Detection of Seizures in Intracranial EEG: UPenn and Mayo Clinic’s Seizure Detection Challenge

Andriy Temko, Senior Member, IEEE, Achintya Sarkar, Gordon Lightbody, Member, IEEE

Abstract—A system for detection of seizures in intracranial EEG is presented that is based on a combination of generative, discriminative and hybrid approaches. We present a methodology to effectively benefit from the advantages each classifier offers. In particular, Gaussian mixture models, Support Vector Machines, hybrid likelihood ratio and Gaussian supervector approaches are developed and combined for the task. This system participated in the UPenn and Mayo Clinic’s Seizure Detection Challenge, ranking in the top 5 of over 200 participants. The drawbacks of the proposed method with respect to the winning solutions are critically assessed.

I. INTRODUCTION

Development of automated seizure detection algorithms has drawn the attention of the research community for a long time. A number of solutions have been proposed where the problem was tackled at feature and classifier levels and combinations thereof. At the feature level, time-frequency distributions are popular for extraction of meaningful features [1]. Methods from the information theory domain that focus on quantification of system (dis)order such as recurrent quantification analysis [2], or entropy-based parameters such as permutation entropy [3] have shown reasonable success. For the binary problem of seizure detection, a variety of classifiers have been found suitable ranging from generative density estimators such as Gaussian Mixture Models (GMM) [4] or 1-class classifiers [5] to more common discriminative methods such as the Support Vector Machines (SVM) [6] or deep neural networks [7].

Despite the fact that a few databases of EEG signals are publically available for seizure detection [8, 9] it is still not easy to identify genuine technical contributions among proposed solutions. The primary cause of this is the different levels of rigor which is followed by the authors. For instance, using a publically available dataset such as Freiburg EEG dataset [8] it is common to exclude worse performing patients using whatever justification possible. A different split of the dataset to train and test the classifier can greatly affect the performance, compromising the reproducibility of results which similarly prohibits the comparison among existing techniques. Only a few studies use statistical routines to assess the performance such as cross-validation, (e.g. leave-one-out) or bootstrap resampling [4, 6]. Moreover, on a given dataset, different tasks can be pursued such as development of a patient-independent [4, 6] or patient-dependent seizure detection system [10] or performing seizure prediction. Comparing solutions evaluated on different datasets is even more difficult as datasets vary in size, may consist of a mixture of intracranial and surface EEG [9]. Reporting different metrics or using different algorithmic latency of post-processing does not facilitate the comparison either.

Under these circumstances, the only way to assess the quality of various solutions is to use fixed metrics, task specifications and more importantly a common dataset. University of Pennsylvania and Mayo Clinic under the sponsorship of American Epilepsy Society have recently organised a Seizure Detection Challenge (SDC) using the Kaggle platform, http://www.kaggle.com/c/seizure-detection/. A dataset of intracranial EEG was made public, the task was clearly explained with the rules and metrics rigorously defined. Over 200 teams participated in the challenge from around the world.

In this paper, we present our approach to seizure detection which is based on a combination of generative, discriminative and hybrid classifiers, which are designed to maximise their inherent advantages in the context of the given data and specified rules. In the SDC, the developed system ranked 5th.

The paper is organized as follows. Section 2 describes the database, metrics and the performance assessment routine used in the study. The developed system is described in detail in Section 3. The results of the system on the provided data are presented in Section 4 where we critically assess its drawbacks in comparison with the winning solutions.

II. DATA AND TASK SPECIFICATIONS

A. Database

The data provided for the SDC consisted of intracranial EEG recordings, obtained from 4 dogs with naturally occurring epilepsy and 8 patients with temporal and extra-temporal lobe epilepsy undergoing evaluation for epilepsy surgery.

In dogs, the EEG was recorded from an implanted device which acquired data from 16 subdural electrodes [11]. Two 4-contact strips were implanted over each hemisphere in an anterior-posterior orientation. Data were recorded continuously at a sampling frequency of 400 Hz and referenced to the group average.

The human EEG recordings [12] were obtained from depth electrodes implanted along anterior-posterior axis of hippocampus, and from subdural electrode grids in various locations. The recordings from human patients have varying
numbers of electrodes and data sampling rates ranging from 500 Hz to 5,000 Hz, with recorded voltages referenced to an electrode outside the brain.

