Towards improved design and evaluation of epileptic seizure predictors

Iryna Korshunova, Pieter-Jan Kindermans, Jonas Degrave, Thibault Verhoeven, Benjamin H. Brinkmann *Member, IEEE,* Joni Dambre

Abstract*—*Objective: **Key issues in the epilepsy seizure prediction research are (1) the reproducibility of results (2) the inability to compare multiple approaches directly. To overcome these problems, the Seizure Prediction Challenge was organized on Kaggle.com. It aimed at establishing benchmarks on a dataset with predefined train, validation and test sets. Our main objective is to analyse the competition format, and to propose improvements, which would facilitate a better comparison of algorithms. The second objective is to present a novel deep learning approach to seizure prediction and compare it to other commonly used methods using patient centered metrics.** ***Methods:*** **We used the competition's datasets to illustrate the effects of data contamination. Having better data partitions, we compared three types of models in terms of different objectives.** ***Results:*** **We found that correct selection of test samples is crucial when evaluating the performance of seizure forecasting models. Moreover, we showed that models, which achieve state-of-the-art performance with respect to commonly used AUC, sensitivity and specificity metrics, may not yet be suitable for practical usage because of low precision scores.** ***Conclusion:*** **Correlation between validation and test datasets used in the competition limited its scientific value.** ***Significance:*** **Our findings provide guidelines which allow for a more objective evaluation of seizure prediction models.**

***Index Terms*—Epilepsy, neural networks, support vector machines, linear discriminant analysis.**

I. Introduction

Epilepsy is one of the most common neurological disorders, as it affects nearly 1% of the world population. It is characterized by the occurrence of spontaneous seizures. In about 30% of cases, medication is not effective in preventing the seizures [1]. The remaining 70% have to take anti-epileptic drugs daily over a period of years to control the seizure frequency. Unwanted side effects of the medication as well as seizure-related injuries and anxiety due to an expectation of seizures, significantly lowers quality of life for the patients [2]. Seizure forecasting systems aim to improve the wellbeing of patients with epilepsy. A system capable of predicting periods with increased risk of a seizure would allow them to avoid potentially dangerous activities at such times. In addition, they could take medication only when needed, significantly reducing the amount of administered drugs and their concomitant side effects. These systems need highly reliable algorithms to detect periods of increased probability of an oncoming seizure.

Seizure prediction can be formulated as a binary classification problem between preictal and non-preictal classes. The preictal state is the period preceding a seizure onset. Across previous studies, its duration varies from a couple of minutes to several hours [3]. A non-preictal class can denote one of three states: interictal (normal), ictal (seizure) and postictal (after seizure) [4]. The main challenge of seizure prediction is to classify between preictal and interictal states.

For a long time, the mere possibility of predicting epileptic seizures was controversial. This was mainly due to statistical flaws [3], caused by the use of short and selected EEG recordings, which made a proper evaluation close to real clinical conditions impossible. The need for subject-specific models is the major reason for working with recordings of a very limited duration: long-term EEG monitoring can often be inconvenient for patients, expensive and time consuming to label.

Recently, several datasets of long-term intracranial EEG (iEEG) recordings from canine subjects have become available and found their usage in a few studies [5], [6]. The reason behind exploiting canine data is that epilepsy in dogs and humans is proven to be highly similar [7] and it is easier to obtain relatively long iEEG recordings from canines.

A new dataset containing over 26,000 hours of iEEG from eight dogs is publicly available on the International Epilepsy EEG portal [1]. A subset of this dataset extended with recordings from two human patients, was used in the American Epilepsy Society Seizure Prediction Challenge organized in August, 2014 on Kaggle.com [2]. Brinkmann et al. [8] describes in detail the setup of the contest, the dataset and a high-level overview of the top 10 algorithms, including our solution.

This contest was supposed to be an important step towards a reproducible seizure prediction research. It aimed at establishing a dataset with predefined train, validation, and test sets so that different algorithms can be directly compared. Ideally, this would bring significant advances to the field, similarly to what contests like ILSVRC [9] and COCO [10] are doing for computer vision. Unfortunately, the results of the Seizure

Manuscript received February 2017; revised April 2017.
This work was supported in part by the Special Research Fund of Ghent University, the Agency for Innovation by Science and Technology in Flanders, and European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement NO 657679.

I. Korshunova, J.Degrave, T. Verhoeven, J.Dambre are with Ghent University, Technologiepark-Zwijnaarde 15, 9052 Ghent, Belgium (correspondence e-mail: iryna.korshunova@ugent.be)
P. J. Kindermans is with Machine Learning group, Technische Universität Berlin, Marchstr. 23, D-10587 Berlin, Germany
B. H. Brinkmann is with Mayo Systems Electrophysiology Laboratory, Departments of Neurology and Biomedical Engineering, Mayo Clinic, Rochester, MN 55905, USA

[1] ieeg.org
[2] www.kaggle.com/c/seizure-prediction

Prediction Challenge cannot be used as a benchmark for future studies and in this paper we analyze why this is the case.

Despite some design flaws in the Seizure Prediction Challenge, its results are still valuable. The contestants used a variety of machine learning models applied to a wide range of features from the time and frequency domains. The general approach, however, was much the same as in many previous studies, confirming that spectral power in discrete frequency bands is a valuable feature for seizure forecasting [5], [6], [11]. Moreover, SVM [12] was the most commonly used algorithm, which follows the trends in the seizure prediction research community [13]. While the competition was limited to using one measure to evaluate the models, in this work we carry out a more in-depth analysis to gain insights into the real-world performance of the proposed algorithms.

Finally, we describe our ninth place solution to the Seizure Prediction Challenge, which is based on convolutional neural networks(CNNs) [14]. While our model was an ensemble of multiple CNNs, here we describe a single refined network, which has comparable competition scores and also suits better for classification of long EEG segments. We show that the CNN-based approach is promising since its predictions are different from those generated by commonly used techniques such as SVM and linear discriminant analysis. Comparison of these three methods showed that no approach is able to outperform others on all the subjects and selected clinically relevant evaluation metrics. Achieving the highest score in the competition with any of the models was not the goal of this paper since doing so would require us to utilize strategies not applicable in practice. For example, top finishing contestants used a non-causal rescaling of predictions, i.e. information from the future was used to correct predictions from the past [8], [15]. This greatly altered the scores on the competition leaderboard and therefore we believe that it is more important to follow correct methodology rather than chasing the leaderboard numbers.

