We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 3

Nonlinear Epilepsy Forewarning by Support Vector Machines

W.S. Ashbee, L.M. Hively and J.T. McDonald

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57438

1. Introduction

Epilepsy is a neurological disorder that changes the observable behavior of an individual to the point of inducing complete loss of consciousness. Pharmaceutical drugs may reduce or eliminate the problems of epilepsy, but not all people respond to pharmaceuticals favorably, and some may find the side effects undesirable. EEG-based epilepsy prediction may offer an acceptable alternative or complementary treatment to pharmaceuticals. Invasive, intra-cranial EEG provides signals that are directly from the brain, without the muscular activity that infests non-invasive, scalp EEG. However, intra-cranial EEG requires surgery, which increases risk and cost of health care, while reducing the number of people able to receive medical attention. Algorithms to predict the seizure event—the ictal state—may lead to new treatments for chronic epilepsy. Finding solutions that involve non-invasive procedures may result in treatments for the largest section of the population.

2. Background

Epilepsy prediction is greater than 1 minute of forewarning before there is any visible indication that a seizure will occur. The physician does not label the pre-ictal periods that precede the seizure—states that may indicate a seizure is near. Event characterization only labels the start time of the seizure. Consequently, labeled data for the pre-ictal state is non-existent, but is necessary to train a Support Vector Machine (SVM). Other researchers address this problem by assuming that the pre-ictal phase occurs immediately prior to a seizure [1]; see Figure 1 for an example.

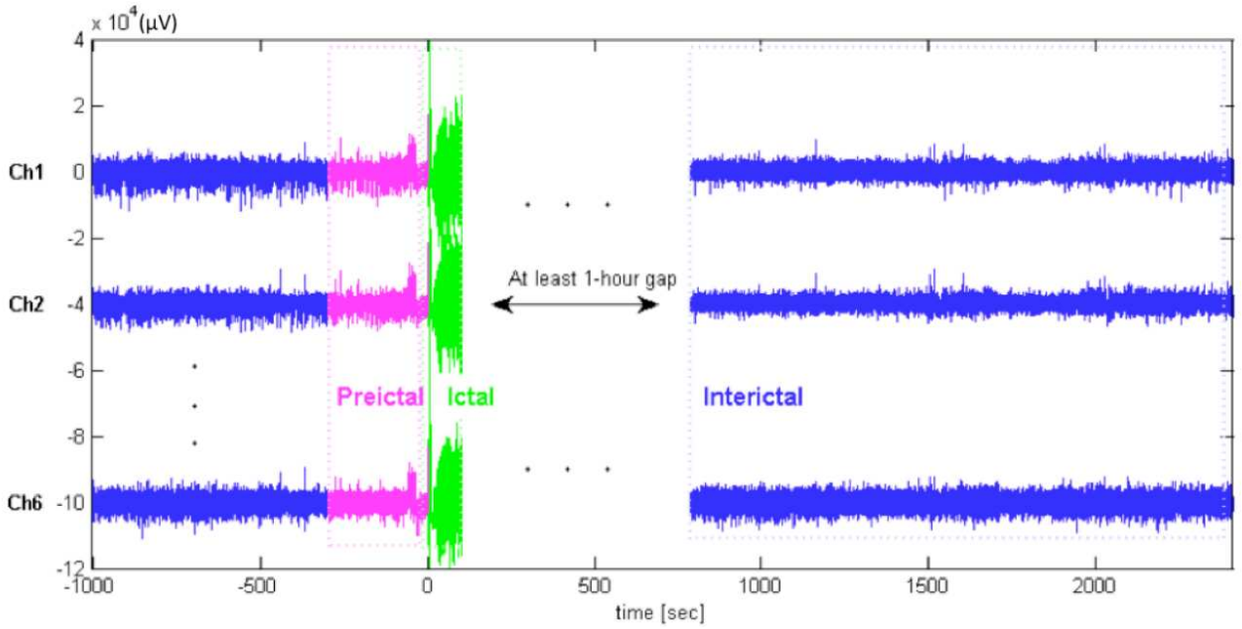

Figure 1. EEG with a seizure (ictal) event and potential class labels [1].

The labeling scheme of Figure 1 results in better than random predictions [1] under the assumption that the pre-ictal region immediately precedes the seizure and may be exploited for epilepsy prediction. This SVM approach provides the most obvious way to label the training and testing data without any extra information being available about the EEG.

Assuming pre-ictal dynamics occur within an hour of the seizure has the added benefit of being more likely to satisfy caregivers' requests to have forewarning within an hour of the seizure event. Netoff et al. achieve a specificity of 77% in classifying the pre-ictal region with no false positives with the above approach [1]. This level of accuracy is not high enough for a marketable prediction algorithm, but suggests that indicators of a seizure occur within an hour prior to a typical seizure. Netoff et al. use a "5 minute prediction horizon" where they label the pre-ictal region. They classify preictal as being within 5 minutes of the seizure and calculate specificities according to that labeling scheme. They assert that the short time frame makes the computational difficulty of the algorithm much more manageable than algorithms that have fewer restrictions on where the pre-ictal region is. They have a second stage of processing as well in which they look for 3 out of 5 pre-ictal indicators in a concentrated bundle in order to achieve prediction [1].

The assumption of pre-ictal indications near the event seems sound because a seizure resembles a dynamical phase transition. More specifically, the brain activity changes from some "normal" phase of brain activity into hyper-synchronous activity. The present work assumes that the brain dynamics within an hour of the seizure are approaching a phase transition, corresponding to measureable change in the scalp EEG. A simple example of a phase transition is liquid water becoming steam due to changes in pressure and temperature. However, scalp EEG exhibits nonlinear, chaotic features that are extremely difficult to predict over long periods and are extremely sensitive to initial conditions. Consequently, seizure prediction in a very

complex system like the brain is very difficult. Indeed, Stacey *et al.* [2] find that no algorithm provides better-than-chance prediction of seizures in statistical tests to date.

One must also choose whether to use monopolar (single channel) or bipolar EEG (difference between two monopolar channels). Mirowski et al. assert that epilepsy can be predicted more effectively with bipolar features [3] because of changes in the brain's ability to synchronize regions during a seizure. Mirowski et al. consider their pre-ictal period to be 2 hours. They assert, "[most] current seizure prediction approaches can be summarized into (1) extracting features from EEG and (2) classifying them (and hence the patient's state) into pre-ictal or inter-ictal". They go on to enumerate more specifically on the bipolar feature set in Figure 2.

(a) Bivariate features are computed on 5s windows (N=1280 samples at 256Hz) of any two EEG channels x_i and x_j.
(b) For M EEG channels, one computes features on $M \times (M-1)/2$ pairs of channels (e.g. 15 pairs for M=6).
(c) Features are aggregated for several consecutive time frames, e.g. 12 frames (1min) or 60 frames (5min).

Figure 2. Approach to using bivariate features in epilepsy prediction [3].

Figure 2 enumerates a feature set from all unique channel pairs [3]. After the enumeration, they use a grid search to find appropriate parameters with their SVM with a Gaussian kernel. Mirowski et al. use intra-cranial EEG data and obtain 100% accuracy for patient-specific machine learning models. However, no single model provides 100% accuracy for all patients [3], so they choose from among a variety of algorithms to achieve high accuracy on a patient specific basis.

By contrast, the present work uses non-invasive, scalp EEG. Moreover, the present work uses a SVM to extract seizure forewarning from the entire patient population. The goal is high accuracy. The long-term objective (not addressed in the present work) is lower health-care cost by using one algorithm for all patients to analyze scalp EEG on a smartphone.

Previous work by Hively *et al.* has used bipolar, scalp EEG and found the best seizure forewarning by using electrodes in the right frontal lobe [4, 5] in the 10-20 system; see Figure 3. The present work uses this same bipolar channel. Additionally, scalp EEG from one bipolar channel facilitates a simple, ambulatory device with two electrodes, which is far more manageable than an EEG headset with many channels. Our hypothesis is that the right-frontal region acts as a filter for pre-ictal condition change—a phase transition in the brain dynamics [6] that can be induced by noise [7]. Pittau *et al.* [8] reviewed the recent technical literature on sound-induced (musicogenic) seizures, which activate the fronto-temporo-occipital area.

Conversely, soothing music (e.g., Mozart's double piano sonata K.448) decreases the intensity and frequency of epileptic seizures [9].

