Mach Learn
DOI 10.1007/s10994-015-5519-7

Multi-task seizure detection: addressing intra-patient variation in seizure morphologies

Alex Van Esbroeck[1] · Landon Smith[1] · Zeeshan Syed[1] · Satinder Singh[1] · Zahi Karam[1]

Received: 15 January 2015 / Accepted: 19 June 2015

Abstract The accurate and early detection of epileptic seizures in continuous electroencephalographic (EEG) data has a growing role in the management of patients with epilepsy. Early detection allows for therapy to be delivered at the start of seizures and for caregivers to be notified promptly about potentially debilitating events. The challenge to detecting epileptic seizures, however, is that seizure morphologies exhibit considerable inter-patient and intra-patient variability. While recent work has looked at addressing the issue of variations across different patients (inter-patient variability) and described patient-specific methodologies for seizure detection, there are no examples of systems that can simultaneously address the challenges of inter-patient and intra-patient variations in seizure morphology. In our study, we address this complete goal and describe a multi-task learning approach that trains a classifier to perform well across many kinds of seizures rather than potentially overfitting to the most common seizure types. Our approach increases the generalizability of seizure detection systems and improves the tradeoff between latency and sensitivity versus false positive rates. When compared against the standard approach on the CHB–MIT multi-channel scalp EEG data, our proposed method improved discrimination between seizure and non-seizure EEG for almost 83 % of the patients while reducing false positives on nearly 70 % of the patients studied.

Keywords Epilepsy · Seizure detection · Multi-task learning

1 Introduction

Epilepsy is a neurological disorder of the central nervous system that predisposes individuals to experiencing recurrent seizures. It affects approximately 1 % of the world's population

Editors: Byron C. Wallace and Jenna Wiens.

✉ Alex Van Esbroeck
alexve@umich.edu

[1] University of Michigan, Ann Arbor, MI, USA

Published online: 11 July 2015

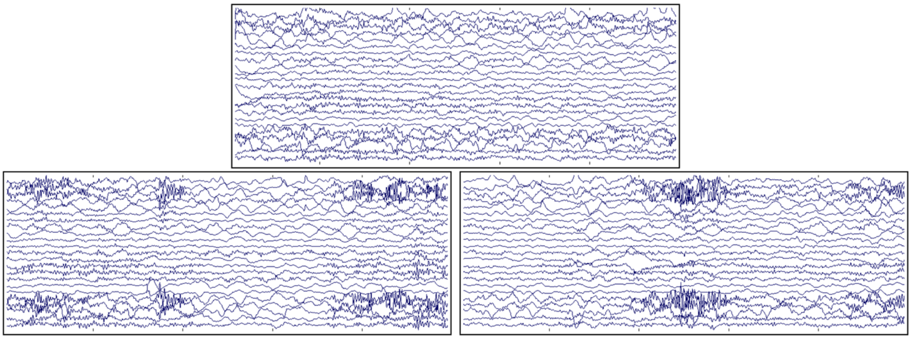

Fig. 1 A dissimilar seizure (*top*) and two similar seizures (*bottom*) from patient 23 of the CHB–MIT data

(Annegers 1997). The unpredictability of seizures can impact all aspects of a patient's life, and epilepsy has been found to be associated with higher rates of suicide, unemployment, and mood disorders such as depression (Strine et al. 2005).

One of the most valuable tools in diagnosing and treating epilepsy is the electroencephalogram (EEG), which records the electrical activity of the brain caused by the firing of millions of neurons, and can be used in an ambulatory setting. The problem of seizure onset detection in EEG recordings aims to automatically detect the start of seizures from the EEG. Accurate and timely seizure detection can be used to notify caregivers, as well as to allow for the development of adaptive nerve stimulation or drug release devices with the potential to reduce the severity of seizures, or prevent them entirely (Shoeb and Guttag 2010).

There has been a great deal of recent work on addressing seizure detection as a classification problem (Shoeb and Guttag 2010; Meier et al. 2008; Mirowski et al. 2009; Gardner et al. 2006). From a machine learning perspective, seizure detection is particularly challenging. As seizures can occur infrequently, there is usually limited ictal (seizure) data available for training a classifier. This scarcity is particularly troublesome in the face of motion artifacts and noise with similar characteristics to ictal data, as well as variability in seizure morphology both within the same patient (intra-patient) and across different patients (inter-patient). Quantifying detector performance presents additional challenges, as there are sharp tradeoffs between the relevant performance metrics (sensitivity, false positive rate, and latency) that make comparison of algorithms difficult.

The earliest work on seizure detection grouped together seizure data from multiple patients to learn a generic seizure onset classifier (Gotman and Gloor 1976; Gotman 1982). These generic seizure detection approaches suffer from aggregating data across patients, as seizure morphology can differ substantially between patients (due to variations in the neuroanatomical and pathophysiological causes of epileptic disease), even when comparing ictal to interictal (non-seizure) data. To account for inter-patient variability in seizure morphology, more recent work on seizure detection has explored patient-specific approaches that train seizure detectors on a per patient basis (Qu and Gotman 1997; Shoeb and Guttag 2010; Mirowski et al. 2009). Other extensions of this work describe how these patient-specific approaches can be implemented in a scalable manner with limited dependence on expert knowledge (Balakrishnan and Syed 2012).