II. Methods

A. Data

In this work we used the dataset provided for the Seizure Prediction Challenge. It consists of iEEG recordings from five dogs with naturally occurring epilepsy and two humans undergoing presurgical wide bandwidth iEEG monitoring for drug-resistant epilepsy. iEEG recordings were collected using a NeuroVista seizure advisory system [16]. For the canine subjects, sixteen subdural electrodes were implanted; the signal was sampled at 400 Hz. The electrodes were placed on two bilateral pairs of 4-contact strips with recorded iEEG voltages referenced to the group average. For human patients, the configuration and number of electrodes were dictated by clinical conditions; the signal was sampled at 5000 Hz, and the voltages were referenced to an electrode outside the brain [8].

Seizures are known to occur in groups, therefore little benefit can be gained by forecasting the follow-on seizures [17]. For this reason, competition organizers included only lead seizures into the dataset. These were defined as seizures occurring at least 4 hours after the previous seizure. Preictal data was extracted from a 66 minute period preceding the lead seizure as shown on Fig. 1. Each one hour sequence was divided into 10 minute clips spaced 10 seconds apart and allowing a 5 minute horizon before the marked seizure onset. Similarly, interictal clips were grouped in sequences of six. Interictal periods were randomly chosen from the whole record such that at least one week is present before or after any seizure. Additionally, for each 10 minute training clip, its relative position within the 1 hour period was known.

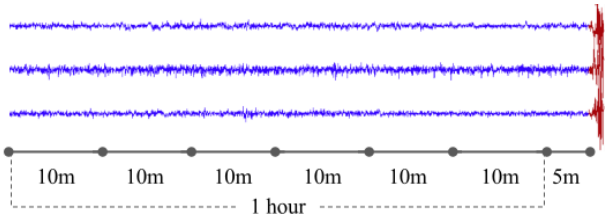


Fig. 1: An example of a preictal EEG sequence consisting of six 10 minute clips. For convenience, we plotted only 3 channels and ommited the 10 seconds gaps between clips.

Competitions on Kaggle.com usually follow the convention of splitting the dataset into three parts: train, public and private. Participants receive no class labels for the public and private sets, so these sets are used for model evaluation, which is done as follows. The Kaggle platform computes two scores based on a set of model predictions: a public and a private score. Public scores are immediately revealed during the competition for every submitted model. Private scores, on the other hand, are available only after the competition ends, so they determine the final ranking. Public scores serve as a source of validation since they can be used to adjust model hyperparameters. Private scores assess the test performance, i.e. model's ability to generalize to out-of-sample data.

To fit the format of the competition, the full EEG record from each subject was partitioned into two parts of approximately equal size: training clips were taken from the first half, and the second half was used to create public and private sets. In the interest of providing the most useful data for training and testing while minimizing the total size of the EEG data bundle, the preictal data periods were sampled more heavily than interictal periods in comparison to the original iEEG recording. As a result, there are more preictal clips than would be present with true random sampling. Clips for the public and private sets were sampled at random, which means that clips from the same 1 hour sequence can end up in different sets.

For the post-competition analysis, organizers prepared a hold-out data, which was used to evaluate the top 10 finishing models [8]. Fig. 2 illustrates the split between train, public, private and hold-out sets. Table I provides the subject-specific characteristics of the iEEG data and the details on each part of the dataset.

B. Task specification and evaluation measures

The goal of the challenge was to classify 10 minute clips of EEG activity. For each clip in the test set, models were required to output a preictal probability, i.e. the probability of a given clip being preictal.

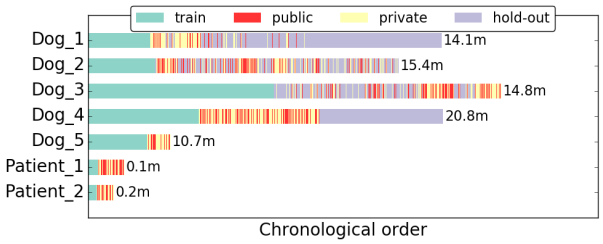


Fig. 2: Partitioning of the record into train, public, private and hold-out sets (time gaps between clips omitted for clear visualisation). For each subject it indicates number of months (m) spent in the recording phase, while the number of clips selected from each such recording is given in Table I.

TABLE I: Per-subject characteristics of the dataset: number of recorded electrodes, number of 10 minute clips in each set: total amount of clips with number of preictal clips given in braces. The number of interictal clips is implied.

Subject	EEG channels	Training	Public	Private	Hold-out
Dog 1	16	504 (24)	176 (11)	326 (13)	2000 (6)
Dog 2	16	542 (42)	404 (42)	596 (48)	1000 (0)
Dog 3	16	1512 (72)	392 (22)	515 (20)	1000 (0)
Dog 4	16	901 (97)	401 (25)	589 (32)	1000 (42)
Dog 5	15	490 (40)	74 (5)	117 (7)	
Patient 1	15	68 (18)	75 (3)	120 (9)	
Patient 2	24	60 (18)	52 (7)	98 (7)	

A key limitation in the competition was that the data processing and classification algorithm had to be identical for all subjects, but the hyperparameters were allowed to vary depending on the data properties, such as sampling frequency.

Before explaining the competition scoring function, we will briefly describe two measures of a binary classification performance. We define 'preictal' as being positive class and 'interictal' to be negative. In this case, sensitivity or a true positive rate (TPR) is a proportion of correctly classified preictal examples: $TPR = \frac{TP}{TP+FN}$, where TP is a number of true positive prediction and FN is a number of false negatives. By analogy, specificity or a true negative rate (TNR) can be calculated as $TNR = \frac{TN}{TN+FP}$.