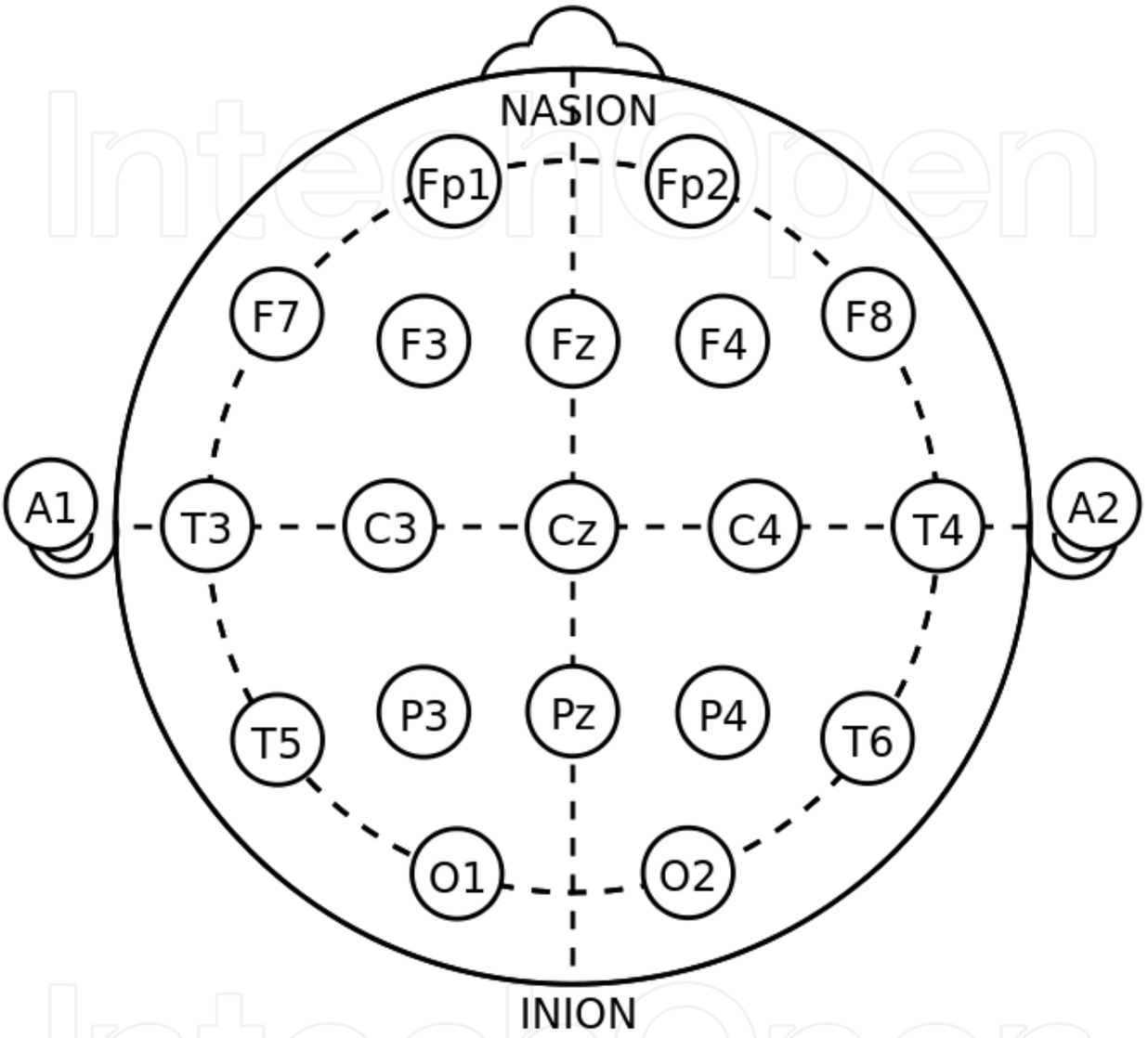

Figure 3. system (EEG) [10].

3. Phase-space analysis

We use one bipolar channel of *scalp* EEG (F8 – FP2) in the 10-20 system, as a measure of the noisy dynamics in cortical neurons over an area of roughly 6 cm^2. Our earlier work obtained channel-consistent forewarning across nineteen EEG channels [11]. The garbage-in-garbage-out syndrome is avoided by rejecting data of inadequate quality [12].

These data were uniformly sampled in time, t_i, at 250 Hz, giving N time-serial points in an analysis window (cutset), $e_i = e(t_i)$. Data acquisition was under standard human-studies

protocols from 41 temporal-lobe-epilepsy patients (ages from 4 to 57 years; 36 datasets from females, and 24 datasets from males). The datasets range in length from 1.4 to 8.2 hours (average = 4.4 hours). Data characterization included patient activity. Forty datasets had seizures, and twenty had no event [13].

A patented zero-phase, quadratic filter enables analysis of scalp EEG by removing electrical activity from eye blinks and other muscular artifacts, which otherwise obscure the event forewarning. This filter retains the nonlinear amplitude and phase information [14]. The filter uses a moving window of $2w + 1$ points of e_i-data that are fitted to a parabola in a least-squares sense, yielding $N - 2w$ points of artifact data, f_i. Essentially no low-frequency artifacts occur in the artifact-filtered signal, $g_i = e_i - f_i$. The value, w, is a parameter that specifies the width in points sampled of the eye blink filter; the value, N, represents the number of points in a cutset, which is represented by a graph; the value, g, is the artifact filtered set of points with eye blinks removed; the value, e, is the set of raw EEG data points; and the value, f, is the set of artifact filter points used to subtract out eye blinks.

A trade-off is required between coarseness in the data to exclude noise, and precision in the data to accurately follow the dynamics. Thus, the artifact-filtered data (g_i) are symbolized into S discrete values, s_i, that are uniformly distributed between the maximum (g_x) and minimum (g_n) in the first base case cutset. Uniform symbols are generated by the form in Eq. (1).

$$0 \le s_i = \mathrm{INT}\left[S\frac{g_i - g_n}{g_x - g_n}\right] \le S - 1 \tag{1}$$

Here, INT converts a decimal number to the closest lower integer. Takens' theorem [15] gives a smooth, non-intersecting dynamical reconstruction in a sufficiently high dimensional space by a time-delay embedding. The symbolized data from Eq. (1) are converted into unique dynamical states by the Takens' time-delay-embedding vector, y_i:

$$y_i = [s_i,\ s_{i+L},\ \ldots,\ s_{i+(d-1)L}] \tag{2}$$

Takens' theorem allows the y_i-states to capture the topology (connectivity and directivity) of the underlying dynamics. The time-delay lag is L, which must not be too small (making s_i and s_{i+L} indistinguishable) or too large (making s_i and s_{i+L} independent by long-time unpredictability). The embedding dimension is d, which must be sufficiently large to capture the dynamics, but not too large to avoid over-fitting.

The states from Eq. (2) are nodes. The process flow, $y_i \rightarrow y_{i+M}$, forms state-to-state links. The nodes and links give a formal, diagrammatic construction, called a "graph." This form gives topologically-invariant measures that are independent of any unique labeling of individual nodes and links [16]. Figure 4 depicts the algorithmic steps to: extract the analysis window from the stream of EEG data; remove the artifacts from scalp EEG; symbolize the artifact-filtered data; and construct the graph nodes and links [4]. The parameter space in Figure 4 enumerates the parameters used to generate the phase space graphs.

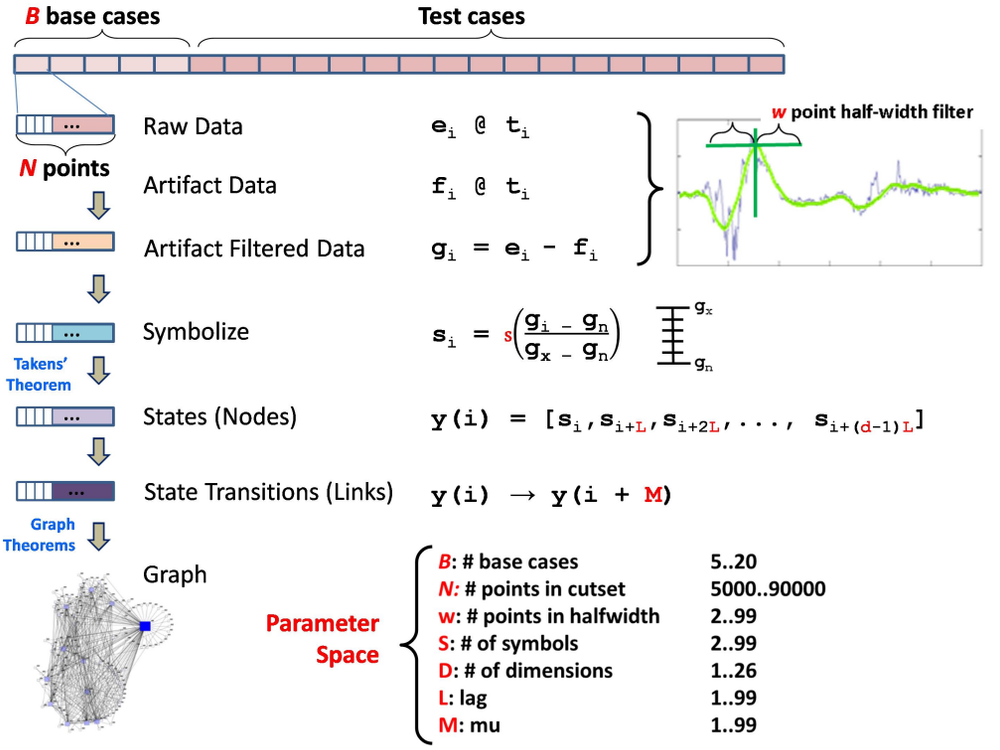

Figure 4. Nonlinear phase space construction [4].

