Available online at www.jbr-pub.org.cn

Open Access at PubMed Central

The Journal of Biomedical Research, 2020 34(3): 228–239

JBR

Original Article

Epileptic seizure detection using EEG signals and extreme gradient boosting

Paul Vanabelle[1,✉], Pierre De Handschutter[2], Riëm El Tahry[3,4], Mohammed Benjelloun[2], Mohamed Boukhebouze[1]

[1]Data Science Department, Centre of Excellence in Information and Communication Technologies, Charleroi 6041, Belgium;
[2]Computer Science Unit, Faculty of Engineering, University of Mons, Mons 7000, Belgium;
[3]Refractory Epilepsy Centre, University Hospital of Saint-Luc, Brussels 1200, Belgium;
[4]Institute of Neuroscience, Catholic University of Louvain, Brussels 1200, Belgium.

Abstract

The problem of automated seizure detection is treated using clinical electroencephalograms (EEG) and machine learning algorithms on the Temple University Hospital EEG Seizure Corpus (TUSZ). Performances on this complex data set are still not encountering expectations. The purpose of this work is to determine to what extent the use of larger amount of data can help to improve the performances. Two methods are explored: a standard partitioning on a recent and larger version of the TUSZ, and a leave-one-out approach used to increase the amount of data for the training set. XGBoost, a fast implementation of the gradient boosting classifier, is the ideal algorithm for these tasks. The performances obtained are in the range of what is reported until now in the literature with deep learning models. We give interpretation to our results by identifying the most relevant features and analyzing performances by seizure types. We show that generalized seizures tend to be far better predicted than focal ones. We also notice that some EEG channels and features are more important than others to distinguish seizure from background.

Keywords: epileptic seizure, electroencephalograms, Temple University Hospital EEG Seizure Corpus, machine learning, XGBoost

Introduction

One of the most common ways to diagnose epileptic seizure is to measure the electrical activity of the cerebral cortex by performing a non-invasive electroencephalogram (EEG)[1]. It is a relatively unexpensive and easy way to proceed compared to other techniques such as MRI or intrusive methods[2]. Consequently, most works to automatically analyze seizures have been done on these signals through time series processing methods.

EEG-based seizure detection has been extensively studied; however, it has rarely reached performance that could durably help neurologists. This is due to several factors. The complexity of the epilepsy and the brain waves is the first one. The EEG signals are non-

✉Corresponding author: Paul Vanabelle, Data Science Department, Centre of Excellence in Information and Communication Technologies, 28 Jean Mermoz Ave., Charleroi 6041, Belgium. Tel: +32-71-159-362, E-mail: paul.vanabelle@cetic.be.

Received 31 January 2019, Revised 04 April 2019, Accepted 03 June 2019, Epub 30 August 2019

CLC number: R742.1, Document code: A

The authors reported no conflict of interests.

https://doi.org/10.7555/JBR.33.20190016

stationary and the statistical features of these signals are different between patients over time. The second factor is the quality of the data sets used for the machine learning task.

An ideal data set needs to be a good representation of all the varieties of EEG signals that could appear during epileptic seizure. Even if high performances were obtained with some of the publicly available data sets, they have typically inherent restrictions. For instance, they can be too specific, regarding a certain type of population as the Children's Hospital Boston (CHB)-Massachusetts Institute of Technology (MIT) data set[3]. Indeed, this data set is only focused on teens and children, as it is generally recognized that EEG signals differ with age[4], it is potentially difficult to generalize the resulting work on adults. It is also particularly more dedicated to the prediction tasks thanks to the long duration of the recordings. Good results on this task are presented[5]. Data sets can also be too small or incompletely documented such as Bonn data set where there is no indication of the type of seizure and patients features[6]. With this one, very high performance with various deep learning architectures using raw data was achieved, but that seems hard to generalize on more sophisticated data sets[7].

Therefore, these experiments do not correspond to clinical conditions, and the results are not representative of current clinical performances. This state of fact initiated the development of the Temple University Hospital EEG Seizure Corpus (TUSZ). Up to now, it is the largest open source corpus of this kind and represents an accurate characterization of clinical conditions[8]. TUSZ is a complex data set offering a great diversity in terms of patients, seizures or EEG montages (the placement of the electrodes). Machine learning models have been proposed[8–9], and these works are based on deep architectures but the results are not yet convincing. Some architectures obtained good specificities but sensitivities are still low, resulting in a low rate of false alarms but at the price of missing a lot of seizures. The first versions of the data set were used for these studies. They were already offering more patients and a wider variety of seizures types for the studies than the other seizures corpora, the latest versions providing even more.

All these observations led us to the following question: Does the increase of data lead to better results? We saw that a common difficulty for generic modeling is the lack of data, both in quantity and quality, to obtain relevant results through machine learning and for having models that apply well to data from new patients.

To increase the amount of training data, we first worked on a more recent release of the TUSZ offering a larger training set, with a classic approach to machine learning and data partitioning. This gives the first method. The second method is based on a leave-one-out approach with iterations, in the manner of cross-validation, to extend the training set with the data of the test set. It means that, at each iteration, we keep a patient of the initial test set for the testing phase and perform the training and validation of the model on the rest of the data. XGBoost is the selected algorithm for our classifier. It is a fast implementation of the gradient boosting classifier which allows us to perform quickly the multiple iterations required by the second method. As it is a decision tree based algorithm, it has also the advantage to be interpretable; the measure of the importance of features on the training set is returned by the algorithm. As inputs of our algorithm, temporal and frequential features (computed on segments of 1 second) are used with some others processed by a specific python library for EEG signals, called pyEEG. The methods are applied to the two most important subsets of the TUSZ in terms of montage leading to multiple experiments.

