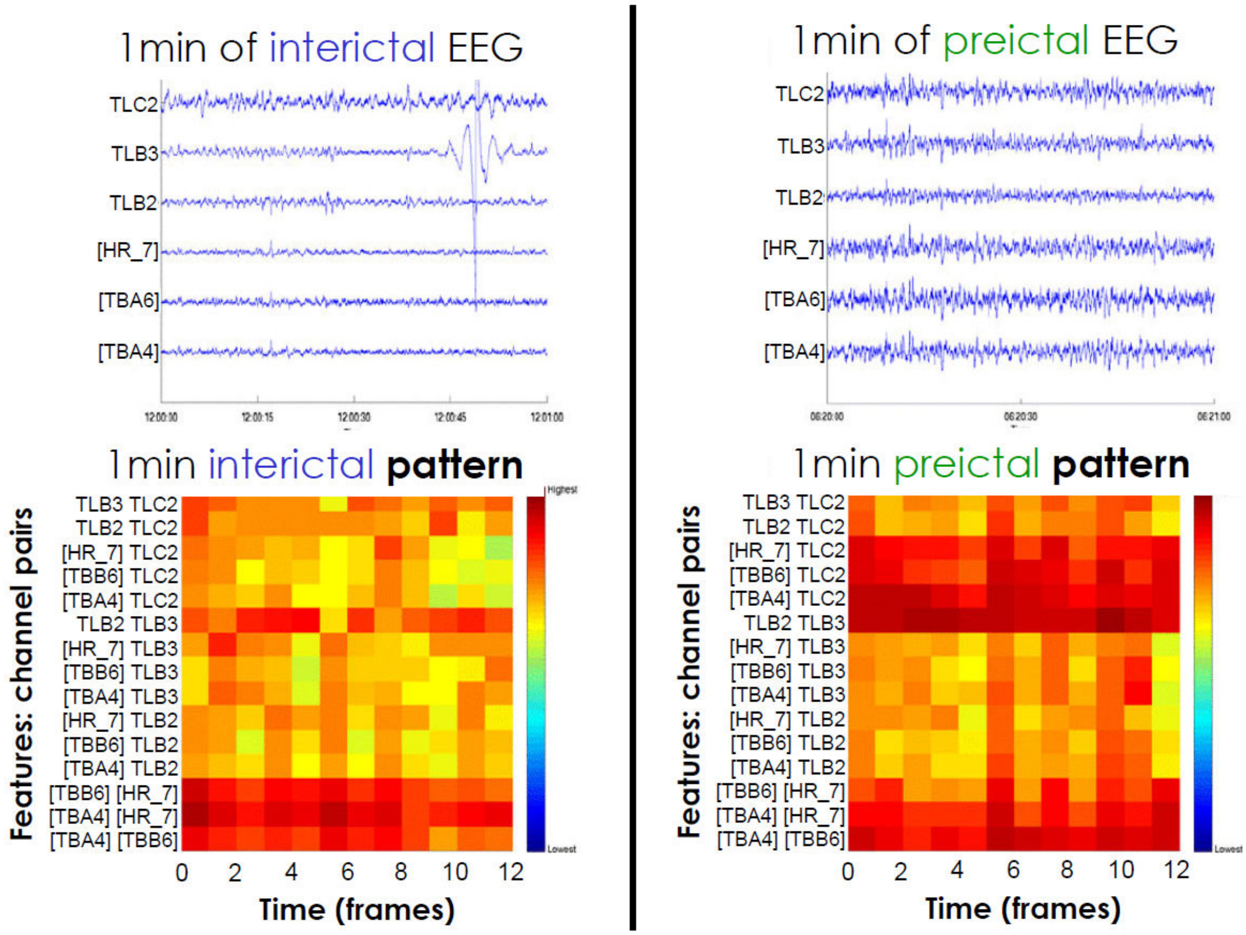

Figure 1. Examples of two 1-minute EEG recordings (upper panels) and corresponding patterns of cross-correlation features (lower panels) for interictal (left panels) and preictal (right panels) recordings from patient 012. EEG was acquired on M=6 channels. Cross-correlation features were computed on 5s windows and on $p=M\times(M-1)/2=15$ pairs of channels. Each pattern contains 12 frames of bivariate features (1min). Please note that channel TLB3 shows a strong, time-limited artifact; however, the patterns of features that we use for classification are less sensitive to single time-limited artifacts than to longer duration or repeated phenomena.