The value, B, is the number of base cases, which establishes a normal range of activity for the patient. The value, N, is the number of sampled points that are in a cutset and graph, The value, w, as mentioned previously is the half-width of the eye-blink filter in sampled point units. The value, S, is the number of bins that the EEG is discretized into in order to create the base-s number represented by the vector y(i) in Figure 4. The value, d, is number of numerals in the base-s number or elements in the d dimensional vector, y(i). L is a time delay embedding that specifies the interval between points sampled in order to create a node. M is a second time delay embedding that specifies the interval between two connected nodes. The parameters mentioned are all used to generate the phase space graphs illustrated in Figure 4.

The dissimilarity measures involve counting unique nodes and links (those not in common between the two graphs): (1) nodes in graph A but not in B; (2) nodes in B but not in A; (3) links in A but not in B; and (4) links in B but not in A. Nodes and links in common between graphs do not indicate change and are not useful. These measures sum the absolute value of differences, which is better than traditional measures that use a difference of averages. Each measure is normalized to the number of nodes (links) in A (for A not in B) or in B (for B not in A). This feature vector, V, is used to classify the EEG as pre-ictal or inter-ictal. The analysis obtains a vector of mean dissimilarities, V, and matching standard deviations,σ, by comparison among the $B(B–1)/2$ combinations of the B base-case graphs, as shown in Fig. 4. Subsequent test-case graphs are then compared to each of the B base-case graphs to get an average dissimilarity vector, v. Our previous approach to obtain forewarning was several successive

instances (K) above a threshold (U_T) for each of J features, $U\left(\bar{V}\right)=\frac{|v-\bar{V}|}{\bar{\sigma}}$. This normalization allows regions of the feature space to be found that forewarn for many patients. Because the graphs are diffeomorphic to the underlying dynamics (from Takens' theorem), changes in the scalp EEG are captured by changes in the graph measures.

The present work uses a SVM approach to obtain forewarning from the normalized dissimilarity measures—namely, we find nonlinear regions in the feature space using a SVM. Figure 5 shows the calculation of the dissimilarity measures [4]. The frequency of nodes and links is not used because Takens' theorem guarantees topology, but not density—meaning Takens' theorem doesn't guarantee useful information in the repetition of nodes or links.

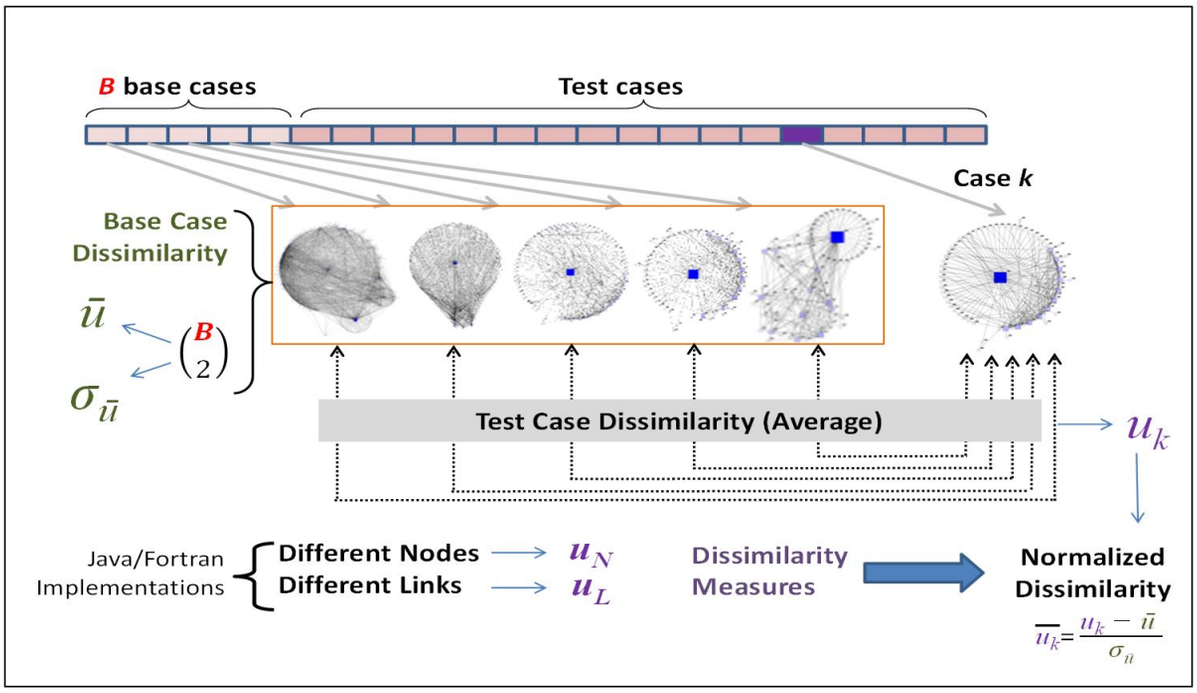

Figure 5. Normalized Graph dissimilarity measures, based on [4].

The dissimilarity measures in Figure 5 capture topology changes between two graphs. While node and link differences are basic graph measures, they quantify the hypothesis in a very simple and general way. Less commonality of nodes and links between two graphs produces larger dissimilarity measures, which are used to capture changes in topology. Topology change is a necessary, but not sufficient condition for a phase transition [17]. Our results show that changes in topology over extended periods indicate a higher likelihood of observing a phase transition as an indicator of an impending seizure. Additionally, the four graph dissimilarity measures from nodes and links rely on two concepts from set theory and Venn diagrams. Node dissimilarity and link dissimilarity are broken into two measures of dissimilarity each. Comparing two graphs (A and B) results in differences in nodes, as well as links. The dissimilarity measures are used as SVM features (for a total of 4 features in the Stage-1 SVM described below), and include the nodes in graph A that are absent from graph B, links in graph A that

are absent from B, nodes in graph B has that are absent from graph A, and links in graph B that are absent from graph A. All four dissimilarity measures are normalized and vary with cutset. Figure 6 shows these dissimilarity measures varying with time and how each cutset results in features and labels ("+" for pre-ictal, and "–" for inter-ictal).

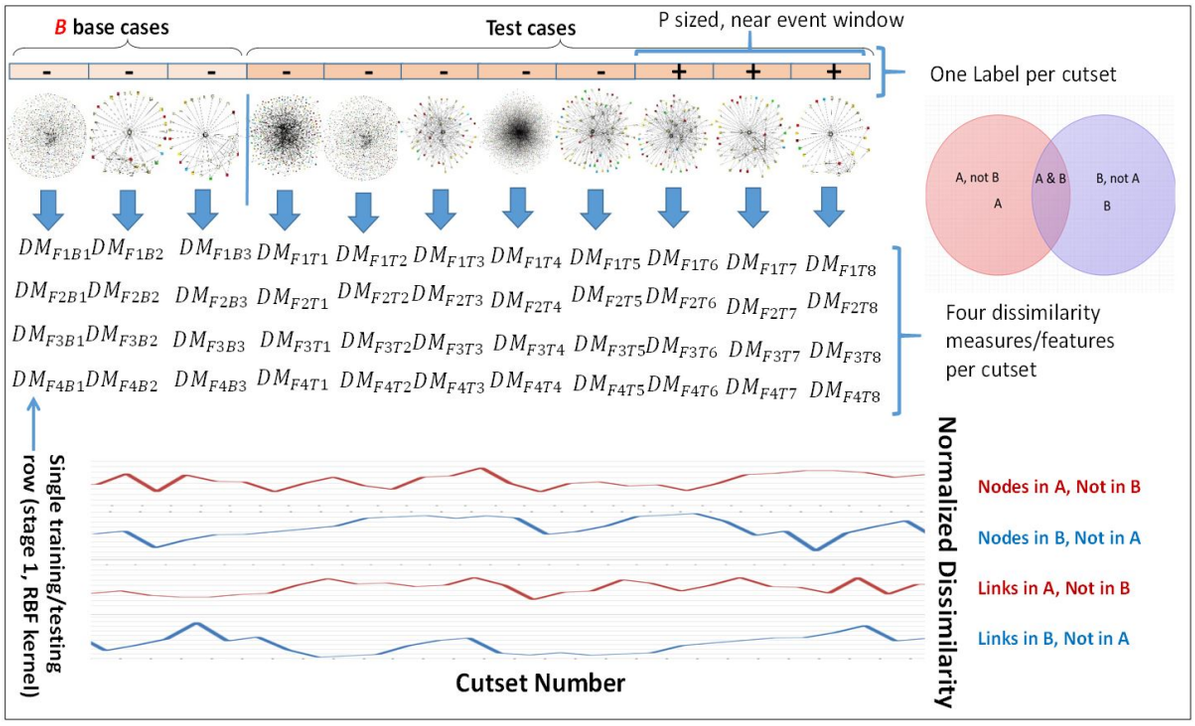

Figure 6. Cutset labels (+/-) and features (four dissimilarity measures).

