Predicting Seizures Using Spectral Clustering and Cost-Sensitive SVM's

Xiyu Duan, Chris Fink, Hao Sun, and Tianpei Xie

1 Motivation

Epilepsy affects an estimated 0.6-0.8 % percent of people worldwide [1]. It is characterized by the unpredictable onset of disabling seizures, and though medication and surgery effectively treat approximately seventy-five percent of all people with epilepsy, it is incurable in the remaining twenty-five percent [1]. There is hope that such people may eventually find relief through the development of brain-implantable devices which would suppress seizures through appropriate electrical stimulation. Before such devices can be implemented, however, there must first be methods for predicting when seizures will occur. Even apart from the development of implantable devices, seizure prediction promises to improve the quality of life of epileptic patients, for it would enable them to engage in certain activities, such as driving, from which they are otherwise prohibited.

2 Background

Seizure prediction has been an active area of research for about three decades. It involves analyzing voltage traces from EEG recordings of brain activity and separating "inter-ictal" states (periods of normal brain activity between seizures) from "pre-ictal" states (periods of activity directly preceding seizures). Until very recently the field has focused upon univariate and bivariate measures of brain activity, such as the total spectral power within certain frequency bands or synchronization between electrodes [1]. This approach led to very simple algorithms which predicted a seizure would occur if one single measure breached a pre-defined threshold.

Such algorithms have not performed well enough to be implemented clinically, so in the last few years some researchers have started to implement machine learning techniques. Seizure prediction is quite amenable to classification algorithms since EEG data need only be sorted into one of two classes, inter-ictal or pre-ictal (it is not necessary to identify the actual seizure state because doing so has no clinical value: physicians do not need an algorithm to tell when someone is experiencing a seizure). Classification algorithms using support vector machines (SVM's) [2] and convolutional neural networks [3] have already been shown to better predict seizures than binary thresholding of various measures.

Performance of these algorithms is quantified through two simple metrics. The first is the **sensitivity**, which is simply the proportion of seizures which are accurately predicted: $\frac{N_d}{N_s}$, where N_d is the number of seizures accurately predicted over a given time interval, and N_s is the total number of seizures a patient experiences in that same time interval. The second performance metric is the false-positive rate, which is quoted in the literature as the number of times per hour that an algorithm incorrectly predicts a seizure will occur. The one tricky aspect of this endeavor is that the designer of the algorithm must arbitrarily specify a "prediction interval," which sets the interval following a seizure prediction that a seizure must occur in order for the prediction to be considered accurate. Longer prediction intervals result in greater specificity, but they carry an important clinical drawback: the longer the prediction window, the more time an epileptic patient will spend "under warning," i.e., anxiously

expecting a predicted seizure to occur. (Consider the limit as the prediction window goes to infinity. In that case, the algorithm would accurately predict every seizure, but the patient would always be under warning.) For this reason, it is better to measure algorithm performance not with false-positive rate, which is unaffected by the size of the prediction interval, but with the **proportion of time under warning (PTUW)**, defined by $\frac{N_f T_p}{T}$,where N_f is the number of false predictions, T_p is the length of the prediction interval, and T is the total seizure-free recording time (assuming that false positive intervals do not overlap). All seizure prediction algorithms balance maximization of sensitivity with minimization of PTUW.

It should be noted that both of these measures are quite different from a classification algorithm's accuracy, which pertains only to individual windows of time. Typically EEG data are partitioned into windows (usually between 20 seconds and 5 minutes), and one feature vector is extracted from each window. These windows are then labeled according to whether they occur during pre-ictal or inter-ictal intervals, and these labels and feature vectors are input to an SVM, for instance. Strictly speaking, the accuracy of an SVM is simply the proportion of feature vectors which are correctly classified, but this value by itself is a poor measure of performance for seizure prediction algorithms. This is because the classification of one single window of time as pre-ictal is all that is necessary to issue a warning to a patient, so that of the many pre-ictal windows that precede a seizure, only one must be classified correctly in order to successfully predict the seizure (see Fig. 1). In terms of individual feature vectors, the sensitivity of an algorithm is given mathematically by the equation

$$\frac{\sum_s \mathbf{1}_{(\sum_w \mathbf{1}_{\ell_{sw}>0})>0}}{N_s}, \tag{1}$$

where ℓ_{sw} is the label assigned by the machine learning algorithm to the w^{th} pre-ictal window preceding the s^{th} seizure (with +1 labeling the window pre-ictal and -1 labeling it inter-ictal), and N_s is the total number of seizures.