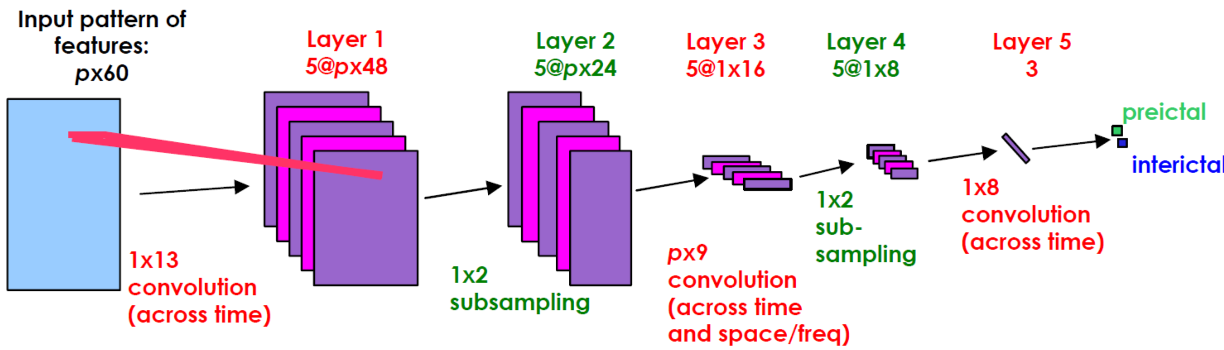

Figure 2. Convolutional network used for the classification of patterns of bivariate features containing 60 consecutive frames (5min) of p simultaneous features. Convolutional networks are a deep neural network architecture with a small number of parameters (weights) that are replicated over large patterns. Convolutional networks behave like successive arrays of small convolution filters. Inputs to hidden layers 1, 3 and 5 result from convolutions and inputs to hidden layers 2 and 4 are result from subsampling. Computations done between hidden layer 5 and the output layer of the convolutional networks correspond to a low-dimensional linear classifier. Thanks to alternated convolutional and subsampling layers, filters on the first hidden layer cover small areas of the input pattern, while filters on layers 3 and 5 cover increasingly larger areas of the original input pattern. For the specific problem of seizure prediction, convolutions are done only across time, with the exception of layer 3, which convolves input from all pairs of channels and all frequencies. Layer 1 can be seen as a simple short time pattern extractor, while layers 3 and 5 perform highly nonlinear spatio-temporal pattern recognition. For M=6 EEG channels, p=M×(M-1)/2=15 for non-frequency-based features and p= M×(M-1)/2×**7**=105 for wavelet synchrony-based features computed on 7 frequency bands.