As a result, with a fast computing approach, we managed to present results that matched the best in the Temple University Hospital (TUH) literature[8]. We showed that increasing the amount of data improves performance. We analyzed the performances by type of seizure and we showed that the seizures categorized as generalized were easier to recognize. Finally, the interpretable property of the XGBoost algorithm shows the features and the EEG channels that are the most discriminant to distinguish a seizure from the background, or ictal from interictal activity. Interpretability is important as much as the collaboration between data scientists and clinicians in the early steps of such a study[10]. Therefore, we report in the discussion section some ideas that resulted from our discussions with neurologists.

As the first official release of the TUSZ was done in April 2017, there is not a lot of available literature on machine learning techniques applied to this data set. Most of the experiments come from Prof. Picone's team at Temple University and have led to deep learning approaches.

Several architectures are extensively described[8] and the results presented were the first reported on this data set. In this paper, the authors report a number of 64 patients on whom algorithms were trained and a number of 50 patients for the testing. Various architectures are applied, including Hidden Markov Models (HMM), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN) combined with MultiLayer Perceptron (MLP) and a hybrid CNN/LSTM model. Dimensionality reduction is usually performed, through (Incremental) Principal Component Analysis [(I)PCA]. The results are based on metrics that are calculated permissively with the Any Overlap method (OVLP); it consists in considering a totally good detection if a detected event touches at least a period of time (a segment) of the real event, instead of calculating the metrics segment by segment. Specificity varies between 70% and 80% except in the hybrid CNN/LSTM model where an outstanding specificity of 96.86% is reached, resulting in a very low false alarm rate of 7 FA/24 hours. On the other hand, sensitivity is always low, 30% for this last model and up to 40% for the others with a false alarm rate at minimum 77 FA/24 hours.

A similar work uses Gated Recurrent Units (GRU), a particular kind of Recurrent Neural Networks (RNN) that is faster but less accurate than LSTM[9]. A CNN/GRU approach is compared to the CNN/LSTM approach and shows the same sensitivity (30%) and a slightly lower specificity (91%). Initialization techniques and regularization methods are also discussed. Although version 1.1.1 of TUSZ was reported to be used in this case, with 196 patients for training and 50 for testing, the results remained in the same order of magnitude compared to the first paper.

Hence, none of as of today state of art models have been able to reach satisfying sensitivities and, except a deep complex CNN/LSTM architecture with many layers, specificities, and false alarm rates are not outstanding neither.

Gradient Boosting is an ensemble method using boosting principles. In short, it fits a classifier (generally a decision tree) to the data given in input and calculates residuals. A new classifier is then adjusted on these residuals. The procedure is repeated while the validation error decreases. There are many advantages to using gradient boosting models based on decision trees; they have the innate property of being robust to correlated features, the normalization of these ones is not required and finally estimates of feature importance can be provided. This last property can give significant insights into the features and on the analysis process.

XGBoost library is an efficient and distributed implementation of the Gradient Boosting algorithm[11]. The main strength of XGBoost is its scalability which allows parallel and distributed computing and makes learning and model exploration faster. Moreover, several other improvements were added such as techniques for reducing overfitting giving to XGBoost better performances than the generic boosting algorithm. We used the XGBoost library through the python package for this work.

Although Gradient Boosting is known to reach high performances in several tasks and is easier to parameterize than deep learning architectures, there was no use of it on "Big Data" seizure detection corpus such as TUSZ, to the best of our knowledge. However, we have to mention[12] where a Gradient Boosting approach is used on Freiburg data set (which is not publicly available anymore) made of 21 patients. Their approach is patient-specific in the sense that they built a custom model for each patient. The data of one patient is actually split in a train subset and a test subset, before feeding a model. The procedure is repeated for all the other patients, resulting in as many models as patients. In the present work, we use the opposite approach which consists of generic modeling.

Concerning the interpretable property we are trying to exploit by means of XGBoost, we must mention that several other works approach the problem in a similar way[13–14]. Both works consider generic modeling on multiple patients and use the Bonn data set for this purpose. The first one uses decision tree-based modelization, including a random forest classifier. The conclusion emphasizes the good performance of the random forest algorithm which is an ensemble method similar to gradient boosting. The second tries to establish rules from a fuzzy logic system. The interpretability of the resulting fuzzy rules is discussed. Bonn data set only provides one EEG channel which limits the scope of interpretation.

Materials and methods

Data set

TUSZ is a subset of the larger EEG signals database from the TUH. TUSZ data set is remarkable by the high number of patients and configurations it contains. Indeed, in the considered version 1.2.1 of this subset released in April 2018, there are 266 patients in the training set (of whom 118 suffering from seizures), 50 in the test set (of whom 38 suffering from seizures), and more than 40 different configurations of electrodes.

TUSZ data set was established by identifying the sessions in the TUH EEG Corpus that were the most likely to contain seizure events. Three annotation tools were used to identify events of clinical interest: Natural Language Processing (NLP), a commercial software, and a three-pass system[15]. Then an annotation team manually annotated the data and showed that the NLP approach had been the most efficient as the first step, revealing again the difficulty to automatically detect seizure events in a large data set of EEG[16].

This data set contains a huge diversity of configurations, in terms of sampling rate, references or montages (placement and connections between the electrodes). However, two major montages predominate in the data set: "Average Reference" (AR) for which the potentials are measured relatively to the average value of a subset of electrodes and "Linked Ears" (LE) for which the reference is located on the ear, electrically quiet. Previous researches offered more statistics about TUH data sets[16–17] and more considerations on the montages and their consequences[18].