While these efforts to address inter-patient variability in seizure morphology are promising, we observe that from a physiological perspective, many epileptic patients have seizure morphologies that vary significantly between events (Kharbouch 2012). This intra-patient variability superimposed on top of inter-patient variability represents an added confounder to efforts to train classifiers for seizure onset detection. An example of this is presented in

Fig. 1, which shows 5 s of multi-channel EEG collected from three different seizures of the same patient. The bottom two show similar activity with bursts of high frequency and high amplitude activity that is not present in the top, indicating that this patient has at least two distinct seizure types.

The seizure detection classification task has previously been thought of as distinguishing between all ictal and interictal data. However, this formulation does not distinguish between ictal data from different seizures or seizure types. Each seizure spans many windows, and this task makes no distinction between windows in different seizures, seeking to optimize the accuracy of individual window detections. As a result, this approach fails to consider the model's performance on the number of different seizures detected. In terms of classification loss, it may be preferable for such models to correctly classify all examples from one seizure type, rather than to correctly classify a fraction of examples from all seizure types. The classifiers used may optimize accuracy on some seizure types at the expense of others, particularly when the number of seizures available for training is limited.

Given the need to alert patients or caregivers to as many seizures as possible, the presence of multi-form seizures forces standard classification models to broadly encompass seizure data, leading to high numbers of false positives. This excess of false positives can place undue burden on caregivers and patients (e.g., alarm fatigue), and substantially increase the power utilization of implantable devices. False positives may also result in excess nerve stimulation or drug delivery in adaptively controlled devices, which can incur undesirable side-effects on patients (Ben-Menachem 2002).

Focusing seizure onset detection on fitting to specific seizure morphologies offers the promise of better modeling a patient's events, potentially reducing the number of false positives incurred in achieving a necessarily high sensitivity. The challenge, however, is that many patients experience seizures infrequently (with the sudden and unexpected onset of these events being associated with substantial burden) and there is often insufficient data to train classifiers specific to each of a patient's seizure types.

We propose to address these limitations of existing methods and data availability by approaching seizure detection with a multi-task learning framework. We consider distinguishing the windows of each seizure from interictal data as a separate task, and learn all individual-seizure discrimination tasks in conjunction. By leveraging a formulation of multi-task learning that couples the parameters of individual tasks (seizures), the proposed approach bootstraps shared knowledge between seizures of different morphologies to identify a discriminative component shared across all tasks for separating seizure EEG data from interictal EEG data. This formulation of the problem allows the classifier to identify common structure present across all types of seizures observed. The discriminative information shared across all seizures can be used in a detector, optimized to achieve high sensitivity for different seizure types with fewer false positives.

The contributions of this paper are: (1) we propose patient-specific seizure onset detection in the presence of intra-patient variability in seizure morphology; (2) we formulate this problem within a multi-task learning framework; (3) we describe the use of shared structure across seizures to address the issue of intra-patient variability during patient-specific seizure detection; (4) we present a task parameter coupled SVM multi-task learning algorithm that can be applied to solve this problem formulation; and (5) we rigorously examine the improvements offered by a multi-task seizure detection approach relative to the best performing current method on a representative real-world EEG dataset.

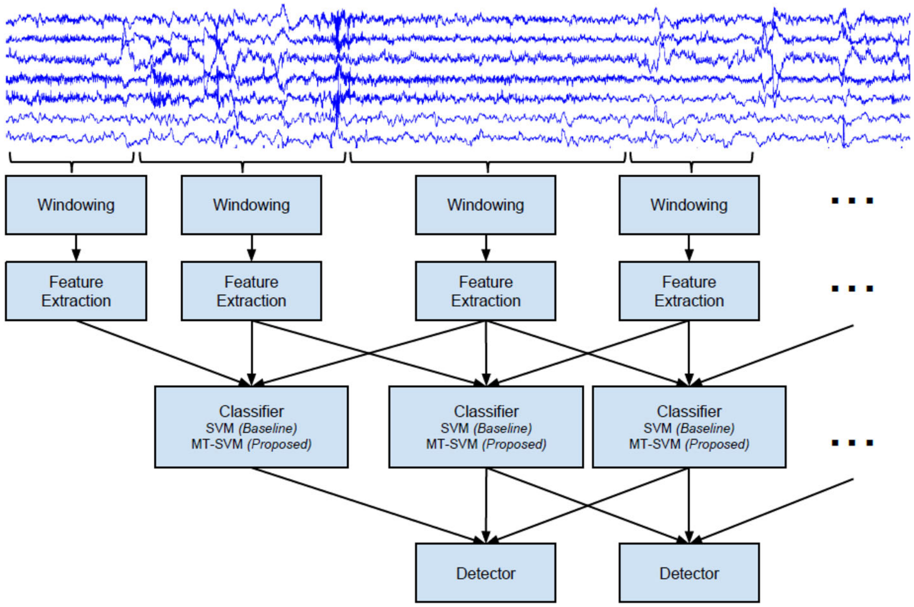

Fig. 2 Patient-specific seizure detection overview

2 Methods

We first present an overview of the standard patient-specific seizure detection framework. We then detail the proposed approach to extend this framework to better handle intra-patient variability.