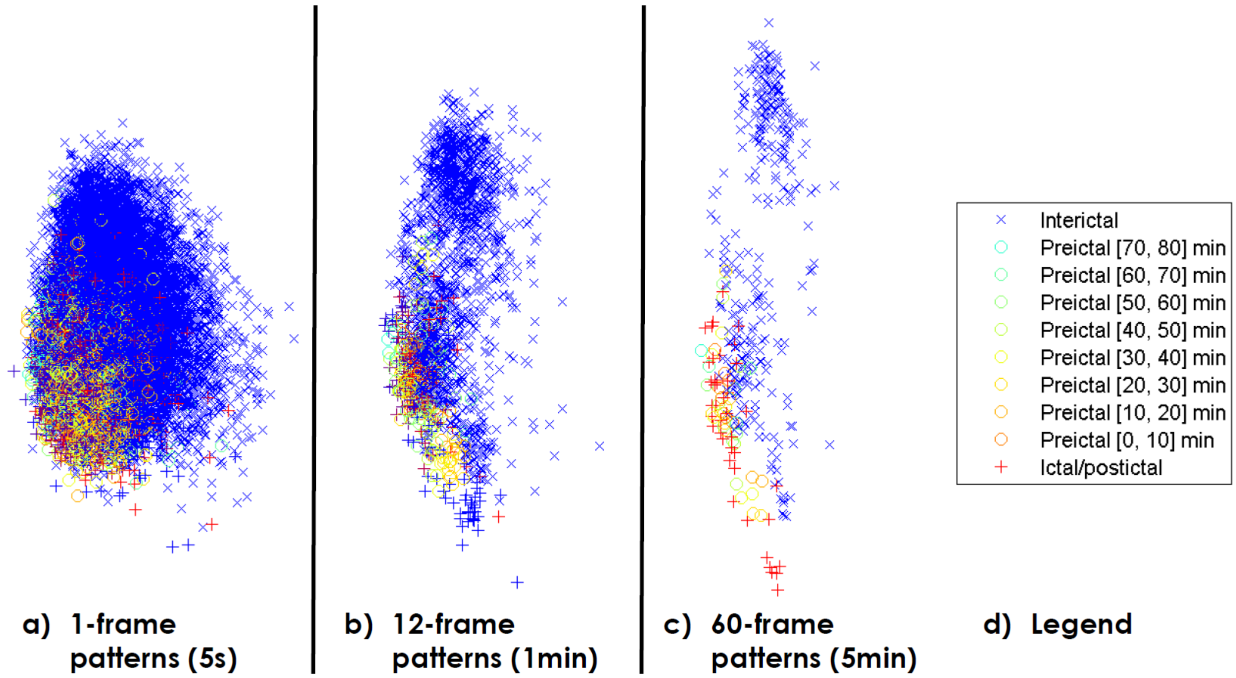

Figure 3. 2D projections of all the interictal, preictal and ictal patterns for patient 001, into the subspace defined by their first 2 principal components. Principal Component Analysis (PCA) was performed on a) 1-frame (5s), b) 12-frame (1min) and c) 60-frame (5min) patterns of wavelet synchrony SPLV features. Patterns a) are vectors containing 15×7=105 elements (15 pairs of channels times 7 frequency bands). Patterns b) are (15×7)×12 matrices containing 1260 elements. Patterns c) are (15×7)×60 matrices containing 6300 elements. As the duration (number of frames) of patterns increases, the separation between the preictal and interictal patterns becomes more apparent; this explains why a simple linear classifier (logistic regression) obtained good seizure prediction results on 60-frame patterns.