3 Objective and Data

Our objective was simple: to develop an algorithm for predicting seizures that would give as high a sensitivity and as low a PTUW as possible. As a benchmark, we planned to compare our results to those obtained in [2]. This study used data from eighteen in patients in what is known as the Freiburg database and obtained a sensitivity of 0.935 and a PTUW of 0.137 (using raw EEG data, and averaging across all patients). Until now the Freiburg database has been considered the standard for seizure prediction studies, but we were excited at the opportunity to analyze new data obtained through Professor Bill Stacey, an epileptologist at the University of Michigan medical school. He helped us to obtain EEG recordings from the new International Epilepsy Electrophysiology (IEE) database. These EEG recordings were better than those from the Freiburg dataset in several ways. First, they contained fewer artifacts; second, they had a larger number of recording electrodes; third, and most importantly, they contained many more seizures per patient. Whereas the Freiburg database had at most five seizures for any one patient, our new data contained recordings from three patients, which had fourteen, twenty-two, and thirty-six seizures. We expected this larger

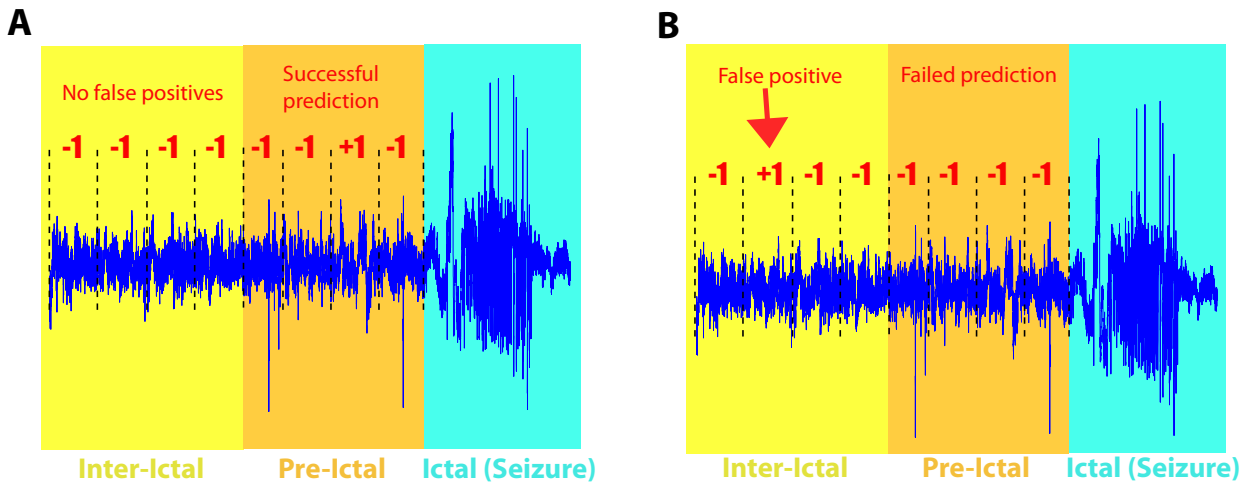

Figure 1: Illustration of performance metrics. The depicted EEG signal is partitioned into an ictal interval (determined by a trained epileptologist), a pre-ictal interval (preceding the ictal segment by a specified time duration), and an inter-ictal (normal) interval. For testing, all inter-ictal and pre-ictal intervals are further partitioned into windows of specified duration, from which feature vectors are extracted and classified. By convention, pre-ictal segments are labeled as +1 and inter-ictal as -1. In **A**, all inter-ictal windows are classified correctly, so there are no false positives. Also, at least one pre-ictal window is classified correctly, so the seizure is considered to be successfully predicted. In **B**, however, one of the inter-ictal windows is misclassified (constituting a false positive), and none of the pre-ictal windows are correctly classified, so the seizure is not successfully predicted.

number of seizures to improve performance through better training of our machine learning algorithm.