Nevertheless, all recordings contain the electrodes of the standard 10/20 placement. We call "channels" the derivations of electrodes obtained with the TCP bipolar montage also called "double-banana". They are shown in ***Fig. 1***, which comes from a previous study[19].

The recordings are split into six file folders: three theoretically for training purpose and the three remaining for the test. Each folder contains one specific montage, either AR, LE, or a slight variant of AR. This variant is less represented and was not considered for this study.

For each patient, there are one or several sessions themselves containing files related to one or more recordings. There are five files for each recording. The .edf file contains the raw EEG signals and a header containing useful information (frequency, duration, date). The .lbl file contains event-based annotations that catch the beginning and end of each event on every channel of the 22 shown in ***Fig. 1***. An event is either background or a particular type of seizure. Practically those annotations may indicate if an ictal phase begins or ends earlier or later on some channels. The .lbl_bi is very similar to the previous one except that the events are just binary (bckg and seiz). The .tse file contains label-based annotations that catch the beginning and end of each event globally for all the channels. These are the most frequently used in Machine Learning research as they allow to give a single overall target for each time step. The .tse_bi file is very similar to the previous one except that the events are just binary (bckg and seiz). For each session, there is also a .txt file that contains general information about the patient, the recordings, in a non-structured way.

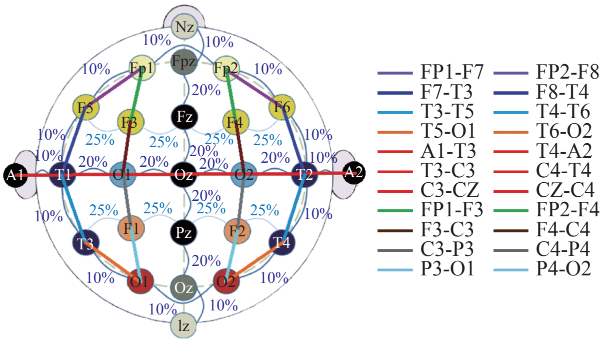

Fig. 1 **Placement and connections of electrodes in a 10/20 placement with bipolar TCP montage.**

Epileptic seizures are usually classified into two main categories. The first category is the generalized crisis which begins bilaterally and synchronously from the outset in both cerebral hemispheres. The second category is the partial or focal crisis which starts in a localized part of the brain. It can be accompanied by a loss of consciousness; the abnormal electrical activity that started from the focused location in the brain can gradually extend to the entire brain; the crisis may end in convulsions.

However, TUSZ is following a more refined classification[20]. In this data set 10 different types of seizures are defined, though some of them are less represented. Several types of seizures are covered by the generalized class or by the focal one. Among the generalized seizures, we find the "absence" which results in a momentary alteration of consciousness, the "tonic crisis" characterized by a strong variation of muscle tone, the "clonic crisis" characterized by jerks and the generalized "tonic-clonic" crisis characterized by convulsions, and also the atonic and myoclonic seizures. Among the focal seizures, we can specifically distinguish the complex and simple partial seizures if respectively there is a loss of consciousness or not. In ***Table 1***, we have extracted from the data set all the information about the distribution in terms of segments of each type of seizures in several folders. The segments have in the present case a duration of one second. We can note in this table that non-specific seizures are dominant.

Motivation of the proposed approach

Since we worked on the TUSZ data set, we tackled the problem of seizure detection through different angles. Prior work was based on the transformation of EEG signals into features and the comparison of several classification models. We noticed during that experiment the good performances of ensemble learning methods as random forest and gradient boosting. Moreover, the use of high-performance version of gradient boosting like XGBoost allowed to considerably reduce the training time, and thus to be more focused on fast iterations and optimization of parameters and hyper-parameters. We also noticed that the results were good when training and test sets were built after mixing non-overlapping ictal and inter-ictal segments on the whole population of recordings. Unfortunately, this method led to a kind of

Table 1 **Seizure segments distribution by type in the different folders**

Seizure code	Seizure type	AR train	AR test	LE train	LE test
FNSZ	Focal non-specific	13 088	23 301	10 186	3 418
GNSZ	Generalized non-specific	6 743	14 226	10 230	122
SPSZ	Simple partial	1 327	364	167	0
CPSZ	Complex partial	8 551	8 114	5 811	784
ABSZ	Absence	14	0	479	339
TNSZ	Tonic	424	857	0	0
CNSZ	Clonic	0	0	0	0
TCSZ	Tonic clonic	795	1 335	938	2 343
ATSZ	Atonic	0	0	0	0
MYSZ	Myoclonic	0	1 178	0	0

AR: average reference; LE: linked ears.

data leakage as a segment of one recording was not really different from its direct neighbor of the same state. Hence, results were less impressive when we tried to classify EEG segments on new patients. This work was using the first version of the data set and limited to 59 patients for the training.

Interested by the abilities and the promises of some deep learning algorithms to directly extract features from complex time-series without preprocessing, we investigated that field without the expected success, as dealing with raw data is really resources-demanding. Another attempt by using EEG features with a deep learning algorithm (Long Short-Term Memory, LSTM) came crashing down on the wall of bad performances related by the TUH literature[8].

Based on these previous approaches, we made several observations. Firstly, the EEG features approach is still relevant for performance issues. Secondly, XGBoost is efficient and allows it to quickly iterate. Furthermore, there is a bad generalization of new patients in the test set.

Taking up the last point, there are two ways to solve this issue. Either we move backward and consider patient-specific modeling, or we increase the amount of training data for generic modeling. The second choice is more relevant and it would be indeed interesting to see to which extent the increase of the amount of data would help to improve performances. EEG features and XGBoost will be the baseline of the two methods introduced hereafter. The first method aims to evaluate the performance of a classic split in terms of train, validation and test sets on version 1.2.1 of TUSZ. With the second method, we search to increase the training data set by replacing the classic split into a leave-one-out approach. It is a common technique used to increase a training set when there is a lack of data.