2.1 Patient-specific seizure detection

By learning seizure detectors on a per-patient basis, patient-specific seizure detectors effectively handle inter-patient differences in seizure morphology. For this reason, patient-specific detectors have been shown to outperform generic seizure detectors in a number of different studies (Qu and Gotman 1997; Shoeb and Guttag 2010; Mirowski et al. 2009).

The training of these detectors almost always comprises the following steps: EEG data are first segmented into windows, spectral energy features are then extracted from each window, and finally these features are used to train a seizure classifier. This classifier can be subsequently be leveraged for detection of seizures. Figure 2 provides an overview of the system and this section briefly reviews each of these steps. The subsequent section describes how we advance this work to address the issue of intra-patient variations in EEG seizure morphology.

2.1.1 EEG windowing

Traditional approaches to segmenting EEG data are grounded in the use of time-epochs. For example, the EEG is often divided into windows that are 2 s long. In this work we build upon recent research demonstrating that this process of EEG segmentation can be improved by using an adaptive segmentation approach that places boundaries where the energy of the signal is changing sharply, resulting in more meaningful windows. This approach has been used previously to improve the performance of seizure detectors based on spectral energy features (Balakrishnan et al. 2010).

We begin with an EEG signal $X[n] = [x_1[n] \ldots x_P[n]]$ with P channels. This segmentation uses the discrete form of the nonlinear energy operator (NLEO) (Agarwal et al. 1998) to identify points in each channel where the signal energy is changing. The NLEO for channel i is defined as:

$$\Psi_i[n] = x_i[n-1]x_i[n-2] - x_i[n]x_i[n-3] \tag{1}$$

Segment boundaries in each channel are identified by using the NLEO with a sliding window. $G_\Psi[n]$ then measures the sum of the absolute difference in frequency-weighted energy between the left and right halves of the length $2N$ window centered at sample n over all P channels:

$$G_\Psi[n] = \sum_{i=1}^{P} \left| \sum_{m=n-N+1}^{n} \Psi_i[m] - \sum_{m=n+1}^{n+N} \Psi_i[m] \right| \tag{2}$$

A high value for $G_\Psi[n]$ indicates a large energy change whereas a value of zero indicates no change. To detect segmentation boundaries a threshold values $T[n]$ is applied:

$$T[n] = \begin{cases} \max\left[G_\Psi\left[n - \frac{L}{2} : n + \frac{L}{2}\right]\right], & n \geq \frac{L}{2} \\ 0, & n < \frac{L}{2} \end{cases} \tag{3}$$

where the parameter L determines how many segment boundaries are created. The final segmentation boundaries are detected by finding the local maxima of the thresholded function $G[n]$:

$$G[n] = \begin{cases} G_\Psi[n], & G_\Psi[n] \geq T[n] \\ 0, & G_\Psi[n] < T[n] \end{cases} \tag{4}$$

This adaptive segmentation provides more natural boundaries than a simple division of the signal into equally-sized time-based windows, capturing the EEG's inherent structure, and resulting in more comparable segments.

2.1.2 Feature extraction

Features for each EEG window are extracted corresponding to the energy of each channel at different time scales. Each channel of EEG is passed through an iterated filterbank structure, and the total energy in each of the four subband signals representing activity at frequencies from 0.5 to 25 Hz is measured, as in the work of Shoeb et al. (2004). The features for each channel are concatenated into a single feature vector for the window. For the final feature vector, we further concatenate the data from the two previous windows, creating a stacked feature vector; this allows the features to incorporate temporal variability information and was shown to improve classification in an earlier study (Shoeb and Guttag 2010).

2.1.3 Classification

At the core of seizure detector is a patient-specific classifier that operates on the extracted features and aims to separate windows containing seizures from those that do not. Typically this step pools all of the seizure features into one class and all of the non-seizure features in another. SVMs are a common choice when training a model for seizure detection (Shoeb and Guttag 2010; Meier et al. 2008; Mirowski et al. 2009; Gardner et al. 2006). The details of the standard SVM model and the proposed approach are detailed in the next section.

2.1.4 Detection

The erratic behavior of EEG and its susceptibility to artifacts results in many seizure-like windows of EEG that would trigger a seizure classification. Raising an alarm at each of these would lead to a detector that has a very high false-alarm rate. Therefore, it is important that seizure detectors not fire alarms at every seizure-like feature vector. The approach of Shoeb and Guttag (2010) uses the following two heuristics to generate alerts in a manner that is more suitable for deployment: the first defines a detection as three consecutive windows classified as positive instances while the second turns off the detector for 5 min after an alarm is raised. The former heuristic reduces the number of erratic false alarms while the latter ensures that once an alarm is raised similar EEG activity that may follow does not trigger repeat alarms.

2.2 Handling intra-patient variability with multi-task learning

In this work, we propose augmenting the classification stage of patient-specific seizure detection (as described in the previous section) with multi-task learning to achieve a seizure detector that is robust not only to inter-patient variability (due to the patient-specific approach), but also to intra-patient seizure variability.