Analysis of graph dissimilarity measures by a SVM allows quantification of the change in topology over time by determining how dissimilar the graphs must be to predict an epileptic event. The details of the forewarning algorithm are in Section 5—with a brief overview of Support Vector Machines in Section 4.

4. SVM with RBF kernels

SVMs are one of the most commonly used supervised learning tools. The SVM approach was originally designed as a two-class (binary) classifier, but has been expanded to single and multiple classes. A SVM without a kernel function performs linear classification by finding a hyper-plane in the feature space that best maximizes a margin of separation between two classes with a given list of features.

SVM kernels define the similarity between two points in the feature space. For example, with a radial-basis-function (RBF) kernel, two points are said to be similar when they are proximate to one another in the feature space. The RBF kernel function for two points in a feature space evaluates to 1 when the distance between the two points approaches zero. The RBF, Gaussian

http://dx.doi.org/10.5772/57438

kernel function evaluates to 0 as the distance between the two points becomes very large. The region where the kernel function evaluates to zero is parameterized by the value of gamma (γ), which is inversely proportional to the width of a multi-dimensional Gaussian function. SVMs with RBF kernels transform the decision boundary from a hyper-plane into much more amorphous decision boundary in the feature space. Figure 7 shows the difference in boundaries found by linear and RBF kernels for a representative SVM example in two dimensions.

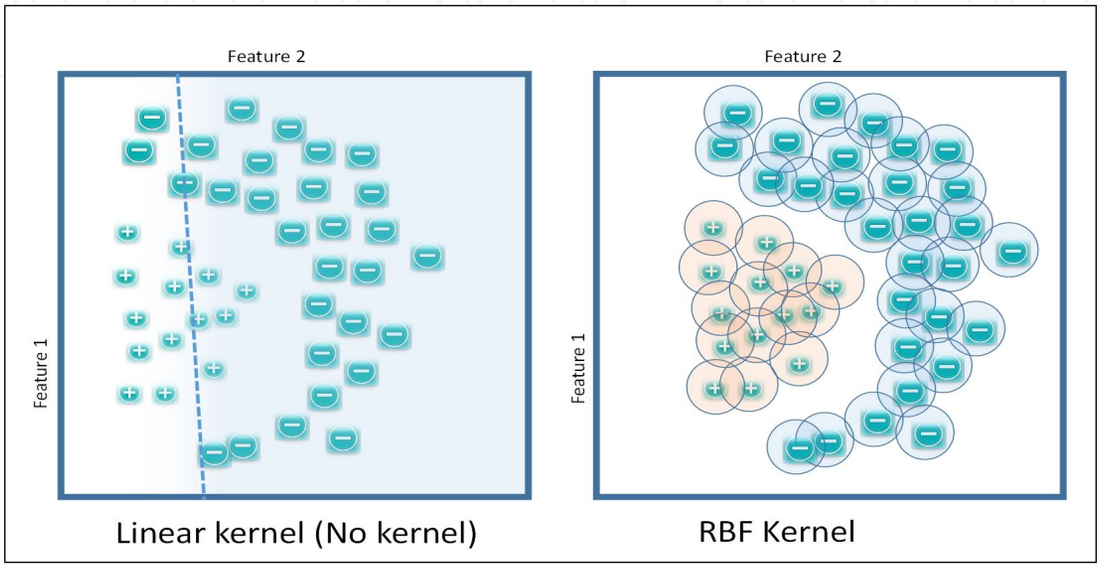

Figure 7. Types of boundaries found with two main SVM kernel types.

Each point in Figure 7 is equivalent to one instance of a class. Positive class values are denoted by positive signs and negative class values are denoted by negative signs. The Cartesian dimensions are the feature values—such as a dissimilarity measure. More than two features (two dimensions) can be used with a SVM, but it is more difficult to visualize when more than 3 dimensions (features) are involved. The main requirement of a RBF kernel is that the training set has a representative sample of the data that will be observed in the future and enough features to make distinctions between classes. Additionally, the range and scale of each feature has a large effect on the value of γ, the results, and the accuracy of predictions. Given a SVM model, future points are likely to be labeled as the class, to which they are most similar in the feature space. Similarity is defined by the kernel function as closeness between points of one class in the feature space. In essence, points from the training set are stored along with a weight associated with that point. A weighted sum of inner products is computed to evaluate how similar a new point is to all of the training data. The SVM training phase minimizes a cost function of several parameters, one of which is a vector, θ, of weights. Eq. (3) [18] gives the SVM cost function to be minimized with a kernel.

$$\min_{\theta} C\sum_{i=1}^{m}\left(y^{(i)}\text{cost}_1(\theta^T f^{(i)}) + (1 - y^{(i)})\text{cost}_0(\theta^T f^{(i)})\right) + \frac{1}{2}\|\theta\|^2 \quad (3)$$

Here, $y^{(i)}$ is the class label of a training point i (positive or negative one); $f^{(i)}$ is the feature vector of the point (i), which compares a single point with the other points in the training set. Cost is an input parameter to the SVM and is a penalty for being in one class or the other.

Once the vector θ is found through the minimization, it can be used to determine the classes of new points. The method of determining the label of new points is shown in Eq. (4) [18].

$$\text{class label of new point} = \begin{cases} \textit{Predict class } 1 \textit{ when } \theta^T f^{(i)} > 0 \\ \textit{Predict class } -1 \textit{ when } \theta^T f^{(i)} < 0 \end{cases} \quad (4)$$

Once the vector θ is obtained from training, it can be used to determine whether a new point in the feature space is of one class or another as given by Eq. (4). For each new point of index i in a test set, a vector $f^{(i)}$ is computed. The vector $f^{(i)}$ is a function of the kernel and the points in the training set. Each row in the vector $f^{(i)}$ is the value of the kernel function when the test set point and the training set points are inputs. The Gaussian kernel function in Eq. (5) will evaluate to 1 for points that are close together in the feature space and to zero when they are not. Eq. (5) gives the kernel function comparing a test set point x_i to a landmark training set point, $l^{(k)}$ [18].

$$f_k^{(i)} = \exp\left(-\gamma \left\| x_i - l^{(k)} \right\|^2\right) \quad (5)$$

Without a kernel, the vector θ has n+1 dimensions for n features. With a kernel, the cardinality of θ and $f^{(i)}$ is m+1 for m training points. Often, this comparison to known points in the training set with a kernel function is referred to as mapping the feature space into a higher dimensional space. Indeed, the dimensionality of f^i and θ increases from approximately the number of features without a kernel to roughly the number of training points with a kernel. The kernel results in linearly separable classes in the higher dimensional space, where the classes are not linearly separable in the feature space. The above description applies to binary classification with a RBF kernel and is intended to give the reader intuition on the types of boundaries that are found in the feature space as illustrated in Figure 7.

LIBSVM makes the details of the linear algebra of the training and testing transparent to the user and is easily used with the intuitions given in this Section [19]. The effect of a RBF kernel is that points proximate to one another will be labeled as belonging to the same class. Multi-class classification is treated as several binary classifications and is beyond the scope of the present work.

5. SVM forewarning algorithm using graph dissimilarity features

Labeling training data as pre-ictal or inter-ictal requires assumptions that must be sound. Current epilepsy prediction algorithms offer guidance about acceptable assumptions. The goal

http://dx.doi.org/10.5772/57438

is enough forewarning to stop or mitigate an event. Patients and caregivers [20] suggested 1-6 hours for safety, planning the day, and "driving myself to the hospital." Non-parent caregivers preferred 25 minutes to 1 hour for travel to the patient's location. Others gave 3-5 minutes, because longer forewarning was seen as more stressful to the patient. These requirements—as well as previous research indicating that these constraints are a reasonable request—led to the labeling scheme used for pre-ictal indications for a SVM. For epileptic event data sets, the pre-ictal region is labeled as being anywhere from 3.3 minutes to 70 minutes before the seizure. Each epileptic patient is labeled as being pre-ictal for the same length of time prior to the seizure. Each plus and minus sign in Table 1 represents a 3.3 minute window (consistent with a cutset length of 49716 points, sampled at 250 Hz). The number of pluses is determined by a parameter (p) that is varied during cross validation. Figure 6 shows how the signs in Table 1 relate to graphs, features, cutsets, and dissimilarity measures.

Patient data type	3.3 minute labeling	70 minute labeling
Event data set	----------+	----------++++++++++++++++++++
Non-event data set	----------	----------
Value of p	1	21

Table 1. Effect of variable p (number of + values) on Stage-1 pre-indication labeling of training data.

The input labeling (e.g., Table 1) assumes only approximate correctness and uses class weights to vary the correctness likelihood. The SVM methodology is implemented in three Stages with 10-fold cross validation. Stage-1 constructs a classifier that can label the pre-ictal state indicators. Stage-2 determines how long a patient must exhibit pre-ictal indicators in order to be certain of a seizure. Stages 1 and 2 establish cross validation accuracy and error. The SVM forewarning algorithm and previous voting method algorithm [4] both imply that patients must be in abnormal states a higher portion of the time before they are likely to have a seizure. Datasets without seizures can have infrequent abnormal states as well. Stage-3 obtains two models that can be used for seizure prediction in an ambulatory setting. Cross validation results in k different classifiers that leave out disjoint sets of data to establish an off-training-set error (OTS error) to avoid overconfidence in accuracy. However, k slightly handicapped classifiers result in either less accuracy than is possible or more complexity in creating an ensemble. Stage-3 avoids this unnecessary choice by performing cross validation to create a final SVM model that includes all of the available data. Accuracy and error rates are statistically stronger, when they are reported from cross validation. The statistical claims are less robust, when one trains and tests on the same data. SVM with a RBF kernel is particularly susceptible to over-fitting—implying the need for cross validation. Figure 8 shows an outline of this three-Stage algorithm.