4 Methods

Following the approach taken in [2], we used cost-sensitive support vector machines (CSVM's) to classify windows of EEG data as either pre-ictal or inter-ictal. CSVM's are appropriate for seizure prediction because they allow investigators to tune the sensitivity versus PTUW trade-off by penalizing false negative windows differently from false positive windows. CSVM's have two user-specified parameters, C and R. In our implementation, C is the penalty term associated with false-negative windows, and R is the factor by which C is multiplied to obtain the penalty factor for false-positive windows. libSVM was used for implementation.

We tested two main ideas for developing a novel seizure prediction algorithm. The first stems from the fact that in previous studies, investigators have arbitrarily chosen a period of time before a seizure and identified this interval as being pre-ictal. In the literature, different studies have chosen this "pre-ictal interval" to be as long as two hours [3] to as short as thirty minutes [2] for the same dataset. This ambiguity is unfortunate but unavoidable because clinicians do not currently know when pre-seizure symptoms emerge or how to identify them

(if they did, the field of seizure prediction would be much more successful than it currently is). In machine learning terms, this means that we are unsure of when the true labels assigned to successive windows should change from -1 (inter-ictal) to +1 (pre-ictal).

We thought that we could perhaps use the classification produced by a machine learning algorithm to estimate the time at which pre-seizure symptoms emerge and thereby more accurately identify when the true labels should transition from -1 to +1. Training the SVM would therefore involve two passes: in the first pass, the CSVM would output predicted labels corresponding to each window of training data, and all windows at the beginning of a pre-ictal interval (initially given the "true" label +1) which were assigned the predicted label -1 (inter-ictal) would have their true labels changed to -1. Then, the CSVM would be re-trained with the updated "true" labels to obtain the final classifier. Using Fig. 1A as an example, the first pass would result in the first two windows of the pre-ictal interval being re-labeled as inter-ictal, and then the CSVM would be trained on the re-labeled dataset to obtain the final classifier. We referred to this training process as the "two-step CSVM" method, and we expected this approach to decrease the PTUW of our algorithm.

Our second idea for algorithm enhancement involved improving the features used for classification. This idea stemmed from one of the first papers to apply machine learning techniques to seizure prediction [3], in which the authors explored the use of several bivariate similarity measures between all possible pairs of recording electrodes to construct feature vectors for a given window of data. The data consisted of $N = 6$ recording electrodes for each patient, so the total possible pairs between electrodes was $N(N-1)/2 = 15$. Feature vectors were constructed using the 15 similarity values for all pairs of electrodes. This dimensionality is sufficiently small for use in machine learning algorithms, which is why the authors of the aforementioned paper were able to use bivariate measures successfully. However, most epileptic patients have dozens or even hundreds of recording electrodes gathering data, in which case it would be impractical to use all pairwise bivariate measurements to construct the feature vectors, since the dimensionality would increase as N^2. A simple solution to this problem was to cluster the bivariate measurements.

Not only would clustering reduce the dimensionality of the algorithm, but there was good reason to believe that it would improve classification performance. This is because decades of research into brain dynamics reveal that while most neurons are inactive at any one instant, the neurons that are active typically form small groups which neuroscientists call *assemblies*, and these assemblies change over time. It is likely that the build-up to a seizure involves a stereotypic evolution of such assemblies [4], and if this is the case, a good clustering algorithm combined with a good classification algorithm should offer excellent seizure prediction performance. A literature search reveals that, to the best of our knowledge, no one has previously applied electrode clustering in conjunction with machine learning techniques to seizure prediction. We used a slight variant of spectral clustering because a recent study showed it to give good results when applied to neuronal data [5].

Briefly, we constructed our feature vectors by first forming a similarity matrix S, in which each element $S_{i,j}$ represented the wavelet coherence between electrodes i and j over a two-

second time interval (we used the complex Morlet wavelet because it has been shown to give superior results with neuronal data [3, 6]). We then performed an eigen-decomposition of S and concatenated the leading eigenvectors for five consecutive windows to form feature vectors representing the evolution of brain dynamics over ten-second intervals (these time-scales were chosen to match those used in other studies [2, 4]). The k^{th} entry in one of the leading eigenvectors could be thought of as indicating the "soft membership" of the k^{th} electrode in the largest cluster at a given point in time. So we were essentially tracking the membership of all recording electrodes in the largest cluster over time.