In the competition, the submissions were judged based on the area under the receiver operating characteristic curve (ROC AUC) which is a curve that plots TPR against $1 - TNR$ at different values of a discrimination threshold. The latter transforms classifiers continuous predictions into binary labels: if the threshold is exceeded, the clip is labelled as preictal and interictal otherwise.

In the competition, AUC was computed over all predictions of all the subjects at once, i.e. if vector $\mathbf{p}_i$ contains predictions for the ith subject, then model's score is $AUC(\bigcup_i \mathbf{p}_i)$, where $\bigcup_i$ denotes concatenation. We will further refer to this measure as aggregated AUC.

When using aggregated AUC, well-calibrated predictions, i.e., predictions that respond similarly to threshold changes, result in better scores. Misalignment of probabilities between subjects can drastically worsen the performance. To explain why this is the case, we will use an alternative interpretation of AUC, which is a probability that a randomly chosen positive (preictal) instance has a higher rank than a randomly chosen negative (interictal) one [18]. For the moment, assume that one subject has the following pairs of predictions and class labels: (0.1;0), (0.2;0), (0.3;1), (0.4;1), its AUC equals to 1, since all positive examples are ranked higher than negatives. Similarly, for the second subject with (0.6;0), (0.7;0), (0.8;1), (0.9;1). However, if we combine these prediction-labels in one group, there are 16 ways to sample pairs of positive and negative examples, and only 12 of these pairs have a higher predicted probability for the positive instance than for the negative one. This yields an aggregated AUC of 0.75. To conclude, the aggregated AUC is higher for the per-subject models which produce comparable probabilities, thus it requires a certain level of robustness against variations of the discrimination threshold.

To soften the requirements of the aggregated AUC, competition rules allowed to rescale test set predictions using the information about the distribution of predictions in the public and private sets. For instance, the simplest solution would be to rescale predictions such that minimum and maximum per-subject test set probabilities are between 0 and 1. In the above-mentioned example, this gives prediction-label pairs of (0.0;0), (0.33;0), (0.67;1) and (1.0;1) for both subjects, and the aggregated AUC becomes 1 again.

In many works, seizure prediction models are trained to classify segments of a fixed length [6], [19]–[21]. In this case, once the preictal probability exceeds a certain threshold, the seizure forecasting system triggers an alarm, and the warning state persists for the same duration as the length of the classified segment. The seizure is considered as predicted if it occurs while the alarm is on [6]. From a machine learning perspective, this formulation makes the problem very well defined.

While having many short-term predictions, quantifying the event of a missed or a forecasted seizure is not always straightforward: to create an event, the subsequent decisions have to be aggregated. It was also noticed in previous studies [21] that there is a difference between the problem formalization seen from clinical and machine learning perspectives. While the majority of algorithms are limited to the classification of short EEG clips, a clinically relevant objective is to correctly classify segments of about 1 hour long [21]. Once we have predictions for 1 hour sequences, event-based metrics can be applied directly. There are two commonly used metrics: lead sensitivity and false positives per hour [6], [22]. Since our dataset already contains only leading seizures and the classification is done per one hour segments, these metrics reduce to traditional sensitivity and specificity. Precision or a positive predictive value ($PPV = \frac{TP}{TP+FP}$) is another binary metric that is valuable for the patients, since it gives a probability of having a seizure, when the system raises the alarm. Unfortunately, studies rarely report PPV [22]. By analogy to PPV, a negative predictive value can be defined. However, due to a large skewness towards the interictal (negative) samples in the distribution of class labels, it is usually very high (greater than 98% for all the considered models), and thus we will not analyse NPV scores.

C. Preprocessing and feature extraction

Simple features from the frequency domain have been shown to be discriminative between preictal and interictal states [5], [11]. We further expanded upon these ideas, so firstly, the 10 minute clips were resampled to 400 Hz and filtered with a band-pass filter between 0.1 and 180 Hz. Each clip was further partitioned into 10 nonoverlapping 1 minute frames which were Fourier transformed. Within each frame, we took the logarithm of the amplitude spectrum. This was averaged within the following frequency bands: delta (0.1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), low-gamma (30-50 and 50-70 Hz), high-gamma (70-100 and 100-180 Hz). The resulting dimension of a data clip was thus equal to $N \times 8 \times 10$ (number of iEEG channels $\times$ frequency bands $\times$ time frames). The training data was standardized by using means and standard deviations calculated over a set of 1 minute frames. These values obtained from the training set were used to standardize test data clips prior to their classification.

D. SVM and LDA classifiers

Many studies reformulate the problem of seizure prediction as a classification of short moving windows from tens of seconds to minutes [4], [6], [11], [21]. This reduces the dimensionality of the feature vectors and increases the number of training examples, thus making many classification algorithms feasible to train. However, if the ultimate goal is to discriminate between longer EEG segments, this approach requires post-processing of the classifier outputs at test time.

In the Seizure Prediction Challenge, this strategy was used in all the best scoring models [8]. For this study, we implemented it by training linear discriminant analysis (LDA) and SVM models on 1 minute clips of $N \times 8$ frequency features. For SVM we used a radial basis function (RBF) kernel with cost and scale parameters C=10 and gamma=0.01 selected on the public set.

To make a single prediction for each 10 minute clip, we took an arithmetic mean of 1 minute predictions, which worked sufficiently well without further post-processing.

E. Convolutional neural networks approach

A convolutional neural network (CNN) consists of a stack of layers, each of which processes the output from the layer below, and passes its output to the layer above. This way, each layer builds a more abstract representation than the layer below. The bottom of such networks is usually composed of convolutional layers, which are sparsly connected, and thus, process only local information. The idea of using sparse connectivity is based on the fact that spatially or temporally nearby features are likely to have mutual information, which is important for the network to grasp in many kinds of image- and signal-processing applications. Many of the current state-of-the-art algorithms in computer vision, artificial intelligence, speech processing rely on using CNNs [23]–[25]. That is why we reckon it can be useful for EEG classification tasks, since in essence, it is a time series classification problem.