I. Stage-1
   a. Obtain 4 dissimilarity measures from phase space analysis (for each cutset in a patient data set). These dissimilarity measures become the features of the RBF kernel;
   b. Labels
      i. Labels for +1: p cutsets immediately before an event;
      ii. Labels for -1: all other cutsets. See Table 1;
   c. Divide datasets into 10 sets of patients: each set contains 4 event patients and 2 non-event patients;
   d. Train RBF Model on 9 sets of patients – obtain SVM model (see Figure 9);
   e. Predict on the remaining, 10th set with RBF Model from (1d) ;
   f. Scan the results from (1e) for max # of contiguous +1. See Table 4-5.
   g. Repeat 1d-1f (and save results): 10 predicted sets. See Table 4-5.

II. Stage-2
   a. Label each patient's feature set from (1g)
      i. Event data sets are labeled as +1;
      ii. Non-event data sets are labeled as -1;
   b. Train on 9 sets of max contiguous pre-ictal indicators from (1g): linear kernel (see Table 5);
   c. Predict on 1 set of max contiguous indicators from (1g): via linear model from (2b);
   d. Get false positive rate (FP) and false negative rate (FN) from (2c);
   e. Get $D_i = \sqrt{\left(\frac{FP}{2}\right)^2 + \left(\frac{FN}{4}\right)^2}$ (We use stratified cross validation);
   f. Repeat (2b) – (2e) 10 times;
   g. Get the average over D_i (average cross validation OTS error rate);

III. Stage-3
   a. Use D(average)=(Average prediction distance) from (2g);
   b. If D(average) < 0.7 create models to use in future via 3c-g;
   c. Use results (see Table 4-5) from all 60 data sets from (1g);
   d. Retrain linear model (similar to Stage-2, but with all 60 patients' max successive indicators) to get the number of successive occurrences to trigger forewarning (see Table 5);
   e. Use data from (1b) to retrain RBF model on all 60 patients' cutsets (a total of 4244 cutsets) to obtain the RBF model (see Figure 10);
   f. Use RBF Model (3e) result to predict +/- on data from (1b); see Table 7;
   g. Use linear model from (3d) to do event forewarning on all 60 patients (see Table 4-5, Table 7, Figure 9, Figure 10) and get $D_{final\ models}$ (represents an optimistic over-fit);

Figure 8. Steps in the three stages for cross validation and final model construction.

Figure 9 shows how Stages 1 and 2 flow together. Stage-3 involves training the RBF Model on all of the Stage-1 cutsets (4244 rows, instead of approximately 90% of it) and training the linear model on the all of the Stage-2 results (60 rows, instead of 90% of it). Then, one predicts on the training data to verify that the model is working as expected to produce $D_{final\ models}$.

Figure 9 shows that event datasets are labeled in Stage-1 as pre-ictal (+) in a window of p cutsets prior to the seizure and inter-ictal (-) outside of this window. All cutsets in non-seizure datasets are labeled as inter-ictal (-). A cost sensitive SVM is used to account for the uncertainty in the pre-ictal and inter-ictal labeling. The motive for this labeling scheme is the caregiver's desire to have forewarning within an hour of the event. Indicators are assumed to be near the event, and the time window is varied by the parameter (p) that is tested during cross validation. The assumption behind the design choice is that the pre-ictal state is a rare occurrence. Because

http://dx.doi.org/10.5772/57438

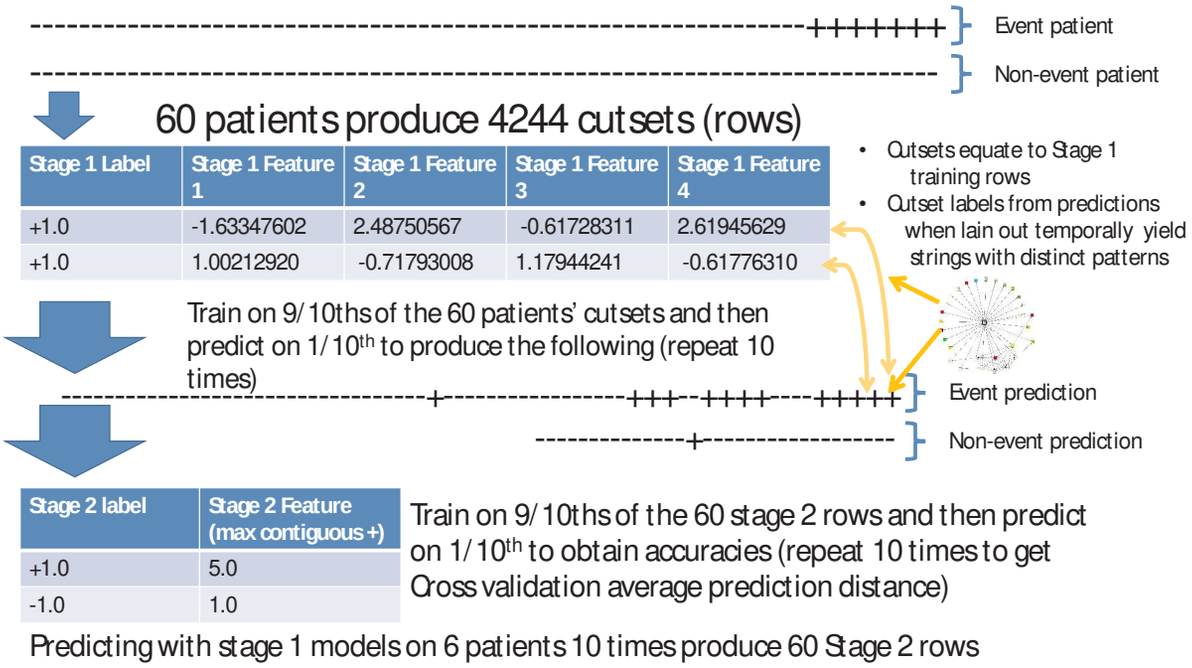

Figure 9. Stage-1 and 2 flow (each + or – represents a 3.3 minute cutset).

there may be similar points in the feature space outside the hour time-frame being labeled as pre-ictal and inter-ictal, the variable for class weights are varied via a Monte-Carlo search over the parameter space during cross validation to determine how to tie-break. Other parameters are also varied randomly during the Monte-Carlo search, as shown in Table 2. To compensate for labeling uncertainty, the cost sensitive SVM adjusts a weight on the class labels to indicate how certain the labeling scheme is for the pre-ictal and the inter-ictal classes. The labeling scheme combined with the features, training set, class weights, and gamma creates regions in the feature space that will be associated with one class or another. Additionally, we use stratified cross validation—maintaining a ratio of 4 event patients and 2 non-event patients in each strata of cross validation. Cross validation is performed on 90% of the patients—54 patients in each training set and 6 patients in each test set—having varying numbers of cutsets due to having varying length observations. This process is repeated 10 times with disjoint sets of patients in each test set.

Successive, contiguous occurrences of pre-ictal indicators trigger an alert (prediction of an event). Non-event datasets have more inter-ictal indicators labeled with negative symbols, while event patients have fairly dense pre-ictal indicators—labeled with plus symbols. A single pre-ictal indicator is usually not enough to make accurate predictions.

The accuracy of the assumptions is reflected in the success rate of the predictions during cross validation. The parameters that appear to be uncertain are left as search variables, recognizing that more free parameters create a more computationally complex search. Too many variables result in a computational explosion in the CPU time to explore the search space. Each point in the parameter space corresponds deterministically with a cross-validation error rate. Assumptions about the certainty of the training or testing example's labels are represented by class

weights in a cost sensitive SVM. Table 2 shows the variables for the SVMs that were searched during the research for this paper.

γ	c_{rbf}	$weigh\,t_+$	$weigh\,t_-$	c_{linear}	**p**
Sets the radius of the contribution of a single point to the decision boundary	Adjusts the pliability of the Stage-1 decision boundary given additional points	Weighs how powerfully the + class influences the decision boundary	Weighs how powerfully the – class influences the decision boundary	Adjusts the pliability of the Stage-2 decision boundary given additional points	Number of cutsets prior to the event that are labeled as being in the positive class.

Table 2. Parameters for the SVM portion of analysis.

Figure 4 and Table 3 lists other variables that might be searched in addition to those of the SVM. Those parameters were fixed in the present analysis because Takens' theorem is sufficiently powerful to show significant changes in topology with many sets of parameters when those topological changes are normalized properly. The dissimilarity measures for a patient reflect relative differences in graph topology. The parameter values for the phase-space graph generation are shown in the Table 3. Recall that these parameters were described near and are contained in Figure 4. Throughout the entire analysis, the parameters to generate the phase space graphs were kept fixed. These values were found in our prior work [4] to give good forewarning using ensemble voting methods mentioned in the background section.