Data were organized in folders containing training and testing data for each human or dog subject, thus allowing and encouraging but not enforcing the development of patient-specific models. The data were provided in 1-second EEG clips. The training data were labelled ‘Ictal’ for seizure clips, and ‘Interictal’ for non-seizure clips, whereas the testing data were labelled ‘Test’. Starting points for the interictal data segments were chosen randomly from the full data record, with the restriction that no interictal segments were less than one hour before or after a seizure. The details of the dataset are outlined in Table I.

Every ictal clip in the training dataset had a defined latency: the time in seconds between the expert-marked seizure onset and the start of the clip. The clips that have latency smaller than 15s are considered as ‘Early’ ictal clips.

### B. Task specifications

The goal of this challenge was to tackle two specific applications. The first one was automated seizure diaries, where latency to seizure onset was not critical. The goal here was to optimize the accuracy of seizure detection by developing algorithms with high sensitivity and specificity. The second application was a responsive neuro-stimulation application where the successful therapy depends on the ability to rapidly detect the onset of seizures. In this task the latency of detected onset was of a particular importance and the goal was to develop algorithms that were accurate in detecting ‘Early’ ictal clips.

The developed algorithmic solution must be the same across all subjects so that the approach is capable of generalizing to new subjects. For instance, one cannot have one set of features and a neural network classifier for subject 1 and another set of features and a SVM for subject 2. The selection of the most suitable system parameters such as classifier, features, etc, if different for each subject must be done in an automated, data-driven way.

The usage of testing data for semi-supervised or unsupervised learning was not allowed in the competition. Any normalization of probability prediction should be designed using only the training data provided.

### B. Metrics

For each tested file (clip) the developed system must produce two probabilities. First, the probability that a given clip is ‘Ictal’ must be estimated. Second, the system must estimate the probability that the clip is ‘Early’ seizure, that is the clip is within the first 15 seconds its respective seizure (latency). The submission file starts with a header line which is followed by probability predictions:

\[
\text{clip, seizure, early}
\]

**Patient_1_test_segment_1.mat, 0.51, 0.34**

The submissions are judged based on the Area Under the Curve (AUC) which is computed from series of sensitivity and specificity obtained from probability predictions for each subtask. The final competition metric is the mean of these two AUCs:

\[
\text{AUC}_{\text{final}} = \frac{1}{2}(\text{AUC}_{\text{seizure}} + \text{AUC}_{\text{early}})
\]  

As follows from Eq. 1 and was further clarified on the competition forum, a single metric is computed to score all the clips of all the subjects in the dataset. This implies that the well-calibrated probabilities will maximize the AUC metric. For instance, it is possible to have a per-subject AUC of 100% and still have a poor value for the final metric if the probability distributions are not properly aligned.

### III. METHOD

#### A. Preprocessing and Feature Extraction

The developed system is outlined in Fig. 1. The EEG signals are filtered using a band pass filter [0.5–128Hz] and down-sampled to 256Hz. The EEG signals are given in the referential montage by default. The bipolar montage was created by subtracting the channels which were believed to lie close to each other. As limited-to-no information was provided about channel locations this was carried out based on the Hamming distance between channel text descriptions. A set of seventy features (Table II) is then extracted from each channel of EEG. Fifty-five of the seventy features have previously been used in our research group for detection of seizures in newborns from surface EEG [4, 6, 13]. The frequency content of neonatal seizures lies within the first 16Hz so these 55 features have been extended with several parameters to capture the higher frequency content which is present in intracranial EEG, such as e.g. subband energies in 3-15Hz, 15-30Hz, 25-128Hz, bandwidth, etc. This large set of features can be seen as a universal descriptor of the intracranial EEG signal, and captures energy, frequency and structural (information theory) signal components extracted from various domains.

![Figure 1. The flowchart of the developed seizure detection system.](image-url)
B. Classifiers

Two classifiers were employed in this study, generative GMMs and discriminative SVMs. Both classifiers are designed in the subject-dependent mode as it was allowed by the competition rules and initial experiments and the literature [10] report substantial improvement when using subject-specific models over subject-independent ones.