First method: EEG Features+standard partitioning+XGBoost

The method is presented in ***Fig. 2***. We first applied a preprocessing step to extract features from the raw EEG signals. These features are temporal and frequential. We based ourselves on the literature review and Python's library PyEEG[21]. A total of 22 features were then considered, which are statistical parameters (mean, variance, skewness, kurtosis, interquartile range, minimum and maximum), other temporal features (Hjorth complexity and mobility, Petrosian fractal) and frequency features (spectral density in the five frequency rhythms: alpha, beta, gamma, delta, theta), the corresponding ratios, spectral centroid, and monotony. As for the frequential features, the five frequency ranges were defined according to previous studies[22]. Delta range was the lowest, between 1 and 4 Hz, theta was between 4 and 8 Hz, alpha between 8 and 13 Hz, beta between 13 and 35 Hz and gamma 35 Hz upwards. Gamma rhythm was rarely modified by a seizure.

We split the raw EEG signal into non-overlapping 1-second segments and computed these features for each of the 22 channels in the bipolar montage, resulting in a 22×22=484 features vector for each segment. This preprocessing step was computed offline and each feature matrix *n_segments*× *n_features* was stored by recording.

After this preprocessing step, the segments of the TUSZ training set were randomly split in 80% of train and 20% of validation. Then XGBoost with 400 estimators of depth 3 was trained. The ratio parameter was used to take into account class imbalance between non-seizure and seizure segments. The learning rate was set to 0.1. To avoid problems of speed and memory, we chose to work with "gpu_hist" as tree method, which is the XGBoost fast histogram algorithm for GPU. The internal evaluation metric for the validation data was set to "AUC".

Once the model converged, the model was tested on

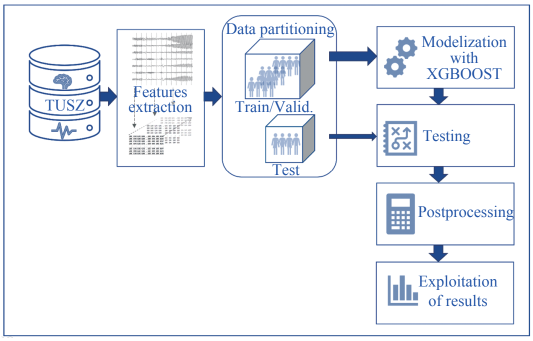

Fig. 2 **First method based on a classical.**

all the recordings of the test set. For each recording, the model tried to predict the class of each segment. A smoothing function was then applied to those predictions to avoid isolated and improbable seizure/non-seizure states. This function was a sliding window on 19 segments applying a simple majority vote. At the end of the test set classification, we had confusion matrices for all the recordings and we could compute confusion matrices by patient or by types of seizures. Similarly, an overall confusion matrix was computed by summing all the recording confusion matrices.

Second method: EEG features+LOOCT+XGBoost

The second method is similar to the first one except that we applied a specific partitioning technique to increase the training data. This method, detailed in ***Algorithm 1***, is close to a cross-validation method called leave-one-out cross validation (LOOCV). Actually, one should rather call it in this case LOOC-Test as explained below, as a single patient is left out for the test at each iteration. Similarly to a cross-validation method, we combined at the end the test results from the multiple rounds to come up with an estimate of the model's predictive performance. This second method is summarized in ***Fig. 3***.

For each recording, we had a matrix of size matrix $n\_segments \times n\_features$ and the binary labels for each time segment (seizure or background). In the algorithm, one can see X as a list of length *number_of_patients* where each component is made of these features matrices for all the recordings of that patient, while y is the corresponding binary target vector. We also needed to store separately the IDs of the patients in the TUSZ train sets and test sets.

Then, the algorithm could be applied. At each iteration, one patient was left out of the training data and was kept for the test. The segments of the other recordings were randomly split in 80% of train and 20% of validation. Then similarly to the first method, a XGBoost model was trained, conserving the same parameters.

Algorithm 1 **Leave-one-out cross test with XGBoost**

```
for i in test_patients_indexes do
    LeaveOneOutPatient= (X, y)[i]
    other-patients: (X, y)\LeaveOneOutPatient
    XGBoost.fit (other_patients, validation_split=0.2)
    for record in LeaveOneOutPatient do
        pred=XGBoost.predict (X[record])
        pred=smooth(pred)
        Confusion_matrix (y[record], pred)
    end for
end for
```

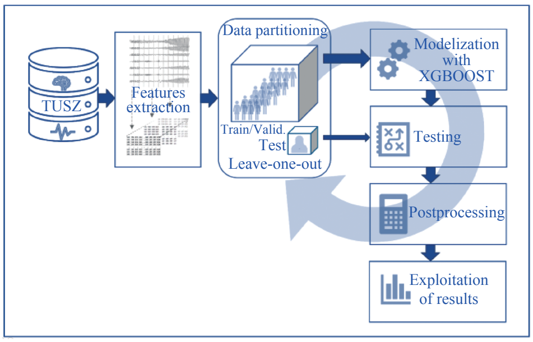

Fig. 3 **Second method based on a LOOCT approach.**

Once the model converged, the model was tested on each recording of the patient left away. Thus, for each recording, the model tried to predict the class of each segment. The smoothing function was still applied as post-processing.

The number of iterations of the LOOCT method was equal to the number of patients in the original test set present in the TUSZ. These were the only patients that are "left out" by the algorithm in order to have a kind of comparison to the first method in terms of population.