B	**D**	**L**	**M**	**N**	**S**	**W**
12	7	56	77	49716	3	29

Table 3. Parameters used in generation of phase space graphs [4].

Statistical validation of forewarning requires measures of success. One measure is the number of true positives (TP) for known event datasets (Ev), to yield the true positive rate (sensitivity) of TP/Ev. A second measure is the number of true negatives (TN) for known non-event datasets (NEv). The true negative rate is TN/NEv (specificity). The goal is a sensitivity and specificity of unity. Consequently, minimizing the distance from ideal (D = prediction distance) is an appropriate objective function for any event type:

$$D = \sqrt{\left[1 - \left(\frac{TP}{Ev}\right)\right]^2 + \left[1 - \left(\frac{TN}{NEv}\right)\right]^2} \tag{6}$$

Eq. (6) is the objective function to be minimized for the OTS error rate and over-fit error rate. The OTS error rate is found from 10-fold cross validation from Stages 1 and 2. The over-fit error rate verifies that the final models in Stage-3 can correctly predict the training examples. Excessive false positives (inverse of a true negative) will cause real alarms to be ignored and needlessly expend caregiver resources. False negatives (inverse of a true positive) provide no

http://dx.doi.org/10.5772/57438

forewarning of seizure events. A Monte-Carlo search is used over variables in Table 2 because the prediction distance has very irregular, fractal behavior—with sparse parameters generating good predictions and the gradients of the parameter space being highly irregular with many local maxima and minima [4].

The algorithm for forewarning is done in two Stages of processing after producing diffeomorphic graphs and their dissimilarity measures for each graph. Stage-1 uses a cost sensitive SVM type (CSVC) from LIBSVM [19] with a RBF kernel. For each iteration of the cross validation, the algorithm labels event data sets as having pre-ictal indicators within a one-hour window prior to the seizure; see Table 1. The length of this window (p cutsets) is part of the Monte Carlo search. All other values—non-event data and event data far away from the seizure (outside the variable window)—are labeled as inter-ictal indicators. The analysis trains on k-1 sets, then predicts on a single, left-out set; this analysis is repeated k times in a k-fold cross validation (with k being 10). Table 4 shows a small sample of Stage-1 predictions for two patient outputs, but in actuality there are 60 sets of predictions—like Table 7 in the results section.

Common prediction output	Example predictions from Stage-1	Maximum Successive occurrences of +
Event Data set	----------+----------+-----++-++----------+++E 53	3
Non Event Data set	----------NE no prediction	0

Table 4. Representative predictions from Stage-1 (E=event; NE=no event).

After making the six predictions on six patients for one of ten cross validation runs, one creates a new set of cross validation folds (representing sets of patients) for the Stage-2 analysis out of the Stage-1's predictions on the omitted set (similar to the middle column of Table 4). For the Stage-2 single feature, one scans each of the predictions made by the Stage-1 for the maximum number of contiguous occurrences of pre-ictal indicators. Specifically, the third column of Table 4 is obtained from scanning the second column of Table 4. Stage-2 of the algorithm has training and testing files that follow the format of the second and third columns of Table 5.

Label meaning	Label	Single Feature (Maximum Successive occurrences of +)
Event Data set	+1.0	3
Non Event Data set	-1.0	0

Table 5. Training/Testing set format for Stage-2.

The analysis labels the Stage-2 training and testing values as either event or non-event data sets. In general, there are fewer pre-ictal indicators (+ in Table 4) in the non-event data sets on successful cross validation runs. One determines the cross validation average prediction

distance by training on values of maximum contiguous, pre-ictal indicators from k-1 subsets at the Stage-2, making k predictions on the omitted subset, and then taking the average of the prediction distances.

Stage-3 obtains the cross-validation prediction distance via two models that predict on the basis of all of the data available instead of 90% of it. Specifically, Stage-3 takes the 4244 cutsets labeled in Stage-1 and the sixty predictions from Stage-2 to create two SVM models for predicting on future patients. Stage-3 involves retraining both the RBF model (from Stage-1) and linear model (from Stage-2). Figure 10 illustrates the flow in Stage-3 after optimal parameters have been discovered via cross validation. Once the two models are obtained, they are used to predict on the original data sets to verify the model's validity.

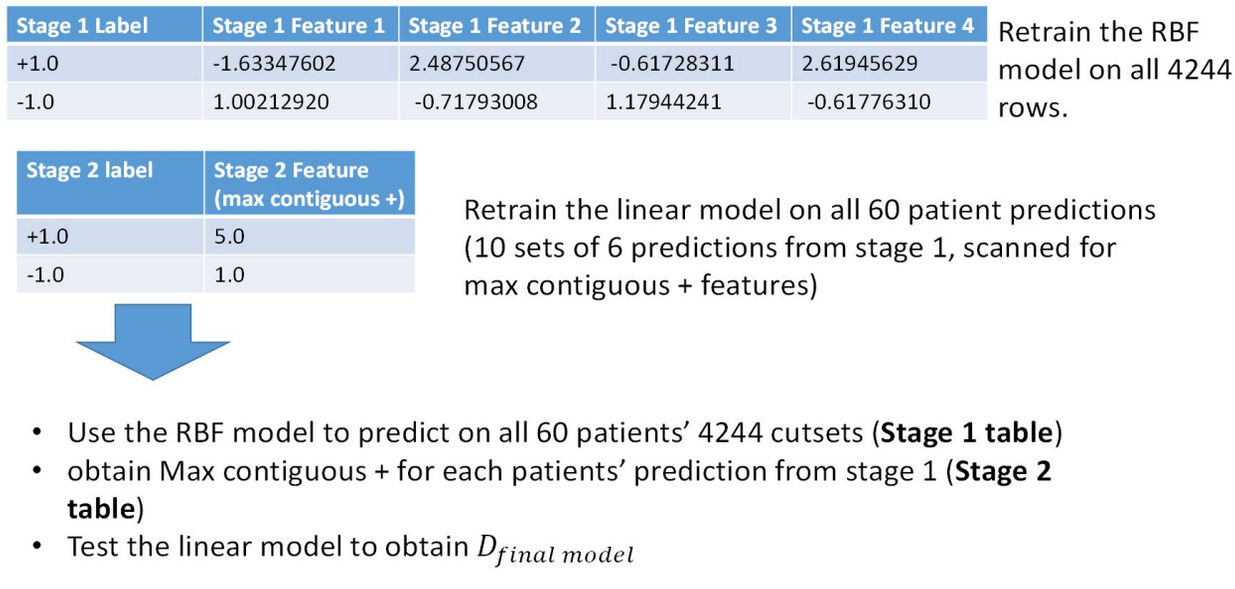

Figure 10. Stage-3 process (builds on Stages 1 and 2).

6. Representative results

From Eq. (6), the scenario of a classifier never getting the answer correct is $D = \sqrt{(1)^2 + (1)^2} \approx 1.41$. A random number generator that is guessing each class with equiprobability will have a prediction distance over time, $D = \sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2} \approx 0.7$. A perfectly ideal classifier will have an average prediction distance (OTS error rate) of 0 during cross validation. A Monte-Carlo search over the parameters for the SVM attempts to find the set of parameters and corresponding SVM models that minimize the prediction distance at Stage-2 for classification of each dataset as an event or a non-event. Cross validation on Stage-2 leads to k prediction distances that are averaged. The average is minimized and corresponding "ideal" parameters for the SVM are found. Once the Monte-Carlo search has found parameters associated with an acceptable OTS error rate, one retrains the Stage-1 model on all of the data.

One then retrains the Stage-2 model on all of the predictions from Stage-1. The retraining process results in two SVM models—one for Stage-1 and one for Stage-2. With these models, one can make predictions on new data.

Table 6 shows representative SVM parameter values (from the Monte Carlo search) and results. N_{OCC} is the number of contiguous, pre-ictal indicators (+ labels) that must be present before forewarning occurs and is found by training the Stage-2 SVM model (and Stage 3g). A value of Nocc of 2 implies that dynamics over a 6.6 minute period must be observed to be abnormal for prediction. Nocc of 1 implies that dynamics over a 3.3 minute period must be observed to be abnormal to have forewarning. D(AVG) is the average OTS error rate during cross validation runs; D(AVG)<0.7 indicates that the algorithm is performing more accurately than random guessing and simple heuristics. D(final) is a value verifying that the final models in Stage-3 have the capability of accurately classifying the training data. D(final) represents an over-fit, but is valuable in narrowing the Monte-Carlo search and determining whether the algorithm has any merit.