5 Results

5.1 Our clustering algorithm applied to IEE data

We applied the foregoing method to construct feature vectors for recordings from three patients in the IEE database, then used a CSVM to classify each ten-second window of data. We used the hold-out method to assess algorithm performance, reserving one-third of the data for training, one-third for validation (optimizing C and R in the CSVM), and one-third for testing. Table 1 summarizes our results for both the standard CSVM and the "two-step CSVM" that we proposed.

	Standard CSVM		2-Step CSVM	
Patient ID	Sensitivity	PTUW	Sensitivity	PTUW
19	0.90	0.463	0.80	0.448
26	0.875	0.811	0.875	0.794
34	0.60	0.058	0.60	0.248

Table 1: Seizure prediction performance on IEE data using clustering to construct feature vectors.

These results were obviously disappointing, as patients 19 and 26 had high sensitivity but also a very high PTUW, while patient 34 had a very low value of PTUW but also a much lower sensitivity. No patient came close to the 0.935 sensitivity and 0.137 PTUW reported for the average patient in [2]. Furthermore, our two-step CSVM idea was not very effective. It led to very slight decreases in PTUW for patients 19 and 26, whose original values for PTUW were already quite high, but for patient 34, whose PTUW was very low to start, the two-step CSVM resulted in the PTUW increasing dramatically. These results are perhaps not so surprising when one simply looks at both the pre-ictal and inter-ictal feature vectors for a specific patient. Fig. 2 shows how there is no discernable difference to the naked eye between pre-ictal and inter-ictal feature vectors for patient 19, indicating that separating the two classes–even with a tool as powerful as the CSVM–is extremely difficult.

Although we attempted many tweaks to our clustering algorithm, such as using different similarity measures and applying different thresholds to the entries within the leading eigenvector, none worked very well. We suspected that our clustering approach probably was not

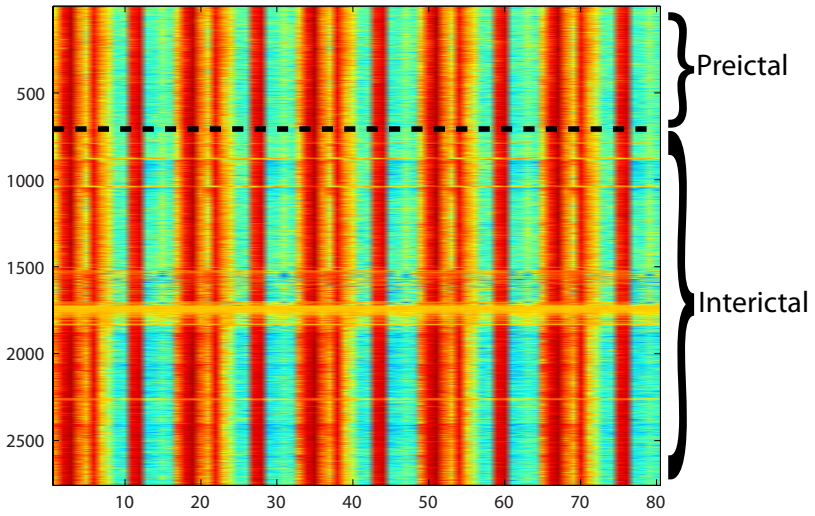

Figure 2: Pre-ictal and inter-ictal feature vectors constructed using spectral clustering for patient 19 from IEE database. Note how difficult it is to see obvious differences between the two classes by eye.

as effective as we had hoped, but we also wondered whether characteristics of the data were adversely affecting performance. We therefore decided to replicate the results of the Park study [2] (which claims to be the most successful seizure prediction study performed to this point in time) and then apply their methods to our data.

5.2 Spectral power method

5.2.1 Reproducing the Park study

In [2], the authors presented a simple method for constructing feature vectors: they calculated the power in each of eight different frequency bands, plus the total power, for each electrode recording within a window of time, and concatenated these nine values over the six electrode recordings for each patient in the Freiburg database. Each twenty-second window of time was therefore associated with a 54-element feature vector used for training a CSVM. Fig. 3 shows the promise of this approach for the first patient in the Freiburg database, whose data exhibit a clear difference between pre-ictal and inter-ictal feature vectors.