At the end of all the iterations, we had confusion matrices for all the recordings and we could compute a confusion matrix by patient or by type of seizure. Similarly, an overall confusion matrix was computed by summing all the recording confusion matrices.

Metrics

The seizure detection is a classification problem, thus the evaluation of the models was based on confusion matrices (***Fig. 4***). The number of true positives (TP) was in this case the number of EEG segments correctly classified as seizure; the true negatives (TN) were the number of EEG segments correctly classified as background; the false positives (FP) were the number of EEG segments classified as seizure that were actually background (False Positives are also known as false alarms in the medical field); the false negatives (FN) were the number of EEG segments classified as background though it is actually seizure.

The confusion matrices are used to compute some interesting ratios. Specificity, also known as true negative rate, is defined by

$$Specificity = \frac{TN}{TN + FP} \tag{1}$$

Sensitivity (or Recall) is the probability to detect an element of the positive class, and it is defined by

		Predicted class	
		Background	Seizure
Actual class	Background	True negatives (TN)	False positives (FP)
	Seizure	False negatives (FN)	True positives (TP)

Fig. 4 **Confusion matrix for seizure detection purpose.**

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

Precision is the fraction of correct instances among all the detected instances, it is defined by

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Finally, the last metric we had been interested in is a measure of accuracy, a metric that evaluates the overall quality of the model. The standard accuracy might not be the best metric to look at, as the classes in the present case are highly imbalanced. We used instead another measure of accuracy, the more appropriated F1-Score, which is the harmonic average of the precision and recall

$$F1 = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \tag{4}$$

In the literature presented above, sensitivity and specificity were mainly used, as well as the false alarm rate. It is obviously an important metric in the case of seizure detection, and we also communicated results in this sense for comparison. Note that, in this literature, sensitivity and specificity were computed using the OVLP[8,23]. OVLP is more a method to calculate metrics in a permissive way than a metric *per se*. OVLP is based on an event decomposition of the EEG signal. Instead of looking at each segment (also called epoch), OVLP only considers the events (*i.e.* continuous sequence of seizure or background) that occur. ***Fig. 5*** shows a potential drift of the method. The signal above shows the reference, divided in segments of equal length while the signal below is the hypothesis, *i.e.* the sequence predicted by the model. While one would count true/false negatives/positives by matching each pair of corresponding segments, OVLP considers that the true positive rate is 100% as soon as the hypothesis overlaps the reference for at least *lambda* segments, where *lambda* is often set to 1 to guarantee that the events do overlap. Thus, in the example, true positive rate is 100% while traditional metrics only counts one segment as a true positive.

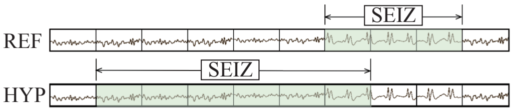

Fig. 5 **Illustration of Any Overlap method.**

In ***Table 1***, we have deliberately retrieved the distribution of types of seizures in terms of seizure segments instead of events, because it seems to us that the impact on the models of the number of segments is more relevant and reliable. Effectively the global ratios of events and of segments are different. For instance, in AR train, there are 77 events of non-specific generalized seizures for 6 743 segments. In the LE train, there are 35 events for 10 230 segments. It means that the non-specific generalized seizures are way longer in the montage LE. This information is hidden if we count by events.

Results

Firstly, to get significant results in reasonable time, we started by testing the two described methods on the AR test set. The results are thus corresponding to 36 patients, 899 recordings containing 562 seizure events, and more than 49 000 seizure segments. The population of each subset is recalled in ***Table 2***. The events and segments are divided between all types of focals and all types of generalized seizures. As we mentioned before, ***Table 1*** presents a more detailed classification, but the two non-specific sets are dominant and the other types are not enough populated, so for a more refined analysis regarding the results and the content of the corresponding training sets, we have decided to consider the two main seizure classes: focal and generalized seizures.

The results of the experiments are summarized in ***Table 3***. The ones concerning the AR Test as target are in the first section and numbered from 1 to 4. As one can see, due in particular to the LOOCT method, we took a few liberties with the initial destinations of each subsets. So, in the first section for instance, AR Test was used in the training set as well. For this round of experiments, sensitivities by type of seizure are summarized in ***Table 4***.

We were expecting a bit more from the second method, that is why we started a second round of experiments with LE Test as test set. This subset was smaller than the AR Test; 82 events against 562 for the AR Test. Note that these subsets have 10 patients in common, and these ones were thus filtered out from the training set for the cases where the AR Test set was used in this role. These experiments are numbered from 5 to 9 in ***Table 3***.

Table 2 **Cardinalities of the selected sets**

	Patients	Recordings	Focal events	Generalized events	Focal segments	Generalized segments
AR train	124	1 220	376	102	22 966	7 976
LE train	129	322	310	100	16 164	11 647
AR test	36	899	312	250	31 779	17 596
LE test	24	49	38	44	4 202	2 804

Table 3 **List of experiments – selected sets, methods and results**

	AR train	LE train	AR test	LE test	First method	Second method	Sensitivity	Precision	Specificity	F1 (%)
1	train		test		x		35.80	31.65	92.46	33.60
2	train	train	test		x		41.74	35.17	92.50	38.17
3	train		train-test			x	52.66	27.55	86.50	36.17
4	train	train	train-test			x	53.87	27.57	86.20	36.47
5	train			test	x		71.65	28.67	73.04	40.95
6	train	train		test	x		66.70	34.06	80.47	45.09
7	train	train	train	test	x		78.72	33.54	76.41	47.04
8	train	train		train-test		x	72.59	39.61	83.26	51.25
9	train	train	train	train-test		x	71.61	40.01	83.76	51.33