As a motivation for our CNN architecture, we used several hypotheses. Since the entire 66 minute period defined as preictal cannot perfectly match the prototypical physiological preictal signature, a preictal sequence of six clips is likely to contain clips without any traces of the epileptic activity. Similarly, we can have both noninformative and informative fragments of EEG at any given time scale within each clip. In order to find relevant regions within the signal, our model needs to extract features in short time windows. Eventually, it has to combine information from different frames to make a single prediction for each 10 minute clip.

The convolutional layers, who first process the signal, are implemented with one-dimensional convolutions through time. The purpose of these layers is to extract the same set of features on every time step. A layer in the neural network, which implements such convolution, gets as input a stack of feature maps $\mathbf{X}^k$ with $k = 1...K$, and convolves each feature map with a set of learnable filters $\mathbf{W}^{k,l}$ to produce a stack of 1-dimensional output feature maps $\mathbf{Y}^l$ with $l = 1...L$:

$$\mathbf{Y}^l = f(\sum_{k=1}^{K} \mathbf{W}^{k,l} * \mathbf{X}^k + b^l) \quad (1)$$

Here, $*$ symbol denotes a one-dimensional convolution, applied along the time axis. f is a nonlinear activation function and b^l is a bias per output feature map l. Multiple of these convolutional layers can be stacked, and can thus form a wide array of non-linear convolutional filters which extract local features from their input.

After convolutional layers have preprocessed the signal, their local features were aggregated inside the network using a global temporal pooling layer. It eliminates temporal information by computing basic statistics over each feature map from the last convolutional layer. It is justified by the fact that exact timing of a preictal symptom within a clip is irrelevant: we are only interested in whether it is present or not.

Multiple studies supported the use of bivariate features, which capture the relationships between pairs of EEG channels [6], [26]. However, instead of using bivariate features, we first made layer convolutional filters to see all frequency bands from all the channels at once. This enables the network to learn the relevant correlations between frequency features from different channels by itself.

Fig. 3 schematically illustrates the CNN architecture, which implements the ideas above. It is very similar to the networks we used in our top ten entry for the Kaggle Seizure Prediction Challenge, except for the presence of intermediate softmax readouts, whose role we explain further.

We found no significant differences in scores, when comparing individual models from our Kaggle ensemble [15] and the model in Fig. 3, when trained on the task of 10 minute clips classification. However, our former models fail to achieve any reasonable performance when trained on 1 hour sequences. The main reason for this, is that there are six times less training samples when 10 minute clips are grouped into 1 hour clips. To cope with this problem, extra regularisation of our model is required, such that the increased risk of overfitting on the fewer but bigger input samples is mitigated. One way of doing this is

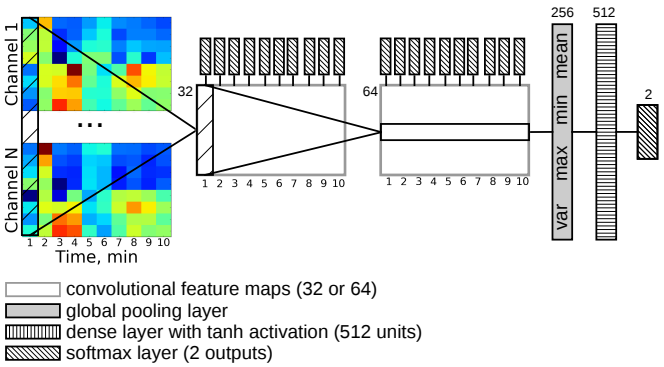


Fig. 3: The architecture of the proposed CNN model. The schematic uses the notation of Krizhevsky et al. [27]. The number of convolutional feature maps and number of units in dense layers are indicated above each layer. The global pooling layer computes the following statistics over 10 activations within each input feature map: variance (var), maximum (max), minimum (min), and mean value.

by adding more prior knowledge of the problem at hand into the network architecture. We did this by including additional softmax outputs at each timestep inside the convolutional layers, as shown on Fig. 3. During training, each of these logistic regression outputs needs to classify the frame correctly as well. Backpropagation through these new outputs forces the features from each time frame to be discriminative for the final classification of that frame. This prevents the lower layers from relaying too much information to the later stages in the model, which the model could use for recognizing specific examples in the train set and thus overfit. During the evaluation phase, these additional readouts are omitted, and the final prediction is only produced by the final softmax layer.

The network presented in Fig. 3 has two convolutional layers with 32 and 64 feature maps respectively, followed by a global pooling layer, a dense layer with a hyperbolic tangent activation and finally a dense layer with a softmax activation. We used ReLU [28] activation functions in the convolutional layers. The model is trained to minimize the cross-entropy loss between labels and predictions. To optimize this loss function, we used the Adam algorithm [29] with a mini-batch size of 32 for 5000 epochs with a learning rate of 0.03. As a form of regularisation, we used L2 regularisation applied to all network weights and dropout [30] applied to the inputs of dense layers.

III. Results

A. Aggregated AUC analysis

In Table II we provided aggregated AUC scores of the CNN, SVM and LDA models evaluated on the hold-out data and the parts of the dataset available in the Kaggle Seizure Prediction Challenge. For comparison, we included the results from the post-competition study, which evaluated the solutions of top scoring teams, including our CNN ensemble [8]. Further, we will not consider ensembles of models since they are difficult to analyse and extra complexity makes them less desirable in practical applications.

TABLE II: AUC leaderboard scores on public, private and hold-out sets. The first column gives the results from a post-competition study [8]: an average AUC over the six top-scoring algorithms together with their minimum and maximum values. The second column contains the score of our CNN ensemble, which was a top-10 finisher in the Kaggle Seizure Prediction Challenge. The scores of our improved CNN architecture, SVM and LDA models are given in the next three columns. The last two colums provide scores of SVM and LDA models, whose per-subject predictions were calibrated.