Because the number of seizures for each patient in the Freiburg database is so small (five or less), the authors performed double cross validation to assess their algorithm's performance. We therefore did the same, partitioning the data into N_s folds (N_s is the number of seizures), where each fold consisted of one pre-ictal interval and $\frac{1}{N_s}$ of the total inter-ictal data. For a given "trial," one fold was left out, while the others were used for training and validation. To determine the best parameters C and R within a trial, we conducted $N_s - 1$ "mini-trials," where one of the $N_s - 1$ folds was left out for validation, and the CSVM was trained on the remaining $N_s - 2$ folds over a spectrum of parameter values. A scalar performance measure

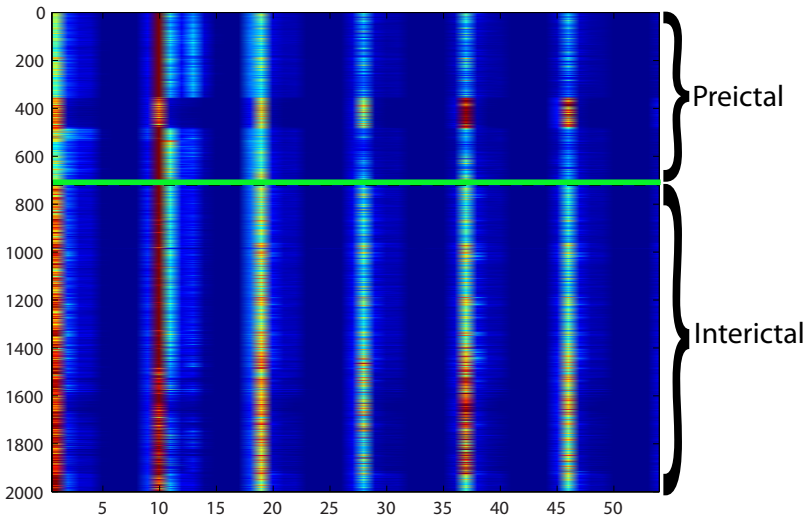

Figure 3: Pre-ictal and inter-ictal feature vectors constructed using spectral power for patient 1 from the Freiburg database. Note how there are several obvious differences between most pre-ictal feature vectors and most inter-ictal feature vectors, especially in the first feature.

F was used to quantify algorithm performance (see [2]) for each value of C and R within a mini-trial. F was then averaged over all mini-trials, and the values of C and R which maximized this average F were chosen as the optimal C and R values for a given trial. The CSVM was then trained on $N_s - 1$ folds and tested on the left-out fold for a given trial. This process was then repeated so that there were a total of N_s trials.

While time constraints prevented us from analyzing the data from all eighteen patients presented in [2], Table 2 displays the results of our attempt to replicate the method of Park et. al. on data from a subset of six patients. We provide the performance results reported by Park et. al. for comparison. Our results are not as good as those presented in [2] because there were a number of methodological details that the paper did not explain and that the corresponding author admitted he did not know when we contacted him. These unknowns included the method used for normalizing features and the value of γ used in the radial basis function. In addition, the method of Park et. al. included an essential post-processing step which applied a Kalman filter to the decision values output by the CSVM (see Fig. 4). We had to guess the parameters of this Kalman filter.

Nonetheless, even though we had to make educated guesses about much of their methodology, our results applying the algorithm of Park et. al. to the Freiburg database were still much better than the results we obtained applying our clustering approach to the data from the IEE database. The next step was to test their spectral power method on the data we had obtained from the IEE database, to see how well its superior performance would generalize to new data. (It should be noted that we applied both our clustering method and our two-step CSVM method to the Freiburg data, and neither gave good results.)

		Results of Park et. al.		Our Results	
Patient ID	No. of Seizures	Sensitivity	PTUW	Sensitivity	PTUW
1	4	1.0	0.000	1.00	0.232
3	5	1.0	0.027	1.00	0.231
4	5	1.0	0.000	1.00	0.000
5	5	1.0	0.580	1.00	0.825
9	5	1.0	0.042	1.00	0.203
14	4	0.75	0.400	0.75	0.999

Table 2: Comparison of our attempt to reproduce results from [2] with the actual performance results reported by Park et. al.

