Matthias Eberlein, Jens Müller, Hongliu Yang, Simon Walz, Janina Schreiber, Susanne Creutz, Ortrud Uckermann, Georg Leonhardt, and Ronald Tetzlaff

Evaluation of machine learning methods for seizure prediction in epilepsy

https://doi.org/10.1515/cdbme-2019-0028

Abstract: Epilepsy affects about 50 million people worldwide of which one third is refractory to medication. An automated and reliable system that warns of impending seizures would greatly improve patient's quality of life by overcoming the uncertainty and helplessness due to the unpredicted events. Here we present new seizure prediction results including a performance comparison of different methods. The analysis is based on a new set of intracranial EEG data that has been recorded in our working group during presurgical evaluation. We applied two different methods for seizure prediction and evaluated their performance pseudoprospectively. The comparison of this evaluation with common statistical evaluation reveals possible reasons for overly optimistic estimations of the performance of seizure forecasting systems.

Keywords: epilepsy, seizure forecasting, seizure prediction, machine learning, deep learning

1 Introduction

Affecting about 1 % of the world population, epilepsy is one of the most common neurological diseases. Although seizures cover relatively short periods in a patient's life, the uncertainty when the next seizure will occur can produce a high level of anxiety [4]. For 70 % of the patients, medication can reduce the frequency of seizures or even abolish them. However, patients report that unwanted side effects of the medication as well as the unpredictability of seizures are the severest handicaps of this disease [13]. A mobile system with the ability to predict seizures can help to relief the patients' anxiety related to the uncertainty of events by enabling them to seek shelter, apply a short acting drug or inform the treating physician about the event. The device might also be used to prevent or mitigate the seizure [12].

Usually, seizure prediction is treated as a binary classification problem of brain activity, recorded as intracranial electroencephalography (icEEG) [8], with the state of impending seizures (*preictal*) being labeled as 1 and periods with a big temporal distance to the next seizure (*interictal*) labeled as 0. In this contribution, we present a new database that has been recorded in our working group. By intensifying the cooperation of clinical research and data analysis we minimize loss of descriptive metadata. For feature extraction and classification of the recorded icEEG signals we employed both, a recently proposed deep convolutional neural network and a feature-based method.

2 Materials and Methods

2.1 Data

The icEEG data set was recorded during presurgical diagnostics from five male patients (32 to 64 years old) who had invasive long-term EEG with subdural stripes, grids, and depths electrodes. The exact position of the electrodes could be identified in a postoperative 3-D T1-weighted MRI. For all patients, the seizure onsets zone could be identified. Patient specific details are depicted in Table 1. All data was subsampled to 200 Hz in order to reduce computational cost. To prevent data contamination by events exceeding the recorded duration, 4 h of data at the beginning and at the end of the recordings were discarded. Furthermore, dysfunctional channels (identified by visual inspection) were excluded from this study.

According to previous work [2, 9], we extracted non-overlapping segments of 600 s. However, we refrained from spacing the segments by gaps of 10 s. All icEEG data records were reviewed and seizures were annotated by an experienced neurologist. Similar to [2, 9], *preictal* was defined as the period starting 65 min and ending 5 min before each seizure onset. A period of 60 min was discarded after every seizure onset in order to minimize data contamination by *ictal* and *postictal* activity. As a consequence, if a seizure occurred between 75 min and 125 min after a preceding seizure, the number of *preictal* segments belonging to the second seizure was reduced.

Matthias Eberlein, Technische Universität Dresden, Faculty of Electrical and Computer Engineering, Institute of Circuits and Systems, 01062 Dresden, Germany,
E-mail: matthias.eberlein@tu-dresden.de
Jens Müller, Hongliu Yang, Simon Walz, Janina Schreiber, Ronald Tetzlaff, Technische Universität Dresden, Faculty of Electrical and Computer Engineering, Institute of Circuits and Systems, 01062 Dresden, Germany
Susanne Creutz, Ortrud Uckermann, Georg Leonhardt, Technische Universität Dresden, Neurosurgery of University Hospital Carl Gustav Carus, Dresden, Germany

Tab. 1: Patient specific characteristics for raw data (left) and segmented, labeled and split data (right). Original amount of channels and seizures are depicted in brackets. Seizures and channels have been discarded as described in Section 2.1. PI denotes the amount of *preictal* and II the amount of *interictal* segments.