	Top Kaggle scores min-max	Kaggle CNN ensemble	CNN	SVM	LDA	SVM calibrated	LDA calibrated
Public	0.81-0.86	0.81	0.78	0.76	0.76	0.80	0.81
Private	0.78-0.82	0.78	0.76	0.74	0.75	0.80	0.80
Hold-out	0.59-0.79	0.77	0.79	0.84	0.59	0.66	0.55

Table II also shows that competition scores on the private set, which was used to determine the winners, are not achievable using original predictions from CNN, SVM and LDA models without combining classifiers or advanced post-processing. However, when their per-subject predictions are calibrated using non-causual rescaling schemes, the scores match those from the top three winners, who also used such a rescaling. The predictions for calibrated SVM and LDA in Table II are calculated by applying logistic function to per-subject predictions $\mathbf{p}_i$, standardized using means μ_i and standard deviations σ_i calculated over all subject's predictions in the coresponding set (public, private or hold-out):

$$\mathbf{p_i}^{\text{calibrated}} = \frac{1}{1 + exp(-\frac{\mathbf{p}_i - \mu_i}{\sigma_i})}. \quad (2)$$

Note that using statistics of the train predictions to calibrate public, private and hold-out predictions does not improve the aggregated AUC scores. Similarly, for CNN the calibration we applied to LDA and SVM predictions, has no apparent effects.

As was shown, non-causal rescaling of SVM's and LDA's outputs yields an improvement of almost 10% in aggregated AUC compared to the original scores. We can, therefore, hypothesize that original per-subject predictions are miscalibrated. This can be verified by analyzing optimal discrimination thresholds for per-subject predictions. The simplest criterion for the threshold to be optimal is that it achieves a minimum Euclidean distance between the ROC curve and the (0, 1) point of the ROC space, which corresponds to zero false positive rate and maximal sensitivity. For LDA, in Fig. 4 we can see that on a joint private and public set it has different optimal thresholds for each subject, while the calibration evens them out.

B. Per-subject AUC analysis

In the previous section, we have shown that aggregated AUC can be altered using non-causual rescaling schemes of per-subject probabilities. While this drastically changed the leaderboard scores in the Kaggle Seizure Prediction competition, strictly monotonically increasing transformations, like the one we used, have no effect on per-subject AUCs.

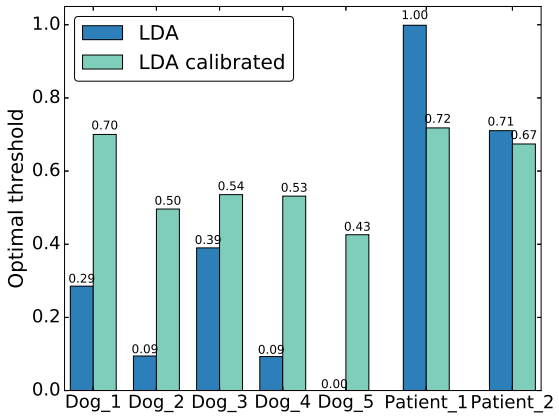


Fig. 4: Optimal per-subject thresholds for private and public predictions of LDA with and without calibration.

In this section, we will analyze per-subject AUC results and see how they change when calculated on different splits of the dataset. This will demonstrate the effect of data contamination due to the fact, that competition contenstants could use the public set for validation, while the split of the EEG record into public, private and hold-out parts was made at random (see Fig. 2).

Since we cannot calculate AUC in cases, when no positive samples are available (e.g. hold-out set for Dog 2 and Dog 3), we will be gradually merging samples from different sets and comparing the results. For the purpose of our analysis, we excluded Dog 5, Patient 1 and Patient 2 from consideration since they did not have a hold-out set. Further, we will consider the following cases.

Case 1: When EEG clips are temporally interweaved, we argue that they should be considered jointly either for validation or testing purposes. In the Kaggle Seizure Prediction Challenge this was not the case with clips from the public (validation) and private (test) sets. Fig. 5 shows a split, were we merged them into one 'test' set. Per-subject AUC on this newly defined set will give us a baseline for the next comparisons.

Case 2: Fig. 5 shows that clips from the hold-out set were also mixed in time with those from public and private sets from Dog 1, Dog 2 and Dog 3. When joining all temporally mixed clips as shown on Fig. 6, we can see that AUC scores are increased compared to the previous case. This increase is the result of adding many samples, which are easier for the model to classify due to their proximity to the public clips, on which the models were validated.

Case 3: Finally, we want to check how the score changes once we take into account clips from the hold-out set, which do not interlace in time with other clips. Based on the available data, this can only be done for Dog 1 and Dog 4 as shown in Fig. 7. As expected, the performance has dropped compared to the AUC results from Fig. 5. This is a result of adding true out-of-sample clips, which are further in time from the training and validation data.

C. Event-based classification

So far, we have provided the results only in terms of AUC scores, which have a limited interpretability. In this section, we will measure the performance of seizure prediction models using clinically relevant event-based binary classification metrics. We will focus on the task of 1 hour sequences classification, which makes it easy to evaluate models in terms of correct predictions, false alarms and missed seizures once the discrimination threshold is chosen.

Our proposed CNN architecture can be trained directly on groups of six 10 minute clips, while SVM and LDA models trained on 1 minute chunks require extra processing of the outputs. Since taking a mean of SVMs or LDAs outputs worked well for 10 minute clips classification, we applied the same technique here: averaging over sixty 1 minute predictions to get a single probability for each 1 hour block of EEG-signals.

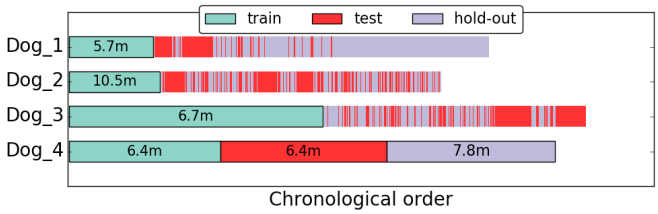
To tune the discrimination threshold, we used stratified cross-validation with four folds, as the number of folds is bounded by the number of positive examples in the train set, which is only four for Dog 1. Cross-validation was done without shuffling the chronological order of the clips. As previously, the optimality criteria for the threshold was a minimum Euclidean distance between the ROC curve and the $(0, 1)$ point of the ROC space.