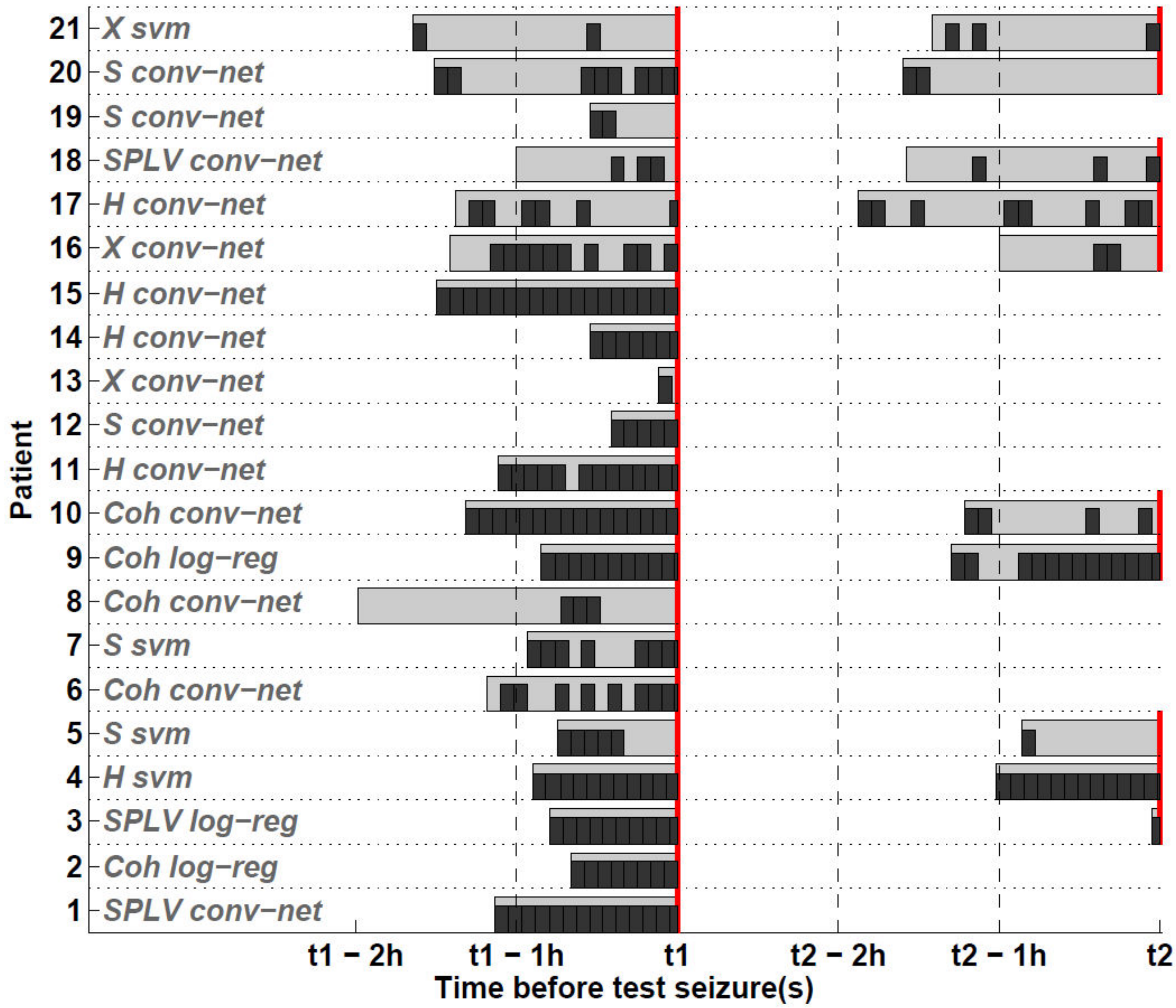

Figure 4. Best results obtained of the Freiburg dataset. For each patient, the figure shows the total duration of preictal EEG recordings (light gray) before each test seizure, and the times of preictal alarms. Some patients had one seizure used for test, other patients two, depending on the total number of seizures available for that patient in the dataset. The type of bivariate features and classifier are indicated on the left.