Table 4 **Sensitivity by type of seizure**

Seizure type	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Focal	23.02%	27.68%	35.02%	37.61%
Generalized	59.50%	66.24%	81.51%	81.10%

Discussion

On overall results

The first observation answers the question of this work: by increasing the amount of data, we improve significantly the detection score. However, by looking at the F1-score, this needs to be discussed as it is mainly true for the first method using the classic split of the data set. As one can see, in the experiment 2, by adding LE Train to the training set, all metrics are improved compared to the experiment 1, where the training set only consisted of the AR Train set; the F1-score increases by 5% to reach 38%. That is less clear with the second method LOOCT, the sensitivities are higher than those with the first method, but we observed a high number of False Positive, which degrades the precision and the specificity. This translates into a stagnant F1-score at 36% for the experiments 3 and 4.

The second round of experiment with LE Test as test set confirmed the interest of adding the LE train set in the training set. The use of the second method shows in this case a good improvement of the F1-score, we are 4% over the best F1-Score obtained with the first method (experiment 7). We also observed that the addition of the AR Test in the training set do not bring the strong improvement that we could expect. Even, in the seventh experiment, we observed a decrease in the precision and the specificity as we had in the first round of experiments. This tends to say that the problem is not the second method but the use of the AR Test set as training material. This observation must be nuanced by the difference in size between the AR Test and LE Test sets. Nevertheless, following a discussion with the TUH Team, we learned that they have visually observed that LE files were much cleaner compared to AR files in terms of signal noise ratio and that it would make sense to get bad performances on AR compared to LE files. Moreover, previous studies discussed the importance of the montage and the reference point as it practically changes the nature of the waveforms even by using the conversion to the bipolar montage[19]. Another explanation for the problem of adding AR Test in the training set could lie in the properties of the inter-ictal seizure signals. ***Fig. 6*** shows a recording in the AR Test set. Our system, based on frames of 1 second, detects a seizure state from the beginning of the record. The TUSZ annotations indicate the beginning of the seizure at 7 seconds. The EEG patterns are the same but these events are more considered as inter-ictal anomalies, as their duration is under 10 seconds. We can imagine the problem of adding these "background" annotated segments in a training set. The intuition is that the accuracy would be penalized by increasing the false negative and false positive rates. On the other hand, it is also probably the case in the other subsets dedicated to the training set, the only difference is that the proportion of patients with epilepsy is higher in the AR Test. These observations also explain why performances are so low on the TUSZ data set. To summarize, this advocates for two things. Firstly, a better strategy at the level of the model for handling these inter-ictal anomalies. Secondly, the training data should be selected cautiously.

Comparatively to the performances reported in the TUH literature, these results are of the same order as the ones of their best model[9–10]. The sensitivity is higher, while the specificity is a bit lower. In terms of false alarm rate for the patients without epilepsy in the test set, 5 events in 3 recordings were detected on a total of 82 recordings in the second experiment; the false alarm rate was then estimated around 8 FA/24 hours. If all the recordings without seizure are considered, the false alarm rate rises to 18 FA/24 hours. This denotes a good capacity of the model to distinguish patients with epilepsy from the others. Furthermore, if we compute the scores by records and not by segments, the precision rises to 64.33%, which means that there's a majority of recordings correctly classified on all the recordings detected with a pathology of epilepsy. The other metrics are unchanged.

On the results by type of seizure

If we look at the sensitivities by type of seizure for the first round of experiments in ***Table 4***, we can make the same observations in terms of improvement in function of the used training sets. Moreover, by making a proportional average of these scores with the amount of focal and generalized segments in ***Table 2***, we can find back the sensitivities in ***Table 3***. The main information that we can draw from this view is that the generalized seizures are detected more efficiently compared to focal seizures. This was expected; nevertheless, we have to point out that the proportion in the training set is in disfavor of the generalized seizures, this one oscillating between 25% (AR Train) and 40% (LE Train). The difficulty to detect focal seizures is due to the fact that they begin in one specific area of the brain with later a potential generalization, while the electrical discharge of the generalized seizures is directly widespread in both sides of the brain.

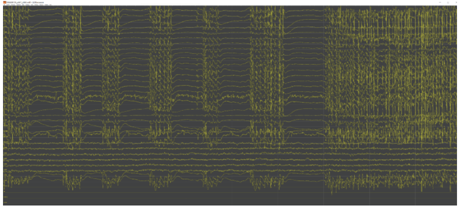

Fig. 6 **Repetitive inter-ictal generalized spike and wave discharges alternated with global EEG signal attenuation, followed by a generalized seizure (7 seconds from the start of the segment).**

On the channel and feature importance

This advocates to analyze the seizure detection problem with a type-specific approach instead of a generic modeling; however, as argument for the generic approach, 11 patients on the 36 of the AR Test encountered both focal and generalized seizures in the same recordings or in several. Nevertheless, some strategies should be adopted to deal with the different locations of the onsets of the focal seizures. For instance, a new metric that takes into account these preoccupations, such as the importance of characterizing the beginning of a seizure, would be meaningful.

The idea of type-specific analysis was recalled by the study of the most important features and channels that were more decisive during the training to discriminate seizures from background. Indeed, the coverage of specific focal seizures in the training set may have an incidence on the importance of the channels. Among all the experiments and the training sets, a feature on two channels emerged, namely the power spectral density on the beta frequency band for the channels P4-O2 and P3-O1. These two particular channels were located symmetrically in the posterior region of the head between the parietal and occipital area. After discussion of our results with neurologists specialized in epilepsy from the University Hospital of Saint-Luc in Brussels, we hypothesized two possible explanations.