For our experiment, we selected Dog 1 and Dog 4, since these subjects have a real hold-out set, which does not overlap with Kaggle's public or private sets as shown in Fig. 6 for Dog 1 and in Fig. 5 for Dog 4. In Fig. 8 we plotted precision (PPV), sensitivity (TPR) and specificity (TNR) as functions of the discrimination threshold. This graph is analogous to ROC and precision-recall curves and can display the complete performance of our models [22].

From these plots, several observations can be made. For both subjects, Dog 1 and Dog 4, CNN's performance curves are mostly flat, which means that CNN's outputs are close to binary values. Therefore, a threshold of 0.5 would suffice for every subject.

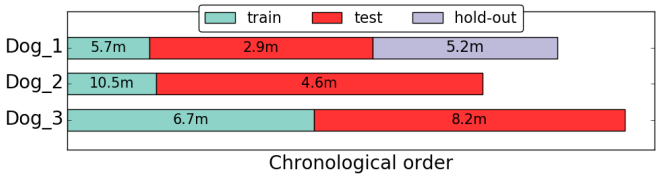
Also, these two cases show a performance decay on the hold-out set. It is logical that clips, which are further in time from the training set, would be misclassified more often due to the nonstationarity of the data. As a result, models need to be regularly retrained. Alternatively, adaptive methods [31], [32] or robust approaches [33]–[35] should be developed to cope with the covariate shift [36].

For Dog 1, Fig. 8(a) shows that LDA's and SVM's performance curves are very dissimilar despite their equality in terms of AUC scores on the test set. LDA's threshold found via cross-validation results in a very low specificity on the test set. Unlike LDA, SVM has an acceptable TNR and TPR for a given threshold, however both methods have almost trivial PPV. Our CNN approach only suffers from a low sensitivity of 25%, which translates into predicting 1 seizure out of 4 seizures from the test set. Therefore, even though these models reach the state-of-the-art performance, all three methods are probably not yet truly useful for patients in a clinical setting. However, it is worth mentioning that Dog 1 was one of the



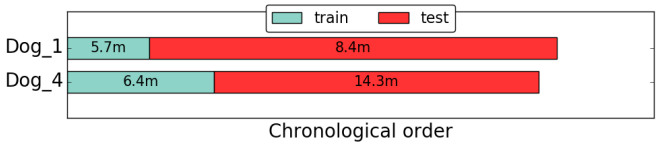
Dog	CNN	SVM	LDA
1	0.78	0.74	0.76
2	0.78	0.81	0.86
3	0.77	0.81	0.75
4	0.90	0.80	0.88

Fig. 5: First variant of repartitioning the dataset, which puts public and private clips into one set. For blocks of clips, belonging to one set, we provided the period in months (m) over which the clips were taken. A table on the right provides per-subject AUC scores of the CNN, SVM and LDA models on this newly defined test set.



Dog	CNN	SVM	LDA
1	0.80	0.76	0.76
2	0.80	0.81	0.87
3	0.80	0.88	0.79

Fig. 6: Second variant of repartitioning the dataset, which puts public, private and temporally interlaced hold-out clips into one set. A table on the right provides per-subject AUC scores of CNN, SVM and LDA models on this newly defined test set.



Dog	CNN	SVM	LDA
1	0.62	0.66	0.54
4	0.81	0.76	0.78

Fig. 7: Third variant of repartitioning the dataset, which puts public and private clips and hold-out clips into one set. A table on the right provides per-subject AUC scores of CNN, SVM and LDA models on this newly defined test set.

difficult subjects in this dataset since it had only three preictal sequences in the training set.

For Dog 4, Fig. 8(b) shows that all models have similar performance with SVM being slightly worse for TNR and PPV, which is caused by a bad choice of the threshold. LDA's threshold, on the other hand, is close to optimal for the test set. Despite a maximal sensitivity and a specificity of almost 90%, PPV is low: only a third of all the alarms predict a seizure.