The basic idea underlying multi-task learning is to solve multiple related classification tasks together with the goal of exploiting shared structure between them. In particular, when there is limited data to build classifiers for each of these tasks individually, this sharing of common structure can significantly improve performance.

We leverage a task parameter coupling formulation of multi-task learning. However, in contrast to the standard application of multi-task learning (which uses this shared structure as a way to identify better solutions to individual tasks) *we focus our attention on the shared structure itself as the goal of the learning*. We use this as a means of bootstrapping shared knowledge between seizures of different morphologies in the absence of sufficient data to train classifiers individually for each seizure type. This allows for the training of a single detector per patient intended to generalize well across seizure types; including those types not encountered during training.

Details of our multi-task learning approach are presented below. For notation, we start first with a brief presentation of the two-class SVM.

The standard two-class SVM classification approach (Cortes and Vapnik 1995) learns a separating boundary between two classes of examples (i.e., positive and negative) in a way that tries to maximize the margin between the data points and the boundary. The SVM uses the sign of the decision function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ to classify a data point $\mathbf{x}$. This problem can be expressed more formally as finding a solution to:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \xi_i + \lambda\|\mathbf{w}\|^2 \qquad \textit{subject to}: \\ y_i\mathbf{w}^T\mathbf{x}_i \geq 1 - \xi_i \qquad \text{and} \qquad \xi_i \geq 0, \quad \forall i \in \{1, 2, \ldots, n\} \tag{5}$$

In our work, we draw upon the multi-task learning method for SVMs proposed by Evgeniou and Pontil (2004). This approach learns solutions for T tasks using a separate classification function for each task t, $f_t(\mathbf{x}) = \mathbf{w}_t^T\mathbf{x}$. The task-specific separating hyperplane $\mathbf{w}_t$ is defined as:

$$\mathbf{w}_t = \mathbf{w}_0 + \mathbf{v}_t, \tag{6}$$

where $\mathbf{w}_0$ is shared across all tasks, and $\mathbf{v}_t$ is specific to each task t. When the vectors $\mathbf{v}_t$ are large relative to $\mathbf{w}_0$, the task-specific components dominate the shared component and each task may have a very different classifier. When $\mathbf{v}_t$ is small relative to $\mathbf{w}_0$, all of the tasks have very similar classifiers. To obtain these separating hyperplanes, one can solve the optimization problem:

$$\begin{aligned} &\min_{\mathbf{w}_0,\mathbf{v}_t,\xi_{it}} \sum_{t=1}^{T}\sum_{i=1}^{m}\xi_{it} + \frac{\lambda_1}{T}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2 + \lambda_2\|\mathbf{w}_0\|^2 \\ &\textit{subject to}\colon\ y_{it}(\mathbf{w}_0+\mathbf{v}_t)^T\mathbf{x}_{it} \geq 1-\xi_{it} \quad \text{and} \quad \xi_{it}\geq 0 \\ &\forall \quad i\in\{1,2,\ldots,n\} \quad \text{and} \quad t\in\{1,2,\ldots,T\} \end{aligned} \tag{7}$$

The regularization parameters λ_1 and λ_2 determine the cost parameter of the SVM as well as enforce the relatedness of the tasks: models trained using large ratios of $\frac{\lambda_1}{\lambda_2}$ ensure that all tasks have highly similar solutions.

It is important to note that the dual form of Eq. 7 is equivalent to the dual form of a standard SVM with the following feature map (for T tasks)

$$\Phi(\mathbf{x},t) = \left(\frac{\mathbf{x}}{\sqrt{\mu}}, \underbrace{\mathbf{0},\ldots,\mathbf{0}}_{t-1}, \mathbf{x}, \underbrace{\mathbf{0},\ldots,\mathbf{0}}_{T-t}\right), \tag{8}$$

where $\mu = \frac{T\lambda_2}{\lambda_1}$. This means that we can solve this multi-task formulation simply by solving a standard SVM, as in Eq. 5, with the feature mapping in Eq. 8. In this formulation, the regularization coefficient λ in Eq. 5 corresponds to λ_1/T. Once solved, the multi-task $\mathbf{w}_0$ and $\mathbf{v}_t$ can be obtained from $\mathbf{w}$ of the standard SVM since $\mathbf{w}$ consists of a stacking of the $\mathbf{w}_0$ and $\mathbf{v}_t$:

$$\mathbf{w} = (\sqrt{\mu}\mathbf{w}_0, \mathbf{v}_1, \ldots, \mathbf{v}_T) \tag{9}$$