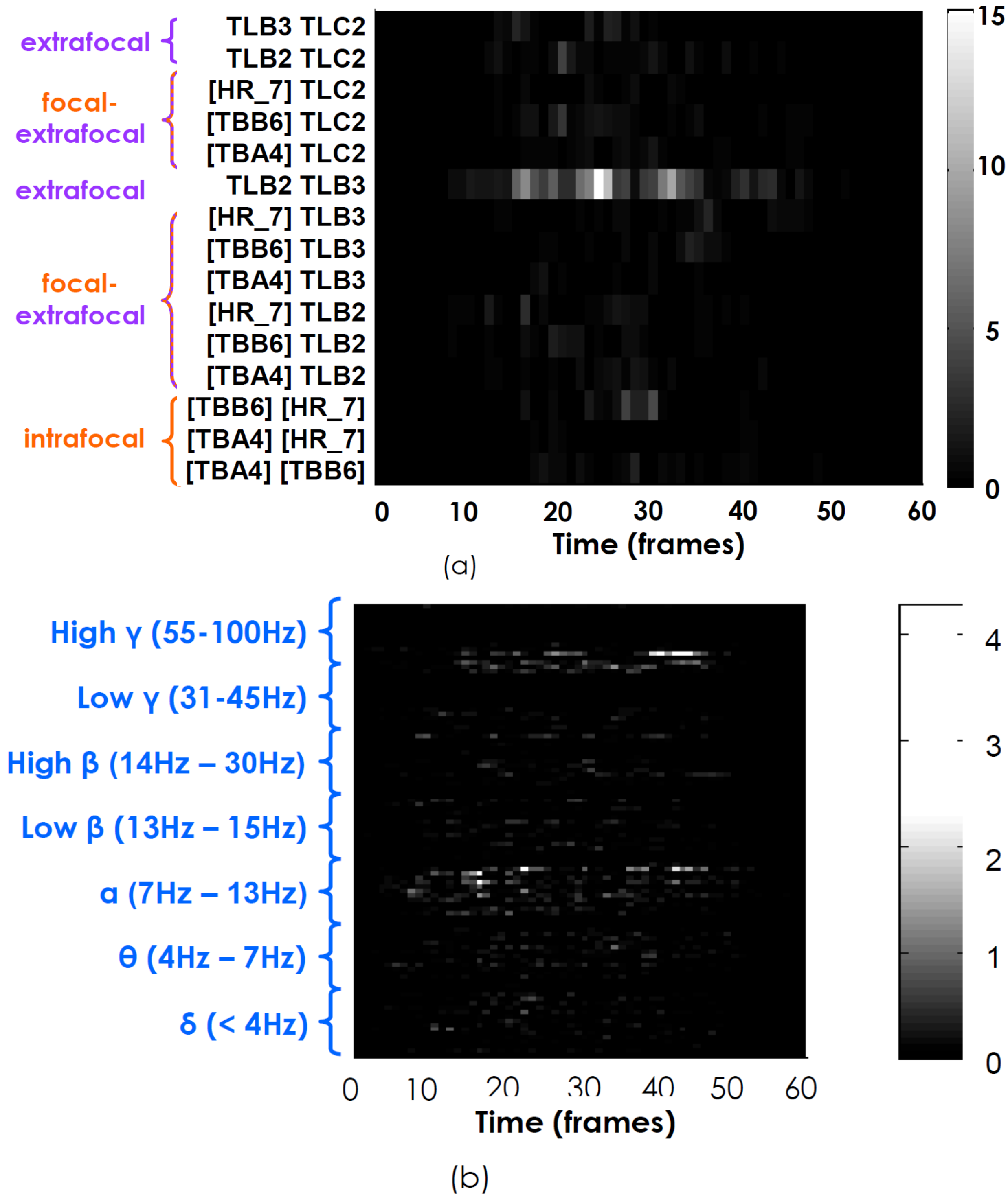

Figure 5. Sensitivity analysis for patient 012 (top panel a) and patient 008 (bottom panel b). Both images represent the input sensitivity of convolutional networks performed on 5min patterns of nonlinear interdependence (a) and wavelet coherence (b) respectively. Wavelet coherence in (b) is the only frequency-specific feature. Classifier (a) appears sensitive to interdependence features measure between two extrafocal EEG channels TLB2 and TLB3, whereas classifier (b) appears sensitive mostly to synchronization features in the high gamma range and in the alpha range.