It is a known advantage of the SVM that it can perform accurate modeling of the separating hyper-plane in a high-dimensional space. To benefit from this property, the features from all channels are concatenated to form a long feature vector. For instance, 1s clip of a subject with 16 EEG channels will result in a vector of 16*70 = 1120 features. These long feature vectors are normalized to zero mean and unit variance and the SVM classifier is trained on the training data of that subject (subject-specific) following a rigorous model selection routine to select the optimal SVM hyper parameters. The output of the SVM is transformed into a probabilistic value using Platt’s sigmoid function [14]. In order to assure that the predicted probability distributions are aligned across subjects, the same priors were reinforced when training the sigmoid function. Two models; ‘Ictal’ vs. ‘Interictal’ and ‘Early Ictal’ vs. ‘Late Ictal’ were obtained for each subject. During the testing stage, a given feature vector, \( x \), is normalized with the same normalization template computed from the training data. The first model outputs the probability of a clip being ‘Ictal’, \( P(\text{Ictal}|x) \). The second model outputs the probability of the clip of being ‘Early’ given that the clip is ‘Ictal’, \( P(\text{Early}|\text{Ictal}) \). Then, the probability of being ‘Early’ is modeled as:

\[
P(\text{Early}|x) = P(\text{Ictal}|x)P(\text{Early}|\text{Ictal})
\]

(2)

In contrast to SVMs, GMMs are not suitable for modeling a probability density function in such a high-dimensional space, as the complexity of the estimation problem increases with dimensionality of the data. However, the main advantage of the GMM is that it can be efficiently trained on millions of datapoints; something that is impossible with the SVM. To benefit from this property, we followed the approach used in speech processing [15]. First, we trained a Universal Background Model (UBM) – a GMM which is trained on all training data from all subjects in the dataset. This allows modeling a large space of possible feature behaviors from all available data. Before training the UBM, the features were not concatenated across channels as in the SVM case, but pooled together. After zero mean and unit variance normalization, the feature matrix is de-correlated with principal component analysis, keeping 98% of the variance. This simplifies the estimation problem and allows using diagonal covariance in the GMM training. After the UBM is trained, a sequence of model adaptation is performed using Maximum a Posteriori optimization criteria [15]. First, the UBM is adapted towards ‘Ictal’ and ‘Interictal’ classes using all interictal and ictal data from all subjects. The data of a particular subject are then used to adapt the generic ‘Ictal’ (‘Interictal’) model towards subject-specific ‘Ictal’ (‘Interictal’) model. Finally, the subject-specific ‘Ictal’ model is further adapted to the subject-specific ‘Early’ and ‘Late’ models with corresponding data. Eventually, four models are kept per subject – ‘Ictal’, ‘Interictal’, ‘Early’ and ‘Late’. In the testing stage, Bayes formula is used to convert the GMM likelihoods \( p(x|\text{Ictal}) \), \( p(x|\text{Interictal}) \), \( p(x|\text{Early}) \) and \( p(x|\text{Late}) \) to posterior probabilities, \( P(\text{Ictal}|x) \) and \( P(\text{Early}|\text{Ictal}) \) with equal priors assumed. Then, Eq. 2 is similarly used to estimate the \( P(\text{Early}|x) \). This showed better performance than contrasting the ‘Early’ model with the ‘Interictal’ model directly.

Apart from the two above-explained approaches, two hybrid classifiers were also constructed that improved the performance of the ensemble. In the context of pure generative and discriminative approaches they can be called hybrid because they are based on both. One is the Gaussian Supervector (GSV) approach which is used in speaker identification/verification tasks [15]. It follows the same process as explained in the previous paragraph, however instead of computing the probability of ‘Ictal’ from the obtained models, the means of the adapted UBM are used as features for a linear SVM.

The second hybrid approach uses the per-channel Log-Likelihoods Ratios (LLR) as features for the SVM classifier. The LLR are computed from the same GMM models with respect to the UBM. For instance, for a 1s clip of 16 channels, the feature vector that is fed to the SVM is of 32 dimensions consisting of 16 LLRs from the ‘Ictal’ model and 16 LLRs from the ‘Interictal’ model. In both hybrid approaches, the same SVM training, model selection and sigmoid probability conversion routines are used.