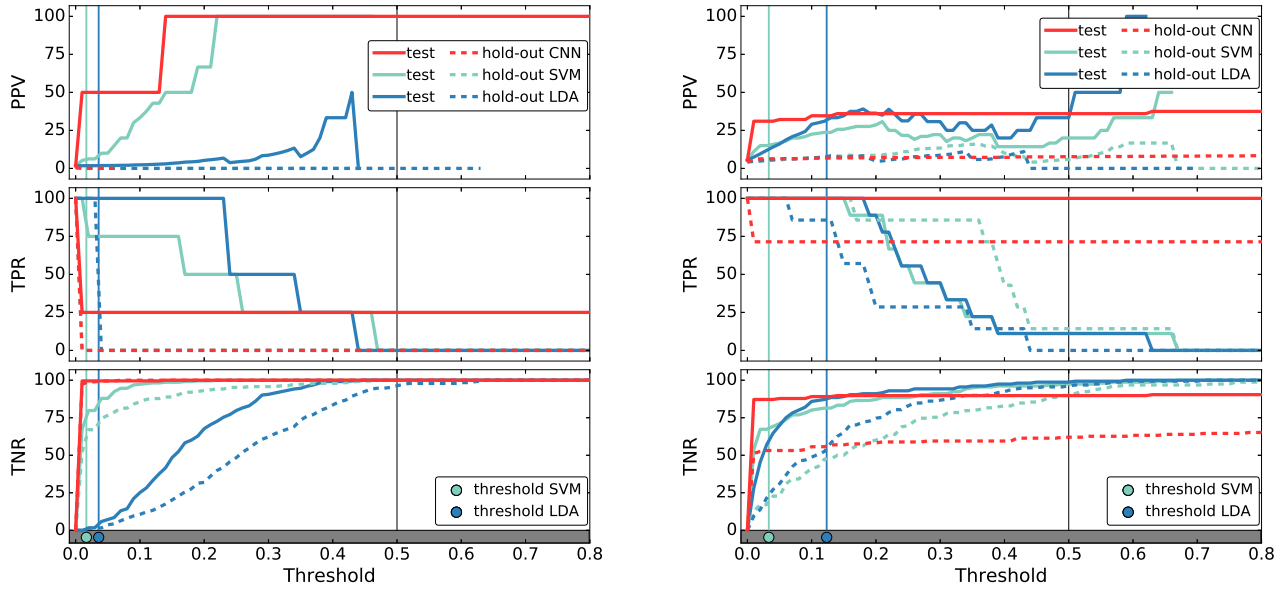
IV. Discussion

We argue that the main reason behind a relatively slow progress in the field of epileptic seizure prediction is the absence of well-established multi-objective benchmarks and publicly available datasets with well-defined train and test sets. Likely, this is the reason why many works cease to compare their approaches against the existing methods [4], [5], [11]. In this paper, we tried to fill this evaluation gap by benchmarking two common classifiers, namely LDA and SVM, against our novel CNN approach on a dataset available in the Kaggle Seizure Prediction Challenge [8].

A fair comparison of the models was complicated by a couple of choices made in the competition design. First of all, we were not able to compare our results to the results of other contestants, because many of them used various schemes to non-causually rescale per-subject test predictions. The second issue was a random split of the dataset into public and private sets. While this is a natural move for independent and identically distributed samples, in case of EEG, it makes the information leak from the validation set into the test set. Therefore, test estimates become overly optimistic as we demonstrated in section III-B when comparing Fig. 5 and Fig. 6. Splitting the continuous record into non-overlapping temporal blocks, such that test data always follows the data, used for parameter tuning, would be a better design, since it simulates the conditions in which seizure forecasting systems are used in practice.

Despite high AUC, TNR, and TPR, positive predictive value of seizure prediction models remains limited. We were able to recalculate PPV for another study, which used long-term recordings from 24 human patients [20]. The mean PPV value across patients in that study was 31%. Furthermore, we would like to stress that this is an optimistic estimate, because no distinction between leading and follow-up seizures was made. We consider PPV to be as important as sensitivity and specificity. It can give a better feeling of how much anxiety a false alarm brings to the patient. For example, with 100% TPR and 90% TNR as in case of Dog 4, there is only a chance of one out of three that a seizure will follow after an alarm. With a higher proportion of interictal data and the same rates of sensitivity and specificity, PPV will become even lower. This may lead to 'alarm fatigue' when patients start to disregard the warnings [37].

Comparing three approaches, SVM, LDA and our novel CNN method, we found that there is a high variance in their performance across test-subjects. We note that it is Dog 4 with most preictal training samples, which performs best, especially



(a) Dog 1

(b) Dog 4

Fig. 8: Precision (positive predictive value), sensitivity and specificity curves on two datasets from Dog 1 and Dog 4. Thresholds for SVM and LDA were optimized via cross-validation on the train set.

using the CNN. This observation is not surprising since CNN has more trainable parameters and therefore, it is likely for this method to benefit most when more annotated data is available. This is in line with the observations made in other fields such as computer vision [38].

Another peculiarity of the CNN model is that it can be trained directly on large chunks (∼1 hour), which is a clinically more relevant objective. For our LDA and SVM models, trained to classify 1-minute clips, we had to apply post-processing of the probabilities, which we chose to do by averaging over 60 probabilities. While this simple approach may constrain the performance of LDA and SVM compared to CNN, choosing among other heuristics, e.g. histogram projection [21], Kalman filters or a 'firing power' method [19], and tuning their hyperparameters, would drastically increase the risk of overfitting to the validation set.

As can be seen in Fig. 8, the CNN-based approach does not produce smooth probabilistic outputs. Instead, the CNN produces extreme probabilities close to binary values. As a result, specificity, sensitivity, and precision do not change significantly over different values of the decision threshold. In some cases, having binary outputs without the need for tuning a threshold, can be an advantage. On the other hand, LDA and SVM approaches do provide us with less extreme predictions which could be translated directly into decision confidence. This can be useful too, particularly in combination with high-quality binary outputs [39]. However, when the goal is to make a binary decision, LDA and SVM require a discrimination threshold, which binarizes the classifier's predictions throughout the testing phase of the device. We have shown that a carefully constructed cross-validation procedure cannot always provide a value of the threshold that would perform reasonably on the test set. Only the retrospective analysis of the performance curves revealed that a good value of the threshold for LDA and SVM would be around 0.2 for both subjects. Unfortunately, many studies that report binary classification metrics avoid explaining how the thresholds were chosen [19], [21], [40], [41].

To summarize, our analysis suggests the following guidelines which allow for a more objective evaluation of seizure prediction models.

- The parts of the EEG record used for model testing should not be temporally mixed with the data samples used for tuning the model parameters. The dataset partition should be causal such that a training set is followed by validation and testing blocks of EEG clips. We have shown that doing this differently will contribute to overly optimistic results.
- Evaluation metrics based on predictions aggregated from multiple subjects, should be used with a great caution. This is because such scores are greatly influenced by the differences between distributions of per-subject predictions.
- Having high AUC scores, as it was in the Kaggle Seizure Prediction Challenge, can lead to the belief that current algorithms are performing well at predicting seizures. To avoid creating this illusion, one should analyse the models in terms of multiple patient-oriented objectives. In this work, having done the event-based analysis of the precision, sensitivity, and specificity, we posit that cutting-edge models for seizure prediction have a limited value for patients with epilepsy.

V. Conclusion

On the example of the Kaggle Seizure Prediction Challenge we studied how important it is to correctly design datasets and choose evaluation techniques when dealing with EEG-based epileptic seizure prediction.

Our analysis suggests that improvements in the field of epileptic seizure prediction are likely to happen once the following is achieved. (1) The community establishes datasets of long-term EEG recordings with predefined train, validation and test sets. Here, it is crucial to split the EEG records into non-overlapping temporal blocks for training, validation and testing. (2) It turns into a common practice to benchmark the proposed algorithms against the existing ones, so it becomes clear what the state-of-the-art results are. (3) Evaluation of the seizure predictors is done not only in terms of standard objectives, but also using clinically relevant metrics.