Patient	Sampling rate	Channels	Seizures	Recorded duration	Seizures train / test	PI segments train / test	II segments train / test
1	1000 Hz	77 (85)	13 (14)	160.6 h	6 / 7	36 / 25	203 / 351
2	1000 Hz	58 (67)	3 (3)	229.8 h	2 / 1	12 / 6	760 / 424
3	500 Hz	107 (123)	7 (7)	256.9 h	5 / 2	28 / 12	416 / 828
4	500 Hz	80 (95)	27 (34)	203.6 h	18 / 9	74 / 27	236 / 206
5	500 Hz	62 (70)	3 (3)	139.2 h	2 / 1	7 / 6	272 / 408

In case of the occurrence of two seizures within 75 min, the whole *preictal* epoch of the second seizure was discarded. As in [9], *interictal* periods were defined to have a minimum gap of 240 min to the next seizure onset. Finally, the data of every patient was divided up into a training and a temporally separated test set, respectively. Details about the patients' and data characteristics are shown in Table 1.

2.2 Data Analysis

In this contribution, we demonstrate two different approaches for the identification of precursors of epileptic seizures in icEEG data sets. The first is based on the extraction of a suitable set of features from the time series followed by a subsequent classification. Here, the selection of suitable features is done individually for each patient. The second approach intends to overcome the strict separation of feature extraction and classification by feeding the raw time series directly to a deep neural network.

2.2.1 Feature-based Classification

We computed various univariate features including band power spectrum, statistical moments as well as error and coefficients of autoregressive models. Furthermore, we included the following bivariate features: Correlation matrices in time and frequency domain, linear coherence, Granger causality, mean phase coherence, and nonlinear interdependence [2, 9, 11, 14]. Every feature was extracted for all channels. We investigated all possible combinations of one univariate and one bivariate feature. This results in $N \times M = 437$ combinations, where $N = 23$ is the number of univariate and $M = 19$ is the number of bivariate features. Depending on the specific combination, the dimensionality of feature space varies from 348 to 7276.

For classification we utilized a Multilayer Perceptron (MLP) comprising three hidden layers with 16, 8 and 4 neurons respectively. Each hidden layer is followed by a batch normalization layer. Additionally, we used the ReLU activation function for each hidden layer while the output layer utilizes a sigmoid function. For each feature combination, we trained 100 networks with different initial weights for each individual patient.

2.2.2 Convolutional Neural Network

We recently proposed three Convolutional Neural Networks (CNN) topologies for seizure forecasting [3], showing that these networks produce promising results on different patients for several long-term data sets. With the *nv1x16*-topology, arbitrary electrode placements in the implantation scheme can be processed. Therefore, we chose this topology for our comparison in this study. The topology comprises subsequent convolution and pooling steps on single channels for feature extraction and two fully connected layers for classification. Details about the topology and training of the network are given in [3].

As the number of channels in this new data set is higher and the recorded duration is considerably shorter, we had to make adjustments to prevent overfitting. We considered two approaches, first a decrease of the number of training epochs from 50 to 10 (subsequently denoted as *10 epochs*) and second, an increased dropout-rate in the fully connected layer from 0.5 to 0.9 (denoted as *0.9 dropout*). For both approaches, 20 neural networks have been trained from scratch for each individual patient.

3 Results

The predictions of both approaches for the test set of Patient 3 are shown in Figure 1.

It can be seen, that the usefulness in a clinical application is limited since both classifiers predominantly predict seizures at a certain time of day. This can be easily explained, because