The goal of using this multi-task SVM approach is to build a patient-specific seizure detector that takes into consideration intra-patient seizure variability in an unsupervised manner, i.e., without the need to explicitly label seizure types or otherwise categorize seizures together. We accomplish this by treating each individual seizure as a separate task. This results in a collection of seizure-specific detectors, corresponding to $\mathbf{w}_t$, the multi-task hyperplanes for each task/seizure. These hyperplanes are a combination of $\mathbf{w}_0$, the discriminative component shared across all tasks, and the task-specific components $\mathbf{v}_t$. To achieve our goal of producing a single classifier that generalizes well to all seizure types, we discard the task-specific vectors $\mathbf{v}_t$ and use only the shared component $\mathbf{w}_0$ for classification on all seizures. This leads to the classification function:

$$f(\mathbf{x}) = \mathbf{w}_0^T\mathbf{x} \tag{10}$$

The task-specific vectors $\mathbf{v}_t$ can be thought of as capturing seizure-specific characteristics, while the shared component $\mathbf{w}_0$ captures the patient-specific structure shared across all types of seizures. Although the training results in a single hyperplane $\mathbf{w}_0$, unlike the standard SVM solution this shared hyperplane is trained to optimize performance across the separated tasks (seizures) while discarding discriminant directions that are specific to particular seizures or seizure types. In standard SVM classification, failure to consider these shared characteristics may cause overfitting to the dominant seizure type, yielding poor generalizability to unseen

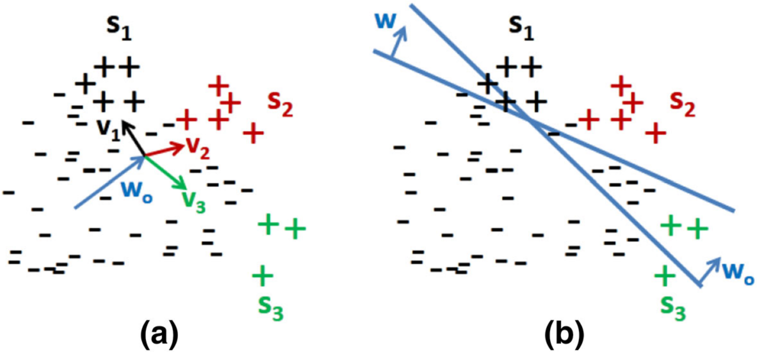

Fig. 3 **a** Sketch of common component $\mathbf{w_o}$ and seizure-specific components $\mathbf{v_{1-3}}$. **b** Sketch of the resultant hyperplane corresponding to standard SVM decision $\mathbf{w}$ and the resultant hyperplane corresponding to our approach $\mathbf{w_o}$

and under-represented seizure types. By tackling the issue of generalization across seizure types, our approach reduces the false positives incurred in achieving high sensitivity.

Figure 3 illustrates our approach through a simple two-dimensional sketch showing examples of windows from 3 seizures, s_1, s_2, s_3, represented by '+' and non-seizure windows represented by '−'. Figure 3a sketches the common discriminant direction $\mathbf{w}_o$ shared among all seizures and the seizure-specific directions $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ learned by the multi-task SVM. Figure 3b sketches the resultant hyperplane when only the shared direction $\mathbf{w}_o$ is used for classification, and contrasts it with the hyperplane resulting from the standard SVM discriminant direction $\mathbf{w}$. The use of the shared direction from multi-task learning allows for a decision boundary that does not favor the more highly represented seizures s_1 and s_2, but instead generalizes to achieve higher accuracy on seizure s_3 as well.

We note that in many datasets there is no grouping of seizures into different types. This is because determining seizure types depends on skilled expertise and the availability of fine-grained clinical labels. It is also highly subjective with considerable disagreement between experts. These issues are significantly more challenging than simply detecting the presence of seizures by EEG technicians.

The proposed method addresses this by treating each seizure as an individual type. While the task coupling formulation presented above is general enough to handle cases where multiple seizures may be grouped into specific types based on available prior knowledge, the representation of each seizure as a separate type allows for the proposed approach to scale to a broad range of datasets in the absence of any additional distinctions between seizures. By leveraging the EEG structure shared across different seizure types, our approach sidesteps the need to explicitly partition seizures into different seizure types for training. By focusing on the core of what separates seizure from non-seizure data for a patient, the resultant classifier generalizes well both to seizure types that may be under-represented in the training data, as well as to potentially unseen types.

3 Experiments

3.1 Data and experimental setup

We evaluated our approach on the Children's Hospital of Boston–Massachusetts Institute of Technology (CHB–MIT) dataset containing scalp EEG data for 23 pediatric subjects (Shoeb

2009). The CHB–MIT dataset is publicly available on Physionet (Goldberger et al. 2000). The evaluation dataset consisted of a total of 969 h of EEG (min 17, mean 42, max 158) and 173 seizures (min 3, mean 8, max 33). The international 20–10 system of electrode placement was used, with all subjects having over 20 channels of data.

To create each patient's classifier we used leave-one-out cross-validation over that patient's approximately hour-long EEG records in the CHB–MIT database, training on all records but one and then testing using the single held out record (which may contain one or more seizures). Due to the abundance of non-seizure data and the large amount of computation needed, we subsampled the non-seizure (negative) instances used for training, using only the first one out of every 15 non-seizure windows. This fixed-period subsampling was chosen over the use of true random sampling to allow reproducibility of the results on the publicly available CHB–MIT database. Additionally, the first 30 non-seizure instances (approximately 1 min) after a seizure were ignored, to account for a post-ictal period. To facilitate learning the earliest part of seizure onset and therefore improve detection latency, only the first 10 windows (about 20 s) of each seizure were used as positive instances. These choices were made to be consistent with previous studies (Shoeb and Guttag 2010). No subsampling of non-seizure instances was used in the validation or testing data, to allow accurate reporting of false positive rates (FPR).