C. Ensemble of classifiers

The ensemble consists of six developed models: two SVM-based classifiers, for bipolar and referential montages (SVM\(_{\text{bip}}\), SVM\(_{\text{ref}}\)), two GMM-based classifiers, for bipolar and referential montages (GMM\(_{\text{bip}}\), GMM\(_{\text{ref}}\)), and two hybrid models (GSV, LLR) which were computed for referential montage only.

It was experimentally derived (on the development data) that the best performance was obtained when predictions that resulted from the systems that used the sigmoid probability conversion function were arithmetically averaged (summation) whereas predictions that resulted from Bayes formula were geometrically averaged (multiplication). The final probability is taken again as a product of SVM-based and GMM-based systems. The prediction for ‘Ictal’ can be expressed as:

\[
P(\text{Ictal}) = e^{\alpha \log \left( \frac{p(\text{Ictal}) + p(\text{SVM}_{\text{bip}}) + p(\text{SVM}_{\text{ref}}) + p(\text{GMM}_{\text{bip}})}{4} \right) + (1-\alpha) \log \left( \frac{p(\text{GMM}_{\text{ref}})}{p(\text{SVM}_{\text{ref}})} \right)}
\]

(3)
This representation allows weighting of SVM-based and GMM-based systems. For estimation of $P_{\text{Ictal}} \alpha=1/2$. The prediction for $P_{\text{Early}}$ is calculated using Eq. 3 with $\alpha=1/6$. It was observed in our system (and also reported on the SDC forum) that for $P_{\text{Early}}$ using the ‘Ictal’ probability instead of the ‘Early’ probability tends to give a small boost in performance for the pure SVM systems due to the high cost of making ‘Early’ errors introduced by the class imbalances (‘Interictal’ vs ‘Early’) that would affect any discriminative classifier. However, we observed that it was not the case with the pure generative models (GMM) for which the properly estimated ‘Early’ probability was used and this explains the 5 times higher weight of the generative models in Eq. 3.

IV. RESULTS AND DISCUSSION

Fig. 2 shows the result of the developed systems separately and in ensemble. The best performing system was based on the original referential montage, concatenation of features across EEG channels and the SVM classifier. In general, the pure SVM systems performed reasonably well followed by the pure GMM systems and the hybrid approaches. Combination of the developed systems resulted in a significant boost of the performance when comparing with the single best system (AUC of 93.09% vs. AUC of 95.44%). Despite the fact that the hybrid approaches performed worse than their pure counterparts, their contribution to the ensemble was important as their removal decreased the performance of the ensemble (as checked by the post-deadline submissions). The models that resulted from the bipolar montage performed consistently worse than their referential montage alternatives however the ensemble score similarly decreases if they are removed. Overall, it was observed that the submission that achieved best results on the development data also performed best on the test data.

The winner of the challenge used a Random Forest classifier and obtained the AUC of 96.29%. The runner-up used a combination of deep neural networks which resulted in the AUC of 95.65%. However, the main difference of their systems to the one presented here was the usage of cross-channel features [16]. These can be as simple as cross-correlation analysis of time-domain EEG signals. Obviously, these features capture seizure-location dependencies which are highly subject-specific. As our main experience lies in the neonatal area [17], where the usage of patient-dependent systems is not possible, we were overly focused on the overall system design rather than on the level of patient dependence introduced. The influence of the cross-channel features cannot be overestimated – both the winner and the runner-up reported the AUC of only 93-94% without using cross-channel features (the development data performance which is usually higher than the test data performance). In contrast, we managed to get the higher AUC of 95.44% on the test data without using any cross-channel features.

The experience we obtained from participating in this challenge also contributed to the improved neonatal seizure detector in our group as we realized the importance of the classifier ensemble in practical terms.

The SDC is open for post-deadline submissions from anyone who wants to check the accuracy of his/her solution. The challenge website provides a platform to benchmark the existing solutions to the seizure detection problem using the fixed dataset, rules, and metrics. Only under these conditions a fair comparison can be performed.

REFERENCES


Figure 2. The performance of the developed systems.