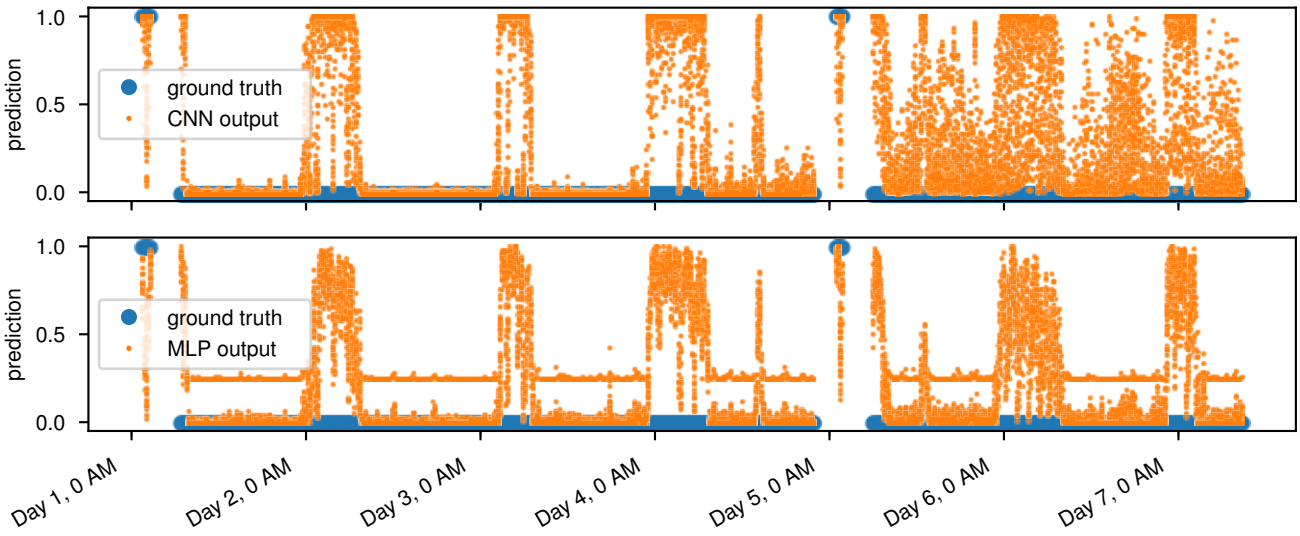

Fig. 1: Comparison of the outputs of 20 CNN (trained for 10 epochs) and 100 MLP (best features) for the test set of Patient 3. The best feature combination was chosen retrospectively based on the mean ROC AUC of the test set. The output during *interictal* and *preictal* periods and the ground truth is plotted over time. 4 out of 100 MLPs did not converge during training to an output of 0 for *interictal* but instead around 0.2. This is visible as a horizontal line in the lower plot.

in the training set of Patient 3, all five seizures occurred between 4:26 AM and 9:29 AM. This is not unusual: There is evidence that for some patients the circadian cycle has influence on the probability of seizure occurrence [5]. Some studies show that the time of day can even be used as a feature for seizure prediction [7].

Additionally, statistical metrics that were used in two recent kaggle competitions [2, 10] are shown in Table 2 for all five patients. For the CNN approaches (left) and for the feature-based approach (right) the arithmetic mean and standard deviation of the Area under Curve of the Receiver Operating Characteristics (ROC AUC) on the respective test set is given. The statistics were computed on the 20 individual runs for both CNN topologies and on all runs of all feature combinations (i.e. 437×100 runs) for the feature-based method. The best feature combination was chosen retrospectively for each patient individually, based on the mean ROC AUC of the test set. We could not identify combinations of features that perform optimally for all patients.

As can be seen in Table 2, for three out of five patients, both approaches perform statistically much better than a random predictor (Patient 1, 3 and 5, group A). For the other two patients, ROC AUC values clearly below 0.5 indicate that the classifier does not behave like a random predictor but instead is predicting the opposed class with above-chance probability (group B). Obviously, both approaches are sensitive to instationarities that correlate with the preictal state in the training data. In group A, these instationarities correlate also with the preictal state in the test data. This is not the case for group B, where both methods learned attributes that obviously correlate with the opposed class in the test set.

Tab. 2: ROC AUC of the compared methods for the test sets of each patient (P)

P	CNN: 10 Epochs	CNN: 0.9 Dropout	features: all	features: best
1	0.74 ± 0.04	0.70 ± 0.08	0.57 ± 0.18	0.79 ± 0.08
2	0.25 ± 0.08	0.29 ± 0.07	0.16 ± 0.11	0.46 ± 0.18
3	0.87 ± 0.02	0.85 ± 0.01	0.88 ± 0.04	0.91 ± 0.01
4	0.32 ± 0.06	0.21 ± 0.05	0.48 ± 0.10	0.65 ± 0.07
5	0.82 ± 0.10	0.91 ± 0.05	0.76 ± 0.13	0.96 ± 0.04

By comparing the statistics of the CNN and the average of all features under consideration, it is evident that a prediction based on a randomly chosen feature set yields a performance that is comparable to the CNN. By reporting results of the best performing features we want to demonstrate how results can be improved by better adapting the algorithms to the individual patient. However, after choosing the optimal feature set based on the performance on the test set, this evaluation is no longer out of sample and does not necessarily correspond to the performance in a clinical setting. To obtain out of sample evaluation for adapted algorithms, methods have to be optimized on a validation set sampled from the training set. Choosing the features that performed best on a validation set did not improve results compared to randomly chosen features. This is due to the fact that there is no correlation between the ROC

AUC values of the validation and test sets. Possible reasons are the instationary nature of the signals and the short length of recording time.