For our multi-task learning based SVM approach, the regularization parameters λ_1 and λ_2 were chosen using a second round of leave-one-out cross-validation. In each round of the cross-validation over a patient's hour-long EEG records used to train and evaluate the models, an additional round of leave-one-out cross-validation was conducted on the training records to select these parameters. Parameters were chosen using a grid search, where λ_1 was selected from [1, 10, 100, 1000, 10,000], and λ_2 was selected from [0.01, 0.1, 1, 10, 100, 10,000]. As there were far more negative instances than positive instances, a cost-sensitive loss function was used in training the SVM, with negative instances weighted by $\frac{N_{instance+}}{N_{instance-}}$. Because seizures had varying lengths (with some seizures lasting fewer than 20 s, and accordingly having fewer positive instances), this weight varied between different tasks (seizures).

We compared our multi-task learning based SVM approach for training a seizure detector to the best reported patient-specific approach on the CHB–MIT dataset, based on traditional two-class SVM classification (Shoeb and Guttag 2010). This approach makes no distinctions between a patient's seizures or seizure types. The cost parameter C was chosen in this case using the cross-validation method described above for the proposed approach, with C ranging over $\{10^{-7,-6,\ldots,8}, 50, 500\}$. The same cost-sensitive weighting was used for the regular SVM. All pre-processing of the data was consistent with the multi-task approach, including the use of adaptive EEG segmentation, the only difference was in the use of a standard two-class SVM for classification. LIBLINEAR v1.8 was used for the implementations of both the standard and multi-task versions of the SVM (Fan et al. 2008).

3.2 Evaluation criteria

To compare the proposed method with the standard SVM, we computed a number of performance metrics both at the classifier level and the detector level. Since the evaluated approaches are patient-specific, classifiers/detectors are trained for each patient using only that patient's data. Therefore, our analyses treat each patient separately and we show that the classifiers/detectors generalize to unseen patient data (and in turn possibly unseen seizure types) by leaving out hour long epochs.

For comparison at the classifier level we calculated the area under the receiver operating characteristic curve (AUC) by combining the predictions from all cross-validations, one run per epoch, and using the decision values output from the classifiers to compute a receiver operating characteristic (ROC) curve. The use of AUC provided a performance metric that was independent of the choice of classifier detection threshold. The ROC describes classifier performance on individual EEG windows, a good measure in the context of classification, however not a good measure of real-world performance. The AUC is therefore not our main result and is shown here for completeness.

For a more meaningful representation of real-world performance we examine the sensitivity, latency, and FPR of the seizure detector, which aggregates the output of multiple consecutive classification outputs. Sensitivity was defined as the percentage of seizures with a detection at any point during the seizure. Latency was defined as the number of seconds elapsed between the onset of the seizure and its first detection. A false positive was any detection occurring outside of a seizure, unless fewer than 5 min had elapsed since the end of the most recent seizure, to allow for brain activity to return to normal. The false positive rate was defined as the number of false positives divided by the total duration of the data, in hours.

Sensitivity, latency, and FPR are all important when evaluating seizure detectors. Fair comparison of the two algorithms across these metrics is particularly difficult due to the complications added by using the SVM classifier predictions in the seizure detector framework. Adjusting the decision threshold does not provide a smooth trade-off between sensitivity/latency/FPR, and moreover one algorithm may perform better along two dimensions and worse along the third. Unfortunately, the trade-off between the three metrics as the decision threshold is shifted makes it impossible to match the methods in terms of FPR and latency and compare using sensitivity, or to match in terms of FPR and sensitivity and compare using latency.

Therefore, to accomplish a fair comparison we compare the false positive rates between the best reported methodology and our multi-task learning approach where the experiments were designed to hold sensitivity and latency as fixed between the models (i.e., to prevent one algorithm from achieving a better FPR by compromising on either sensitivity or latency). We determined a priori that (1) it was essential not to miss any seizures, and therefore required both methods to detect all of the seizures in the recording (i.e., 100 % sensitivity), and (2) to support prompt intervention the methods should operate at the lowest possible latency that could be supported by both approaches. The classifier thresholds for flagging seizures were set accordingly, and the resulting FPR are reported in our paper. For cases where there existed no exact match of latencies with 100 % sensitivity, the nearest match was chosen.

4 Results

Table 1 shows the results of the experiments. At a per-window classification level, the use of multi-task learning led to small but consistent increases in AUC for 19 of the 23 cases studied. At a per-seizure level, the use of multi-task learning led to improvements in FPR when the two approaches were matched by sensitivity and latency as described above. The relative reductions in FPR were more pronounced than the AUC changes, with 15 of the 23 cases showing an improvement in FPR greater than 10 % (compared to only six of the cases showing a worsening in FPR of more than 10 %). The median overall improvement in FPR was 27 %.