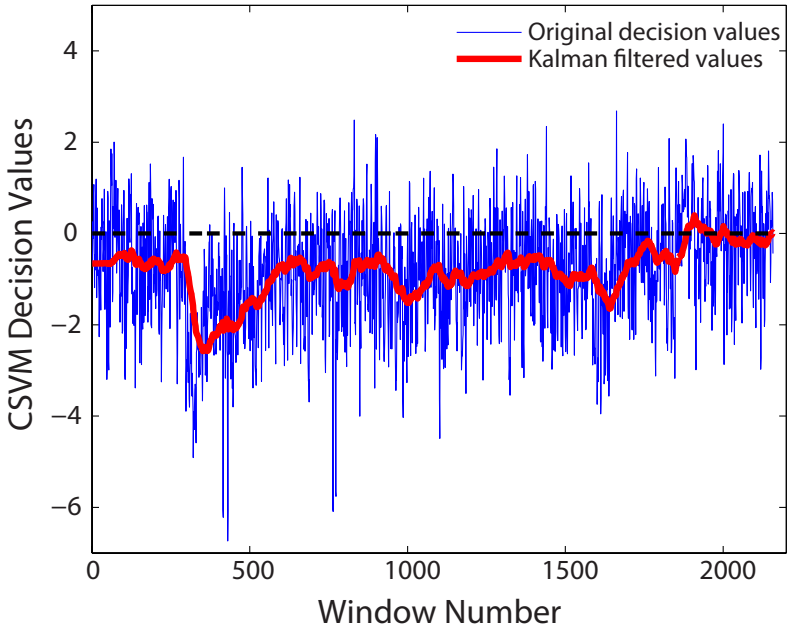

Figure 4: Example of Kalman filtering applied to the CSVM decision values of inter-ictal data for patient 1. Park et. al. noticed that the CSVM in their method often misclassified inter-ictal windows as pre-ictal windows (they tuned the parameter R so that false positives were more prevalent than false negatives), but that these false positives often occurred very briefly in time. They therefore applied a filter to the time series of CSVM decision values for all data in order to eliminate these misclassifications. You can see in this plot that the raw CSVM decision values are often greater than 0, implying that such windows would incorrectly be classified as pre-ictal and that the patient would be under warning for almost the entirety of this depicted inter-ictal interval. The *filtered* decision values, however, remain below zero for almost the entire time interval, dramatically reducing the PTUW.

5.2.2 Applying the spectral power method to IEE data

We suspected that part of the reason the spectral power method worked so well in [2] was a sample size effect–that perhaps the fact that each patient's data contained only four or five seizures contributed to the superior results that were reported. We therefore decided

to test the spectral power method on the IEE data not by using the hold-out method on all available data for each patient (as we had initially done when we applied our clustering method to the IEE data), but to instead randomly select the inter-ictal and pre-ictal intervals corresponding to just five seizures for each patient and then evaluate performance using double cross validation. If we obtained good results for a small number of seizures, but worse results for a larger number of seizures, we could be fairly certain that a sample size effect was in play.

However, the results were not good even for a small number of seizures. Even after trying many different parameter values, we could not do any better than the performance listed in Table 3. As in the previous studies, looking at the feature vectors provides a good indication why the performance is so bad (see Fig. 5). It therefore appears that small sample size most likely did not contribute to the superiority of the results published in [2], but that there must be some other explanation for why the spectral power method worked so well for data in the Freiburg database but not for data in the IEE database. We thought that perhaps the patients from the Freiburg dataset experienced different kinds of seizures than the patients from the IEE dataset, but after searching the information on seizure types recorded within the databases and talking to Professor Stacey, this does not appear likely.

Patient ID	Sensitivity	PTUW
19	1.0	0.919
26	1.0	0.806
34	1.0	0.893

Table 3: Seizure prediction performance on IEE data using spectral power to construct feature vectors. These values were obtained by performing double cross validation on five randomly-selected seizures for each patient.

6 Conclusion

Before further discussing explanations for this discrepancy, we must note that we only tested the spectral power method on three patients outside the Freiburg dataset, and that the spectral power method did even give exceptional performance for all the patients in the Freiburg database (four out of eighteen patients had PTUW values over 0.25). It is therefore possible that the spectral power method could yield good results for a high proportion of patients outside the Freiburg dataset, and that we just happened to pick three that did not. Still, the magnitude of the failure of the spectral power method when applied to these three patients (none had PTUW values under 0.80) indicates that the spectral power method most likely will not generalize well to other datasets.