It is remarkable that the insufficient performance for Patient 3 that is observable in Figure 1 is not revealed by a mere ROC analysis. There are other statistical metrics like precision or specificity that are more sensitive to the high false positive rate. Nonetheless the observation that for this patient, both approaches seem to be more sensitive to the circadian cycle instead of impending seizures will not be revealed without including further clinical information about the data set.

4 Discussion

We want to point out several methodological challenges that seizure prediction research is currently facing. First, the main condition for scientifically valuable results is a true out of sample evaluation with a temporally separated test set [8]. Strictly speaking, the test set must only be used once for final evaluation of the model. Iterative testing and tuning of the algorithm causes data leaks from the test set. Besides that, statistical evaluation has to be reconsidered. We are aware that pseudo-prospective evaluations – as shown in Figure 1 – are not always feasible, depending on the recorded duration of the data set and the available meta data. Nonetheless, results should always be validated by comparing them with naive classifiers ([6, 15]), by testing with surrogate data [1] or by providing at least a more extensive statistical analysis of the findings – especially if new data sets are being investigated.

None of these recommendations are new, but still have not found their way into a widespread best-practice methodology in seizure prediction. Our aim is to create awareness to them and emphasize the consequences of methodological flaws in study design.

5 Conclusion

Although the results shown here are not statistically representative due to the small amount of seizures used for the evaluation, we demonstrate that feature-based methods and deep learning algorithms can both be sensitive to hidden properties of the data and highlight the necessity of proper evaluation of methods to improve the meaningfulness of future studies.

Acknowledgement: This work was supported by the European Regional Development Fund (ERDF) and the Free State of Saxony (project number: 100320557).

We thank the Center for Information Services and High Performance Computing (ZIH) at TU Dresden for generous allocation of computing time.

References

[1] Ralph G. Andrzejak et al. “Testing the null hypothesis of the nonexistence of a preseizure state”. In: *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 67.1 (2003), p. 4.

[2] Benjamin H. Brinkmann et al. “Crowdsourcing reproducible seizure forecasting in human and canine epilepsy”. In: *Brain* 139.6 (2016), pp. 1713–1722.

[3] Matthias Eberlein et al. “Convolutional Neural Networks for Epileptic Seizure Prediction”. In: *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*. IEEE, 2019, pp. 2577–2582.

[4] Dean R Freestone, Philippa J Karoly and Mark J Cook. “A forward-looking review of seizure prediction”. In: *Current Opinion in Neurology* 30.2 (2017), pp. 167–173.

[5] Philippa J. Karoly et al. “Interictal spikes and epileptic seizures: their relationship and underlying rhythmicity”. In: *Brain* 139.4 (2016), pp. 1066–1078.

[6] Philippa J Karoly et al. “The circadian profile of epilepsy improves seizure forecasting”. In: *Brain* 140.8 (2017), pp. 2169–2182.

[7] Isabell Kiral-Kornek et al. “Epileptic Seizure Prediction Using Big Data and Deep Learning: Toward a Mobile System”. In: *EBioMedicine* 27 (2018), pp. 103–111.

[8] Iryna Korshunova et al. “Towards Improved Design and Evaluation of Epileptic Seizure Predictors”. In: *IEEE Transactions on Biomedical Engineering* (2018).

[9] Levin Kuhlmann et al. “Epilepsyecosystem.org: crowdsourcing reproducible seizure prediction with long-term human intracranial EEG”. In: *Brain* 141.9 (2018), pp. 2619–2630.

[10] Levin Kuhlmann et al. “Seizure prediction — ready for a new era”. In: *Nature Reviews Neurology* 14.10 (2018), pp. 618–630.

[11] F. Mormann et al. “Seizure prediction: the long and winding road”. In: *Brain* 130.2 (2007), pp. 314–333.

[12] Vivek Nagaraj et al. “Future of seizure prediction and intervention: closing the loop.” In: *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society* 32.3 (2015), pp. 194–206.

[13] Andreas Schulze-Bonhage and Anne Kühn. “Unpredictability of Seizures and the Burden of Epilepsy”. In: *Seizure Prediction in Epilepsy*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2008, pp. 1–10.

[14] R. Tetzlaff and V. Senger. “The Seizure Prediction Problem in Epilepsy: Cellular Nonlinear Networks”. In: *IEEE Circuits and Systems Magazine* 12.4 (2012), pp. 8–20.

[15] M. Winterhalder et al. “The seizure prediction characteristics: A general framework to assess and compare seizure prediction methods”. In: *Epilepsy and Behavior* 4.3 (2003), pp. 318–325. arXiv: 1311.7129v1.