The first one refers indeed to the type of seizure from the location point-of-view. The locations of these features correspond to the place where the focal seizures of occipital origin start. Generalized seizures can also induce ictal changes in this area due to their widespread nature, they are nevertheless a minority in the training sets, and focal seizures are most frequently observed. Considering the general statistics of epilepsy among all populations, occipital seizures have been reported to constitute 8% of the population with epilepsy and are occasionally subject to a secondary generalization[24]. The dominant type of the focal seizures is generally considered to be the Temporal Lobe Epilepsy (TLE) which represents more than 60% of all focal seizures[25]; this type of seizure rarely spreads to the occipital lobes. Taken together, all these statistics do not directly explain the location and the importance of these channels. Moreover, neurologists reveal that a bias can exist in the statistics of hospitals specialized in epilepsy compared to the whole population with epilepsy, as patients with refractory epilepsy (pharmacoresistant epilepsy) are more likely to be redirected through these institutions. The Temple University Hospital is indeed accredited as a level 4 epilepsy center, which means that it has the professional expertise to provide the highest level medical and surgical treatment for patients with refractory epilepsy[26]. To sort out this question, we need to determine the exact locations of all the included seizures from the data set. This information is not explicitly given in the TUSZ, but it could be retrieved by processing the .lbl files presented before, which can be the subject of future work.

The second potential explanation concerns the general physiological activity in the occipital lobe. When reviewing the EEG, neurologists first consider the occipital region; as a normal background activity is generally observed in these channels, with EEG signals less polluted by artefacts like muscle or eye movements, it helps to interpret the onset of an eventual seizure. If this empiric method reveals indeed visual discrimination of ictal and interictal phases, it is likely that the algorithm has found the same rule. Furthermore, a study from the TUH aims to measure the performance of their model if fewer channels are considered[27]. The removed channels were chosen by using domain knowledge instead of using a proper automated and optimized selection process. If only two channels were to be preserved, one of them would be located in the occipital region and the other bound to the CZ electrode for its central location. The removal of the other channels is justified either by their susceptibility to noise or artefact interference, or by their less strategic location. So, to conclude this point, experts seem to be concerned by the activity in the posterior region of the head and our interpretable models suggest this location for the most important features. The validation of this relation will also be the subject of future work.

Below the power spectral density on the beta frequency band for the channels P4-O2 and P3-O1, we can find other significant combinations. As there were 22 features for 22 channels, instead of enumerating them, we did a global analysis presented in two figures. ***Fig. 7*** shows the averaged channel importance over all the features and ***Fig. 8*** shows the averaged feature importance of all the channels. For the channel importance, we find first the two already mentioned channels between the occipital and parietal area, then some others between the central and the parietal area or between the temporal and occipital area or purely temporal. Two things are interesting to note. Firstly it is not symmetric, as if the channels in the right hemisphere are more important; secondly, the less important channels are related to the frontal area, indicating that EEG is less sensitive to abnormalities in the frontal region. Indeed, the frontal lobe is a big lobe and many regions, especially deep mesial ones are not well captured by the EEG[27–28]. For the feature importance, the power spectral density on the beta frequency band confirms to be the first on average, closely followed and far above the others by the Petrosian Fractal Dimension. Various studies have shown the interest of using fractals for biosignals analysis and seizure detection[29–30]. For the remaining features, we note the good performance of the other power spectral densities in alpha, gamma, and theta frequency bands and of the variance.

About the perspectives

In this paper, we analyzed the problem of seizure detection using machine learning techniques on the TUSZ with as main goal to understand whether increasing the amount of data increased performance of seizure detection. By using both a classic machine learning approach and a leave-one-out method on two subsets, we generated experiments based on training sets that have the characteristic of being large, heterogeneous and not commonly used in the current state-of-the-art. Using an efficient implementation of the gradient boosting algorithm called XGBoost, we showed the potential of improvement that can be done by adding more training data, even if this one should be selected cautiously. We presented results of the same order of magnitude than the corresponding references based on deep learning architectures. The performances by type of seizure were also analyzed and showed that generalized seizures are easier to recognize, even if they were under-represented in the training set. Finally, the interpretable property of the XGBoost algorithm showed that a feature linked to two specific EEG channels between the parietal and occipital area was statistically the most discriminant to distinguish a seizure from a non-seizure (considering segments of one second and a generic modeling done with the proportions by type of seizure proposed by the TUSZ).

In further works, we would like to extend this approach to the future releases of the TUSZ with more data. A larger data set means that the LOOCT method would not be necessary anymore. Based on the actual findings, we are planning to study an architecture that would be more able to discriminate ictal phase from