Table 6 shows that the best cross validation accuracy with an average prediction distance of 0.287 and a final model prediction distance of 0.056. Table 6 shows additional representative cross validation averages—D(Avg)—and final model prediction distances—D(final). Hundreds of runs resulted in cross validation prediction distances of <0.5. The best cross validation accuracy of.287 achieved thus far also has a fairly decent over-fit error rate of 0.056—which has a value of the objective function, $D=\sqrt{\left(\frac{1}{20}\right)^2+\left(\frac{1}{40}\right)^2}\approx .056$, corresponding to one false positive and one false negative in the final model's ability to predict the original training data. Although, due to the fact that the number of events and non-events was unbalanced, the probability of having a false negative was twice as likely as false positive, which is still desirable.

Ex#	γ	c_{rbf}	$weigh\,t_+$	$weigh\,t_-$	N_{occ}	c_{linear}	D(AVG)	D(final)	size of + label window
1	3.929	9.732	19.160	24.860	1	3.723	0.287	0.056	21
2	2.640	87.825	32.867	81.832	1	2.945	0.342	0.025	22
3	2.860	10.022	85.200	50.843	2	8.033	0.374	0.075	18
4	0.964	85.719	71.740	33.237	2	9.259	0.396	0.125	19
5	7.709	2.587	28.705	26.903	2	2.347	0.413	0.050	20
6	2.207	75.920	52.493	36.102	2	8.994	0.438	0.125	18
7	6.783	18.035	49.974	30.441	2	6.985	0.456	0	21

Table 6. Summary of typical best results to date.

Example 7 in Table 6 has an average cross validation accuracy of.456 with D(final)=0 (perfect prediction). Cross validation average accuracy or error rate is the more valid statistical claim. The best cross validation accuracy represented in Table 6 is in the same realm of accuracy as Netoff et al.'s intracranial methodology. Recall that Netoff et al. claim a specificity of 77.8% with no false positives, which is approximately a prediction distance of approximately.22 ($D_{Netoff} \approx \sqrt{(1 - .78)^2 + (0)^2} \approx .22$) [1]. Our best cross validation accuracy is of the same order of magnitude. Given that we are using scalp EEG, and Netoff et al. are using intra-cranial EEG, this is a statistically significant result that cannot be said to be an over fit. Mirowski et al. claim 100% accuracy, which would be a prediction distance of zero ($D_{Mirowski} = 0$). However, Mirowski et al. are making patient specific machine learning models that are tailored to individual patients. We are creating a novel algorithm that can be applied to a group of patients with non-invasive EEG. Very little research is being done to advance non-invasive EEG prediction algorithms in this way. Furthermore, comparing cross validation accuracies and error rates dispels any arguments of overconfidence due to over-fitting.

Figure 11 shows a plot of forewarning times (typically less than 1.5h) for the final-model of Example 7. The number of successive contiguous indicators to trigger forewarning was found to be 2 successive + values. Table 7 shows the Stage-1 predictions (Stage-3g of Figure 8 for all 60 patients) to produce the distribution of forewarning times in Figure 11. See Table 7 for the cutset indications (+ or -) that correspond to the parameters in Example 7 from Table 6.

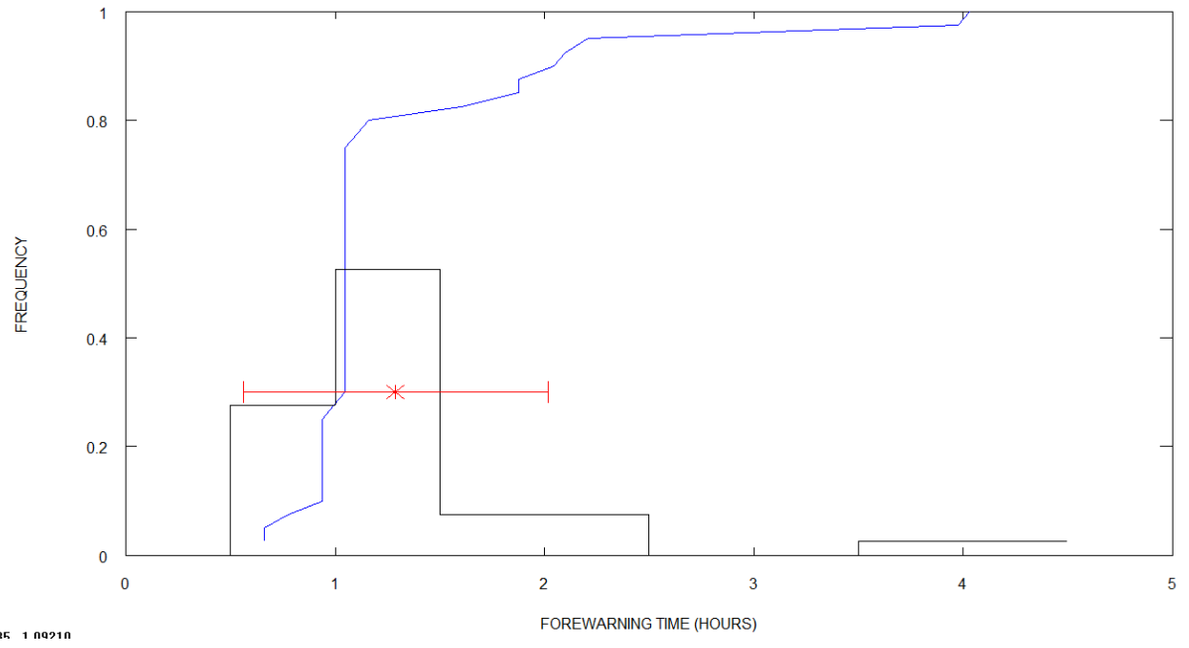

Figure 11. Distribution of forewarning times for Example #7 in Table 6.

http://dx.doi.org/10.5772/57438

Patient	
1	----------NE no prediction
2	----------NE no prediction
3	----------NE no prediction
4	-+----------+----------+----------NE no prediction
5	----------NE no prediction
6	----------NE no prediction
7	----------NE no prediction
8	----------NE no prediction
9	----------NE no prediction
10	----------NE no prediction
11	----------+----------NE no prediction
12	----------+----------+----------+----------NE no prediction
13	-------+----------+----+----NE no prediction
14	----------NE no prediction
15	----------NE no prediction
16	----------+----------+----------NE no prediction
17	----------+----------NE no prediction
18	----------+----------NE no prediction
19	----------+----------+----------+----------NE no prediction
20	----------+----------NE no prediction
21	--+----------+----------++++++++++++++++++++E 62
22	----------+++---++--+-+++-+--++E 62
23	----------+----------++++++++++++++++++++E 62
24	----------+-++++++++++-++-+++++E 56
25	----------+----------+--+---++-++++++++++-++E 46
26	----------+----------+-+-+-++-++++++++++++++++E 56
27	----------+-++--++++++-++--+-++E 56
28	----------+-------+----------+----------++--++++-+-+++-+-++++E 62
29	----------+-------+-------+----++-++++++++++E 39
30	----------+--+-++++++++++++++++++++E 56
31	----------+----------+----------+-------+++++++-++++++++-+-++E 62
32	----------+---+-+----------+----------++++++++++---++++++++E 62
33	----------+-+-+++++++++-++++-++-+--+E 66
34	----------++++++++++++--++++++E 62
35	----------++++--+-++-+++E 39
36	----------+----------+++-++++++-----+--+--E 62
37	----------++++++++++++++++++--E 62
38	----------+----------++----------+-----+++++----------++++++++-++++++++++-+E 238
39	----------+----------++++-+++++++++++++++++-E 62
40	----------++----------+----------+---+----------++++++++++++++++++++-E 241
41	----------+++-++++++++++-+++-+++E 62
42	----------++++++-++++++++-+---++++E 69
43	----------+-------++++++++++++++++++++E 62
44	----------+----------++-++++++++-+++++++++E 62
45	----------++-++-+++++++-+--++++E 62
46	----------+---+-++-+-++-+-+-++-+++-E 56
47	----------+++---+----++++++++++++++-+--+++++++++E 125
48	----------+-------+----------+-----+-------++----++----++-----E 56
49	----------++++++++++++++++++++E 62
50	----------++--+++--+-+-----++++++++-++-+-+++++++++E 122
51	----------+----------+++---++++-+++-+--++E 59
52	----------+---------+---------+--+--+-++-++-++-+++++++----E 56
53	--------++----------++-+++++++-++++++++++-E 132
54	----------+----------+-++++--+-----+-+++-++++++++++++++++++++E 112
55	----------+----------+++++-++++++++++++++-E 62
56	----------+-----+-----++----++++-----++++++++++-+-++++++++++E 112
57	----------++++++++++++++++++++E 62
58	----------++++++++++++++++++++-E 62
59	----------++-------++++++++++++++++++--+++E 96
60	----------++++++++++++++++++++E 62

Table 7. Stage-3g predictions on all 60 patients for Example 7 in Table 6 (E=event; NE=no event; number following "E"=forewarning time in minutes).