The table shows that even though the latency on the majority of the patients is below 15 s, the FPR are quite high for several patients. This result can be attributed, in part, to

Table 1 Performance of the standard SVM and the proposed method on each subject in the CHB–MIT dataset, as well as the number of seizures (# Sz.) per subject

Subject	# Sz.	SVM			Proposed approach			Δ
		AUC	Latency	FPR	AUC	Latency	FPR	%ΔFPR
01	7	.996	10.29	0.05	.997	10.29	0.00	**100.00**
02	3	.988	6.67	1.78	.983	6.67	3.92	−120.31
03	7	.973	13.43	21.37	.974	13.43	20.03	**6.27**
04	4	.952	17.50	16.37	.953	17.50	6.73	**58.88**
05	5	.991	10.80	0.25	.993	10.80	0.00	**100**
06	10	.981	2.80	4.15	.981	2.80	2.23	**46.39**
07	3	.998	2.00	0.00	.999	2.00	0.04	$-\infty$
08	5	.934	12.00	0.67	.936	12.00	3.23	−378.57
09	4	.992	8.50	0.49	.993	8.50	0.13	**73.53**
10	7	.999	3.43	0.17	.999	3.43	0.08	**55.56**
11	3	.992	10.67	0.09	.992	10.67	0.03	**66.67**
12	33	.821	3.88	1472.55	.824	3.82	827.58	**43.80**
13	12	.972	8.17	1.62	.973	8.17	1.23	**24.00**
14	8	.973	2.25	3.35	.977	2.25	2.44	**27.17**
15	20	.976	11.90	5.05	.976	11.80	5.10	−0.99
16	8	.832	23.50	66.40	.847	23.23	60.89	**8.30**
17	3	.895	21.33	5.28	.916	21.33	6.59	−25.00
18	6	.911	9.33	17.99	.934	9.33	14.27	**20.68**
19	3	.995	15.33	0.03	.999	15.33	0.00	**100**
20	8	.936	6.75	0.70	.936	6.75	1.19	−70.00
21	4	.956	3.00	1.21	.957	3.00	1.52	−25.64
22	3	.994	12.00	0.54	.994	12.00	0.25	**52.94**
23	7	.996	8.29	2.27	.996	8.29	0.80	**64.52**
Median	7	.976	9.33	1.62	.977	9.33	1.52	**27.17**

Bold values indicate a reduction in false positive rate using the proposed approach.
Area under the ROC curve (AUC), latency (seconds), false positives per hour (FPR), as well as the percentage decrease in the FPR from the standard SVM to the proposed approach are shown. All results are generated by detectors with 100 % sensitivity

the availability of scalp data in the CHB–MIT repository (which has more artifacts than intra-cranial or deep-brain data) and a conservative definition of false positive which treats contiguous false positives as multiple occurrences despite being part of a single artifact. These contiguous time-localized regions containing large numbers of false positives were also observed in earlier studies (Kharbouch 2012). Therefore, detectors typically include a built-in no-trigger zone of several (1–5) min after a seizure is detected (Shoeb and Guttag 2010; Kharbouch 2012).

Unfortunately, incorporating such a zone complicates the relationship between sensitivity, latency, and FPR, making a matching of sensitivity and latency between the two approaches impossible in many cases. However, when a 1 min no-trigger zone was included in the seizure detectors compared here, our proposed detector using multi-task learning achieved 100 % sensitivity with a latency of less than 15 s for 15 patients and an average FPR of 0.42. In contrast, the SVM approach previously proposed for patient-specific seizure detection achieved 100 % sensitivity with a latency less than 15 s for 14 (one less than the proposed) patients and a higher average FPR of 0.58. This indicates that after including the no-trigger zone the proposed method reduced the false positive rate to nearly a third of the standard approach, with matching sensitivity and latencies. These numbers are potentially more meaningful indicators of the real-world performance of the classifiers. It is worth noting that there is no general consensus on what FPR/latency/sensitivity is acceptable, as this is highly dependent on the situation (e.g., seizure rate/severity, intervention cost/type).

5 Conclusion

In this study, we addressed the issue of detecting epileptic seizures in the presence of inter-patient and intra-patient variability. We focused, in particular, on an approach based on multi-task learning that treats patient seizures as separate tasks and learns a discriminant direction common to all for use in detection. Failing to do this yields detectors that often overfit to certain dominant seizure types at the cost of generalizability to unseen or under-represented seizure types. The lack of generalizability in turn leads to detectors with high false positive rates, limiting their real-world utility.

When evaluated on real-world EEG data from 23 epileptic patients, our multi-task learning approach outperformed the standard methodology for patient-specific seizure detection in the majority of the patients. This improvement was present at both the per-window (AUC improvement in 19 of 23 cases) and per-seizure (FPR reduction of greater than 10 % in 15 cases using our multi-task approach vs. 6 using the standard approach) level. Moreover, when combined with the use of a no-trigger zone the use of multi-task learning improved seizure detection both in terms of the number of patients with a latency of less than 15 s and the average FPR. These improvements hold the opportunity to decrease alarm fatigue and streamline the delivery of therapy, and while they may appear small, they are substantial from the perspective of clinical studies and given the prevalence of epilepsy in the general population translates into potential improvements of hundreds of thousands of patients in the U.S. alone.