As a technical contribution of this paper, we have proposed a novel CNN architecture This novel approach was compared to LDA and SVM methods. While it was not able to strongly outperform traditional methods, it is important for two reasons. (1) It offers a valuable alternative because of its ability to produce almost binary predictions for relatively long EEG segments. (2) We have demonstrated that proper regularisation enables training of complex neural networks on limited amounts of brain data.

References

[1] S. Shorvon and D. Goodridge, "Longitudinal cohort studies of the prognosis of epilepsy: contribution of the national general practice study of epilepsy and other studies," *Brain*, vol. 136, no. 11, pp. 3497–3510, 2013.

[2] A. Schulze-Bonhage and A. Kühn, *Unpredictability of Seizures and the Burden of Epilepsy*. Wiley-VCH Verlag, 2008, pp. 1–10.

[3] F. Mormann *et al.*, "Seizure prediction: the long and winding road," *Brain*, vol. 130, no. 2, pp. 314–333, 2007.

[4] C. A. Teixeira *et al.*, "Epileptic seizure predictors based on computational intelligence techniques: A comparative study with 278 patients," *Computer Methods and Programs in Biomedicine*, vol. 114, no. 3, pp. 324 – 336, 2014.

[5] J. J. Howbert *et al.*, "Forecasting seizures in dogs with naturally occurring epilepsy," *PLoS ONE*, vol. 9, no. 1, pp. 1–8, 01 2014.

[6] B. H. Brinkmann *et al.*, "Forecasting seizures using intracranial EEG measures and SVM in naturally occurring canine epilepsy," *PLoS ONE*, vol. 10, no. 8, pp. 1–12, 08 2015.

[7] E. E. Patterson, "Canine epilepsy: An underutilized model," *ILAR Journal*, vol. 55, no. 1, pp. 182–186, 2014.

[8] B. H. Brinkmann *et al.*, "Crowdsourcing reproducible seizure forecasting in human and canine epilepsy," *Brain*, 2016.

[9] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[10] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," *Computer Vision ECCV 2014. Lecture Notes in Computer Science*, vol. 8693, pp. 740–755, 2014.

[11] Y. Park *et al.*, "Seizure prediction with spectral power of EEG using cost-sensitive support vector machines," *Epilepsia*, vol. 52, no. 10, pp. 1761–1770, 2011.

[12] V. Vapnik, *Statistical Learning Theory*. New York: Wiley-Interscience, 1998.

[13] T. N. Alotaiby *et al.*, "EEG seizure detection and prediction algorithms: a survey," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 1, 2014.

[14] Y. LeCun *et al.*, "Gradient-based learning applied to document recognition," in *Intelligent Signal Processing*. IEEE Press, 2001, pp. 306–351.

[15] I. Korshunova and O. Thas, "Epileptic seizure prediction using deep leaning," 2015. [Online]. Available: http://lib.ugent.be/catalog/rug01:002214009

[16] K. A. Davis *et al.*, "A novel implanted device to wirelessly record and analyze continuous intracranial canine EEG," *Epilepsy Research*, vol. 96, no. 12, pp. 116 – 122, 2011.

[17] S. R. Haut *et al.*, "Seizure clustering during epilepsy monitoring," *Epilepsia*, vol. 43, no. 7, pp. 711–715, 2002.

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[19] C. Teixeira *et al.*, "Output regularization of SVM seizure predictors: Kalman filter versus the Firing Power method," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2012, pp. 6530–6533.

[20] M. Bandarabadi *et al.*, "Epileptic seizure prediction using relative spectral power features," *Clinical Neurophysiology*, vol. 126, no. 2, pp. 237 – 248, 2015.

[21] V. Cherkassky *et al.*, "Reliable seizure prediction from EEG data," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–8.

[22] A. Temko *et al.*, "Performance assessment for EEG-based neonatal seizure detectors," *Clinical Neurophysiology*, vol. 122, no. 3, pp. 474 – 482, 2011.

[23] S. Dieleman *et al.*, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," *Monthly Notices of the Royal Astronomical Society*, vol. 450, no. 2, pp. 1441–1459, 2015.

[24] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb 2015.

[25] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," *ArXiv e-prints*, 2016.

[26] P. Mirowski *et al.*, "Classification of patterns of EEG synchronization for seizure prediction," *Clinical Neurophysiology*, vol. 120, no. 11, pp. 1927–1940, November 2009.

[27] A. Krizhevsky *et al.*, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105.

[28] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[30] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[31] P.-J. Kindermans *et al.*, "A P300 BCI for the masses: Prior information enables instant unsupervised spelling," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 719–727.

[32] P. Kindermans *et al.*, "Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller," *Journal of Neural Engineering*, vol. 11, no. 3, p. 035005, 2014.

[33] P. von Bünau *et al.*, "Finding stationary subspaces in multivariate time series," *Physical Review Letters*, vol. 103, no. 21, p. 214101, 2009.

[34] W. Samek *et al.*, "Stationary common spatial patterns for brain–computer interfacing," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026013, 2012.

[35] W. Samek, "Divergence-based framework for common spatial patterns algorithms," *Biomedical Engineering, IEEE*, vol. 7, pp. 50–72, 2014.

[36] M. Sugiyama and M. Kawanabe, *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press, 2012.

[37] E. Dolgin, "Technology: Dressed to detect," *Nature*, vol. 511, no. 7508, pp. S16–S17, Jul 2014.

[38] D. Mishkin *et al.*, "Systematic evaluation of CNN advances on the ImageNet," *ArXiv e-prints*, 2016.

[39] A. Temko *et al.*, "Clinical implementation of a neonatal seizure detection algorithm," *Decis Support Syst*, vol. 70, pp. 86–96, Feb 2015.

[40] T. Netoff *et al.*, "Seizure prediction using cost-sensitive support vector machine," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Sept 2009, pp. 3322–3325.

[41] M. J. Cook *et al.*, "Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study," *The Lancet Neurology*, vol. 12, no. 6, pp. 563 – 571, 2013.