In Figure 11, the solid black line is the occurrence frequency (arbitrary units) in half-hour bins. The blue line is the cumulative distribution of forewarning versus time. The red H-bar with the star in the middle indicates the mean value of the forewarning times (approximately 1 hour) and the sample standard deviation. The result in Example 7 of Table 6 is better than random guessing or biased heuristics with D(final)=0, despite poorer cross validation accuracy than other examples. This example shows most of the forewarning times of ≤1.5 hours with a statistically significant accuracy. One can visually make a prediction by scanning Table 7 from left to right and looking for 2 contiguous plus values. When 2 values are found, the seizure is highly likely to occur. One may be tempted to reduce the forewarn time by increasing the value of positive values that trigger a forewarning, but that would likely result in bad OTS error, which is why it is not the value found by the second and third stage SVMs.

7. Discussion

Ideally, we would like to achieve an average cross validation OTS prediction distance of zero, final model prediction distance of zero, and all forewarning times <1 hour. In order to achieve this goal, additional features will need to be explored that exploit topology and the distance metric that Takens' theorem guarantees. Some modifications for Stage-2 improve the results (e.g., the choice of p cutsets prior to the event as a variable, and use of a RBF kernel instead of a linear one). More search parameters in Stage-1 (e.g., those in Table 3) should lead to better results with enough CPU time. Additional graph dissimilarity measures may be helpful. We have discovered more features for Stage-1 that may be of use. More data is needed for a robust statistical validation of the model.

The choice of optimal features is very difficult. Theorems guide the choice of parameters, features, and the algorithm. Some combination of theorem-based feature selection and occasional intuition derived from experimentation is the only way to keep the cost of the research initiative practical. Feature selection is one of many hard problems involved in epilepsy prediction. When one adds features, one often needs more data to make meaningful statistical assertions. Other important choices involve the type of kernel and the thresholding strategy. Linear kernels and thresholding strategies may perform well while Radial Basis Function (RBF) kernels perform poorly and vice versa. Our previous work [4] used a voting method that performed well. There is no guarantee that a set of features will behave similarly with different kernels and strategies to determine the threshold. Other measurement functions are possible under Takens' theorem to create the phase-space states. Use of a single-class or multi-class SVM could also prove fruitful.

The results in Tables 6-7 and Figure 11 are encouraging, despite several limitations, which are discussed next. (1) We analyzed 60 datasets, 40 with epileptic events and 20 without events. Much more data (hundreds of datasets) are needed for strong statistical validation. (2) These data are from controlled clinical settings, rather than an uncontrolled (real-world) environment. (3) The results depend on careful adjustment of training parameters. (4) Only physician-selected portions of the EEG are available, rather than the full monitoring period. (5) The

http://dx.doi.org/10.5772/57438

present approach uses retrospective analysis of archival data on a desktop computer. Real-world forewarning requires analyst-independent, prospective analysis of real-time data on a portable device. (6) The results give forewarning times of 4 hours or less. A time-to-event estimate is needed. (7) All EEG involved temporal lobe epilepsy; other kinds of epilepsy need to be included. (8) A prospective analysis of long-term continuous data is the acid test for any predictive approach. Prospective data were unavailable for the present analysis. Clearly, much work remains to address these issues.

8. Conclusions

The present work uses Support Vector Machine analysis to extend earlier work by Hively *et al.* [4] for forewarning of epileptic seizures. The previous work obtained a prediction distance of 0.0559 and a maximum forewarning time of 5.1 hours. The present analysis divides the continuous data stream from one bipolar channel of scalp EEG into contiguous, non-overlapping windows; removes the muscular artifacts with a novel zero-phase quadratic filter; converts the artifact-filtered data into discrete symbols; applies the time-delay-embedding (Takens') theorem to create unique phase-space states that capture the brain dynamics; forms a graph from the nodes (phase-space states) and links (dynamical state-to-state transitions); extracts dissimilarity measures by pair-wise comparison of graphs (e.g., nodes in graph A that are not in B); uses these dissimilarity measures as features for a novel Support Vector Machine to classify the data as forewarning of a seizure event or not (i.e., not characteristic of the baseline, but characteristic of data near the event). The present work obtains a prediction distance as small as zero (sensitivity =1, specificity =1) with most of the forewarning times ≤1.5 hours. The best off-training-set error rate was.287 using 10-fold cross validation.

Our non-invasive (scalp) EEG analysis resulted in cross validation error rates comparable to other invasive EEG approaches. Additional accuracy could be obtained by applying this methodology to specific patients on a per patient basis for custom EEG models if the data were available. Modifications could conceivably be made to the algorithm to improve the computational feasibility of per patient machine learning models. A research team could allow patients to start with group-based models that are less accurate while the patients collect and upload ambulatory data from their devices as they use them in real-world settings. Furthermore, businesses could be compensated for creating patient specific models from patients' ambulatory data. Additionally, our algorithms have other applications as well, such as failure forewarning in machines [21] and bridges [22].

Acknowledgements

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government

Author details

W.S. Ashbee[1], L.M. Hively[2] and J.T. McDonald[3]

1 School of Computing at University of South Alabama, USA

2 Computational Sciences and Engineering Division at Oak Ridge National Lab, USA

3 School of Computing at University of South Alabama, USA

References

[1] Netoff T, Park Y, Parhi K. Seizure prediction using cost-sensitive support vector machine. Engineering in Medicine and Biology Society (2009) EMBC 2009 Annual International Conference of the IEEE, pp. 3322–5.

[2] Stacey W, Le Van Quyen M, Mormann F, Schulze-Bonhage A, "What is the present-day EEG evidence for a preictal state?" *Epilepsy Res.* (2011);97(3):243–51.

[3] Mirowski PW, LeCun Y, Madhavan D, Kuzniecky R. Comparing SVM and convolutional networks for epileptic seizure prediction from intracranial EEG. Machine Learning for Signal Processing (2008) IEEE Workshop on MLSP, pp. 244–9.

[4] Hively LM, McDonald JT, Munro NB, Cornelius E, "Forewarning of Epileptic Events from Scalp EEG," peer-reviewed proceedings paper for *Biomedical Science and Engineering Conference at ORNL* (May 2013).

[5] Hively LM, Protopopescu VA, Munro NB, "Enhancements in epilepsy forewarning via phase-space dissimilarity," *J Clin Neurophysiol*. (2005);22(6):402–9.

[6] Percha B, Dzakpasu R, Żochowski M, and Parent J, "Transition from local to global phase synchrony in small world neural network and its possible implications for epilepsy," *Phys. Rev. E* (2005);72, paper #031909

[7] Van den Broeck C, Parrondo JMR, and Toral R, "Noise-induced nonequilibrium phase transition," *Phys. Rev. Lett.* (1994);73, 3395-3398

[8] Pittau F, Tinuper P, Bisulli F, Naldi I, Cortelli P, Bisulli A, et al., "Videopolygraphic and functional {MRI} study of musicogenic epilepsy. A case report and literature review," *Epilepsy & Behavior* (2008);13(4):685 – 692.

[9] Jenkins JS. "The Mozart effect," *J. Royal Soc. Med.* (2001); 94:170-172

[10] "10-20 System (EEG)."*Wikipedia*. Wikimedia Foundation, 22 July 2013. Web. 16 Aug. 2013.

[11] Hively LM, Protopopescu VA, "Channel-consistent forewarning of epileptic events from scalp EEG," *IEEE Trans Biomed Eng*. (2003);50(5):584–93.

[12] Hively LM, "Prognostication of Helicopter Failure," *ORNL/TM-2009-244*, Oak Ridge National Laboratory, Oak Ridge, TN (2009).

[13] Protopopescu VA, Hively LM, Gailey PC, "Epileptic event forewarning from scalp EEG," *J Clin Neurophysiol*. (2001);18(3):223–45.

[14] Hively LM, *et al.*, "Nonlinear Analysis of EEG for Epileptic Seizures," *ORNL/TM-12961*, Oak Ridge National Laboratory, Oak Ridge, TN (1995).

[15] Takens F, "Detecting strange attractors in turbulence," In: Rand D, Young L-S, editors. *Dynamical Systems and Turbulence*, Warwick 1980 [Internet]. Springer Berlin Heidelberg; 1981. p. 366–81.

[16] Bondy JA, Murty USR, *Graph Theory*, Springer (2008).

[17] Franzosi R, Pettini M, "Topology and phase transitions II. Theorem on a necessary relation," *Nuclear Physics B* (2007);782(3):219 – 240.

[18] Ng A. "Machine Learning: SVM Kernels." Coursera. Stanford University, n.d. Web. 10 Oct. 2013.

[19] Chang C-C, Lin C-J, "LIBSVM: A library for support vector machines" *ACM Trans. Intell. Syst. Technol.*(2011);2(3):27:1–27.

[20] Arthurs S, Zaveri HP, Frei MG, Osorio I, "Patient and caregiver perspectives on seizure prediction," *Epilepsy Behav*. (2010);19(3):474–7.

[21] Protopopescu V, Hively LM, "Phase-space dissimilarity measures of nonlinear dynamics: Industrial and biomedical applications," *Recent Res. Dev. Physics* (2005);6, 649-688.

[22] Bubacz JA, Chmielewski HT, Pape AE, Depersio AJ, Hively LM, Abercrombie RK, "Phase Space Dissimilarity Measures for Structural Health Monitoring," (2011) *ORNL/TM-2011/260* (Oak Ridge National Laboratory).

IntechOpen