We propose that the authors of [2] in a very general sense overfit the Freiburg data. We do not mean that they overfit any specific set of parameters, but rather that they picked a methodology that may work only for a specific dataset. They probably tried many different ways of constructing the feature vectors, then chose the one that gave the best results. Of the

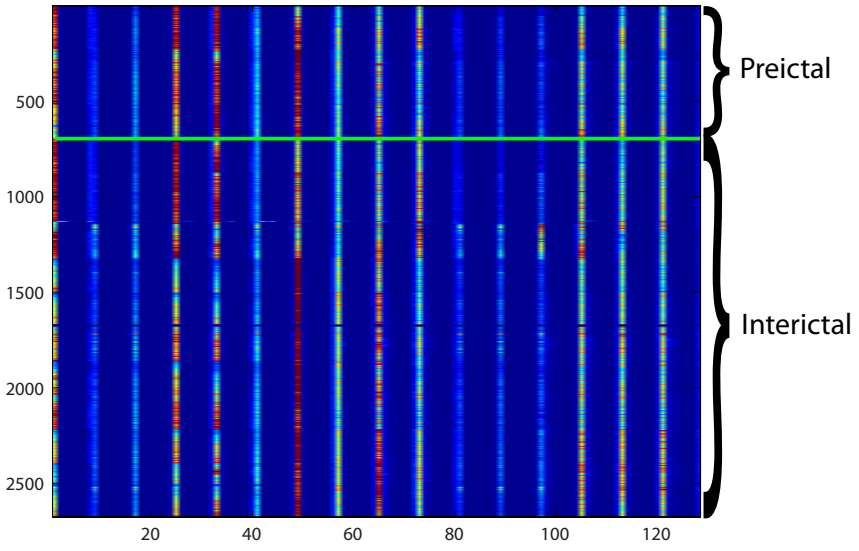

Figure 5: Pre-ictal and inter-ictal feature vectors constructed using spectral power for patient 19 from the IEE database. Note how difficult it is to tell differences between pre-ictal feature vectors and inter-ictal feature vectors by eye.

myriad ways of partitioning the frequency bands and normalizing the power in each band, they undoubtedly chose the method that gave the best performance. The decision to apply a post-processing filter, as well as the details of its implementation, were probably influenced by the specifics of the dataset.

We do not mean to be too hard on the authors, because they did the best they could with the data available to them. If anything, this study points to the necessity of building a very large database with data from hundreds to thousands of epileptic patients. Without this, it will be very difficult to design seizure prediction algorithms that work generally. Machine learning algorithms are powerful, but they are only as good as their input data.

It is also worth noting that while our original clustering method did not produce very good results on the IEE data, these results were arguably better than those produced by applying the spectral power method to the same data (compare Table 1 to Table 3). So while our clustering method did not produce good results in this study, it may be worth considering modifications of this approach in the future. Our "two-step CSVM" idea, on the hand, was not successful when applied to either the Freiburg or IEE datasets, and therefore does not look promising for future application.

7 Contributions

Xiyu worked mainly to obtain and process the data. Hao and Chris wrote code for implementing the machine learning methods. Tianpei found code for and figured out how to use libSVM. We all frequently brainstormed together to solve problems.

References

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

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

[3] P Mirowski, D Madhavan, Y LeCun, and R Kuzniecky. Classification of pattners of eeg synchronization for seizure prediction. *Clinical Neurophysiology*, 120:1927–1940, 2009.

[4] P Jiruska, J Csicsvari, AD Powell, JE Fox, WC Chang, M Vreugdenhil, X Li, M Palus, AF Bujan, RW Dearden, and JGR Jefferys. High-frequency network activity, global increase in neuronal activity, and synchrony expansion precede epileptic seizures *in vitro*. *The Journal of Neuroscience*, 30:5690–5701, 2010.

[5] X Li, D Cui, P Jiruska, JE Fox, X Yao, and JGR Jefferys. Synchronization measurement of multiple neuronal populations. *Journal of Neurophysiology*, 98:3341–3348, 2007.

[6] M Le Van Quyen, J Foucher, JP Lachaux, E Rodriguez, A Lutz, J Martinerie, and F Varela. Comparison of hilbert transform and wavelet methods for the analysis of neural synchrony. *Journal of Neuroscience Methods*, 111:83–98, 2001.