There are several limitations to the present work. While the proposed approach reduced FPR in over two-thirds of the patients, for six patients the rates increased non-negligibly. This discrepancy was not clearly associated with demographic characteristics of the patients or with the number of seizures. The clinical information on patients in the CHB–MIT dataset is limited, and with richer metadata it may be possible to better distinguish between patients who benefit from a multi-task learning approach. Additionally, the presence of outlier tasks is known to negatively impact the performance of multi-task learning approaches (Gong et al. 2012). This is a likely concern in seizure detection, where partial seizures may have widely varying characteristics. Unsupervised identification of outlier seizures and subsequent removal of these tasks from the training procedure could reduce noise in training improve the method’s performance. Previous work on seizure detection has found that for some patients the use of non-linear classifiers improves prediction. The multi-task SVM method applied in this paper can be used with kernels to learn non-linear seizure detectors, and investigation of the method’s utility in this setting is an area for future investigation. When incorporating the classifier into a seizure detector three consecutive positively classified windows were required to constitute a detection. Invesigation of different detection schemes could yield further improvements in FPR.

We believe that our work, though presented in a seizure detection context, has value in other chronic disease settings where abnormal events occur multiple times. Due to intra-patient variations in many of these settings (e.g., cardiac arrhythmias/ischemia, epileptic seizures, asthma etc.) these diseases remain a challenge diagnostically, but efforts such as our work provide a systematic way of reducing the complexity of these conditions.

References

Agarwal, R., Gotman, J., Flanagan, D., & Rosenblatt, B. (1998). Automatic eeg analysis during long-term monitoring in the icu. *Electroencephalography and Clinical Neurophysiology*, *107*(1), 44–58.

Annegers, J. F. (1997). *The treatment of epilepsy: Principle and practice*. Baltimore: Williams and Wilkins.

Balakrishnan, G., Shoeb, A., & Syed, Z. (2010). Creating symbolic representations of electroencephalographic signals: An investigation of alternate methodologies on intracranial data. In *2010 Annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 4683–4686). IEEE.

Balakrishnan, G., & Syed, Z. (2012). Scalable personalization of long-term physiological monitoring: Active learning methodologies for epileptic seizure onset detection. In *International conference on artificial intelligence and statistics (AISTATS)* (pp. 73–81).

Ben-Menachem, E. (2002). Vagus-nerve stimulation for the treatment of epilepsy. *The Lancet Neurology*, *1*(8), 477–482.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Evgeniou, T., & Pontil, M. (2004). Regularized multitask learning. In *Proceedings of the tenth ACM international conference on knowledge discovery and data mining (SIGKDD)* (pp. 109–117).

Fan, R., Chang, K., Hsieh, C., Wang, X., & Lin, C. (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, *9*, 1871–1874.

Gardner, A., Krieger, A., Vachtsevanos, G., & Litt, B. (2006). One-class novelty detection for seizure analysis from intracranial eeg. *The Journal of Machine Learning Research*, *7*, 1025–1044.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220.

Gong, P., Ye, J., & Zhang, C. (2012). Robust multi-task feature learning. In *Proceedings of the ACM international conference on knowledge discovery and data mining (SIGKDD)* (pp. 895–903).

Gotman, J. (1982). Automatic recognition of epileptic seizures in the eeg. *Electroencephalography and Clinical Neurophysiology*, *54*(5), 530–540.

Gotman, J., & Gloor, P. (1976). Automatic recognition and quantification of interictal epileptic activity in the human scalp eeg. *Electroencephalography and Clinical Neurophysiology*, *41*(5), 513–529.

Kharbouch, A.A. (2012). *Automatic detection of epileptic seizure onset and termination using intracranial eeg*. Ph.D. thesis, Massachusetts Institute of Technology.

Meier, R., Dittrich, H., Schulze-Bonhage, A., & Aertsen, A. (2008). Detecting epileptic seizures in long-term human eeg: A new approach to automatic online and real-time detection and classification of polymorphic seizure patterns. *Journal of Clinical Neurophysiology*, *25*, 119–131.

Mirowski, P., Madhavan, D., LeCun, Y., & Kuzniecky, R. (2009). Classification of patterns of eeg synchronization for seizure prediction. *Clinical Neurophysiology*, *120*(11), 1927–1940.

Qu, H., & Gotman, J. (1997). A patient-specific algorithm for the detection of seizure onset in long-term eeg monitoring: Possible use as a warning device. *IEEE Transactions on Biomedical Engineering*, *44*(2), 115–122.

Shoeb, A. (2009). *Application of machine learning to epileptic seizure onset detection and treatment*. Ph.D. thesis, Massachusetts Institute of Technology.

Shoeb, A., Edwards, H., Connolly, J., Bourgeois, B., Ted Treves, S., & Guttag, J. (2004). Patient-specific seizure onset detection. *Epilepsy and Behavior*, *5*(4), 483–498.

Shoeb, A., & Guttag, J. (2010). Application of machine learning to epileptic seizure detection. In *Proceedings of international conference on machine learning (ICML)*.

Strine, T. W., Kobau, R., Chapman, D. P., Thurman, D. J., Price, P., & Balluz, L. S. (2005). Psychological distress, comorbidities, and health behaviors among U.S. adults with seizures: Results from the 2002 national health interview survey. *Epilepsia*, *46*(7), 1133–1139.