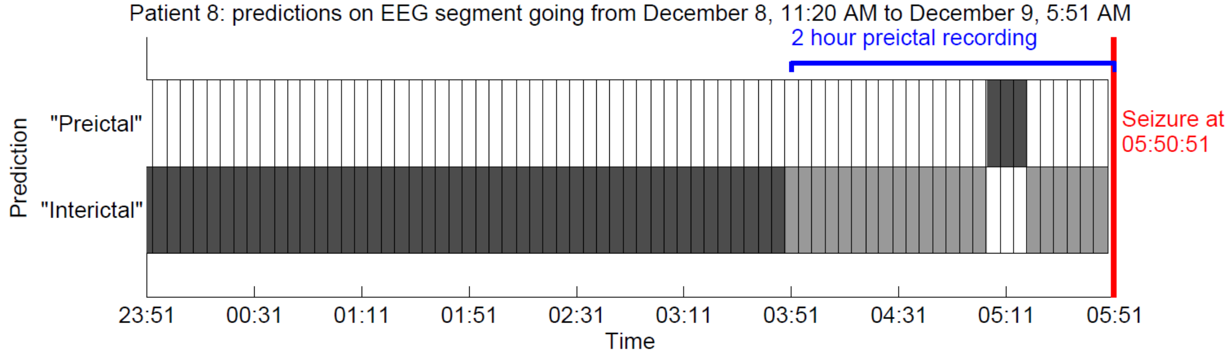

Figure 6. Detail of the performance of the seizure prediction system on patient 8, on a test dataset comprising a segment of EEG going from December 8, 11:20 AM through December 9, 5:51 AM. Only the segment after 23:51 PM is shown. The classifier was a convolutional network, and the 5min-long patterns consisted of wavelet coherence features. Dark grey boxes show successful predictions for each pattern (true negatives when the pattern is interictal and true positives when the pattern is preictal). Light gray boxes show false negatives (missed preictal alarms). There were no false positives/alarms.

Perfect seizure prediction (test set)	C			S			DSTL	SPLV			H			Coh		
	log reg	*conv net*	*svm*	*log reg*	***conv net***	*svm*	*svm*	***log reg***	***conv net***	*svm*	*log reg*	***conv net***	*svm*	***log reg***	***conv net***	*svm*
	4	9	4	3	**10**	5	1	**10**	**13**	7	9	**11**	7	**11**	**15**	8
	19%	43%	19%	14%	**48%**	24%	5%	**48%**	**62%**	33%	43%	**52%**	33%	**52%**	**71%**	38%

Table 1. Number of patients with perfect seizure prediction resuts (no false positives, all seizures predicted) on the test dataset, for each combination of feature type and classifier.

Perfect seizure prediction (test set)	Type of bivariate features					
	No frequency information			Frequency-based		
	C	***S***	*DSTL*	***SPLV***	*H*	***Coh***
	11	**19**	2	**14**	11	**13**

Table 2. Number of patients with perfect seizure prediction results on the test dataset, as a function of the type of EEG feature.

Perfect seizure prediction (test set)	Type of classifier		
	log reg	***conv net***	*svm*
	14	**20**	11

Table 3. Number of patients with perfect seizure prediction results on the test dataset, as a function of the type of classifier.

feature	classifier	pat 1		pat 2		pat 3			pat 4			pat 5		
		fpr	ts1	fpr	ts1	fpr	ts1	ts2	fpr	ts1	ts2	fpr	ts1	ts2
C	og reg	x	x	x	x	x	x	x	x	x	x	x	x	x
	conv net	0	68	0	40	x	x	x	0	54	61	0	25	52
	svm	0,23	68	0	40	x	x	x	x	x	x	x	x	x
S	og reg	x	x	x	x	0	48	3	0	54	61	x	x	x
	conv net	0	68	0	40	0	48	3	0	54	61	x	x	x
	svm	0,23	68	0	40	x	x	x	0,13	39	61	0	45	52
DSTL	svm	x	x	x	x	x	x	x	0	39	51	x	x	x
SPLV	og reg	0	68	0	40	0	48	3	0	54	61	x	x	x
	conv net	0	68	0	40	0	48	3	0	54	61	x	x	x
	svm	0,12	68	0	40	0	48	3	0	54	41	x	x	x
H	og reg	x	x	0	40	0	48	3	0	54	61	x	x	x
	conv net	0	68	0	40	0	48	3	0	54	61	x	x	x
	svm	0,23	68	0	40	0	48	3	0	54	61	x	x	x
Coh	og reg	0	68	0	40	0	48	3	0	54	61	x	x	x
	conv net	0	68	0	40	0	48	3	0	54	61	0	45	52
	svm	0,12	68	0	40	0	48	3	0	54	61	x	x	x

feature	classifier	pat 6		pat 7		pat 8		pat 9			pat 10			pat 11	
		fpr	ts1	fpr	ts1	fpr	ts1	fpr	ts1	ts2	fpr	ts1	ts2	fpr	ts1
C	og reg	x	x	0	46	x	x	x	x	x	0	79	73	x	x
	conv net	x	x	0	56	x	x	x	x	x	x	x	x	x	x
	svm	0,12	66	0	36	x	x	x	x	x	0,12	79	73	x	x
S	og reg	x	x	0	56	x	x	x	x	x	x	x	x	x	x
	conv net	x	x	0	56	x	x	0	51	78	x	x	x	0	67
	svm	0,12	16	0	56	0	9	0,13	51	43	0,12	79	73	0,25	67
DSTL	svm	x	x	x	x	x	x	x	x	x	0,24	9	3	x	x
SPLV	og reg	0	66	0	56	x	x	0	51	78	x	x	x	0	57
	conv net	x	x	0	56	0	39	0	51	78	0	79	73	0	67
	svm	0,12	66	0	56	x	x	0	51	78	0,24	79	73	0	27
H	og reg	x	x	0	56	x	x	0	51	78	x	x	x	0	67
	conv net	x	x	0	56	x	x	0	51	78	x	x	x	0	67
	svm	0,12	66	0	56	x	x	0	51	78	0,24	79	73	0	27
Coh	og reg	0	66	0	56	x	x	0	51	78	x	x	x	0	37
	conv net	0	71	0	56	0	44	0	51	78	0	79	73	0	67
	svm	0,12	66	0	56	x	x	0	51	78	0,24	79	73	0	32

feature	classifier	pat 12		pat 13		pat 14		pat 15		pat 16		
		fpr	ts1	fpr	ts1	fpr	ts1	fpr	ts1	fpr	ts1	ts2
C	og reg	0	25	0	2	x	x	x	x	x	x	x
	conv net	0	25	0	7	x	x	x	x	0	65	25
	svm	0	25	x	x	x	x	x	x	0	60	20
S	og reg	0	25	x	x	x	x	x	x	x	x	x
	conv net	0	25	x	x	x	x	x	x	x	x	x
	svm	x	x	x	x	0,13	33	0,12	90	0	55	55
DSTL	svm	x	x	x	x	x	x	x	x	x	x	x
SPLV	og reg	0	25	x	x	x	x	x	x	x	x	x
	conv net	0	25	x	x	x	x	0	90	x	x	x
	svm	x	x	x	x	0,26	33	0	80	x	x	x
H	og reg	0	25	x	x	0	33	0	70	x	x	x
	conv net	0	25	x	x	0	33	0	90	x	x	x
	svm	x	x	x	x	0,13	33	0	85	x	x	x
Coh	og reg	0	25	x	x	x	x	0	45	0	60	10
	conv net	0	25	x	x	x	x	0	90	x	x	x
	svm	x	x	x	x	0,26	28	0	85	0	60	5

feature	classifier	*pat 17*			pat 18			*pat 19*		pat 20			*pat 21*		
		fpr	ts1	ts2	fpr	ts1	ts2	fpr	ts1	fpr	ts1	ts2	fpr	ts1	ts2
C	og reg	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	conv net	x	x	x	x	x	x	x	x	*0*	*91*	*96*	x	x	x
	svm	x	x	x	x	x	x	x	x	x	x	x	0,12	99	70
S	og reg	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	conv net	x	x	x	x	x	x	*0*	*28*	0	91	96	x	x	x
	svm	x	x	x	x	x	x	x	x	x	x	x	x	x	x
DSTL	svm	x	x	x	x	x	x	x	x	x	x	x	x	x	x
SPLV	og reg	x	x	x	x	x	x	x	x	x	x	x	*0*	*99*	*75*
	conv net	x	x	x	0	20	70	*0*	*28*	x	x	x	x	x	x
	svm	x	x	x	x	x	x	x	x	x	x	x	0,12	99	80
H	og reg	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	conv net	*0*	*78*	*113*	x	x	x	x	x	x	x	x	x	x	x
	svm	x	x	x	x	x	x	x	x	x	x	x	0,12	14	75
Coh	og reg	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	conv net	x	x	x	*0*	*25*	*90*	x	x	0	99	20	x	x	x
	svm	x	x	x	0,23	15	90	x	x	x	x	x	0,12	99	75

Table 4. Seizure prediction results on the test dataset, as a function of the type of EEG feature and type of classifier. For each patient, the false positives rate (in false alarms per hour) as well as the time to seizure at the first preictal alarm (in minutes), for one or two test seizures, are indicated. Gray crosses mark combinations of EEG feature type and classifier type that failed to predict the test seizures or that had more than 0.3 false positives per hour.