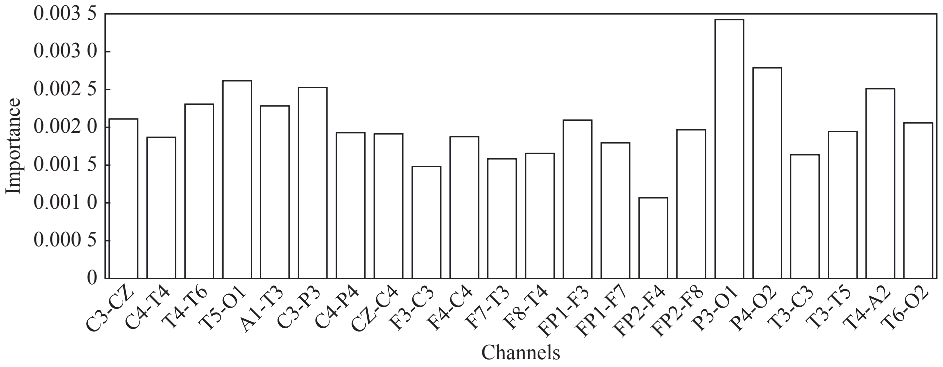

Fig. 7 **Channel importance-mean on all feature types.**

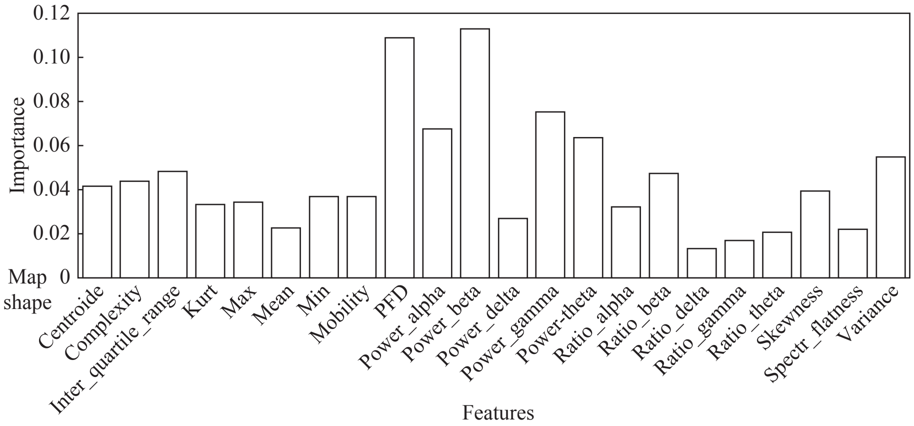

Fig. 8 **Feature importance-mean on all channels.**

inter-ictal, and also a type of seizure from another, either by using a multiclass classification approach by distinguishing focals and generalized seizures, either by leaving aside the generic approach or by focusing on type-dependent models. The use of interpretable models is still relevant and it would be interesting to see how the importance of the features evolves in function of the focal seizures and their onset locations. We are also interested to do some process mining based on the by-channel annotations by looking at the order of (dis) appearance of a seizure on particular channels in order to identify channels that are the most critical. One of the goals would be to converge to a system capable of early detection. Such a fast detection is critical for clinicians because it allows nurses to react as quickly as possible with the appropriate actions.

From a technical point of view, deep learning models could be tested again. Studying other sources of signals such as electrocardiograms (ECG) would also be interesting but once again, we suffer from a lack of data resources.

Acknowledgments

This work was carried out as part of the BIG DATA PLATFORM project (https://www.cetic.be/Big-Data-Platform), supported by the Région Wallonne (convention number 7481).

We warmly thank Professor Susana Ferrao Santos, neurologist and epileptologist (Refractory Epilepsy Centre, University Hospital of Saint-Luc, Brussels), and Simone Vespa, Research Assistant (Institute of Neuroscience, Catholic University of Louvain, Brussels), for their interest in our work and for the constructive discussions we had.

We would also like to thank Professor Joseph Picone and his colleagues Vinit Shah and Meysam Golmohammadi (Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA, USA) for their valuable feedbacks to our questions related to the TUH corpus.

References

[1] Elger CE, Hoppe C. Diagnostic challenges in epilepsy: seizure under-reporting and seizure detection[J]. *Lancet Neurol,* 2018, 17(3): 279–288.

[2] Smith SJM. EEG in the diagnosis, classification, and management of patients with epilepsy[J]. *J Neurol Neurosurg Psychiatry,* 2005, 76(Suppl 2): ii2–ii7.

[3] Shoeb AH. Application of machine learning to epileptic seizure onset detection and treatment[D]. Massachusetts: Massachusetts Institute of Technology, 2009.

[4] Bell MA. The ontogeny of the EEG during infancy and childhood: Implications for cognitive development[M]// Garreau B. Neuroimaging in Child Neuropsychiatric Disorders. Berlin, Heidelberg: Springer, 1998: 97–111.

[5] Tsiouris KM, Pezoulas VC, Zervakis M, et al. A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals[J]. *Comput Biol Med,* 2018, 99: 24–37.

[6] Andrzejak RG, Lehnertz K, Mormann F, et al. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state[J]. *Phys Rev E,* 2001, 64(6): 061907.

[7] Ahmedt-Aristizabal D, Fookes C, Nguyen K, et al. Deep classification of epileptic signals[C]//Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Honolulu, Hawaii, USA: IEEE, 2018.

[8] Deep Learning Approaches for Automated Seizure Detection from Scalp Electroencephalograms[M]//Obeid I, Selesnick I, Picone J. Signal Processing in Medicine and Biology. Cham: Springer, 2020: 235–276.

[9] Golmohammadi M, Ziyabari S, Shah V, et al. Gated recurrent networks for seizure detection[C]//Proceedings of 2017 IEEE Signal Processing in Medicine and Biology Symposium. Philadelphia, PA, USA: IEEE, 2017: 1–5.

[10] Itani S, Lecron F, Fortemps P. Specifics of medical data mining for diagnosis aid: a survey[J]. *Expert Syst Appl,* 2018, 118: 300–314.

[11] Chen TQ, Guestrin C. Xgboost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2016: 785–794.

[12] Zhang YL, Zhou WD, Yuan SS, et al. Seizure detection method based on fractal dimension and gradient boosting[J]. *Epilepsy Behav,* 2015, 43: 30–38.

[13] Wang GJ, Deng ZH, Choi KS. Detection of epileptic seizures in EEG signals with rule-based interpretation by random forest approach[C]//Proceedings of the 11th International Conference on Advanced Intelligent Computing Theories and Applications. Fuzhou, China: Springer, 2015: 738–744.

[14] Jiang YZ, Deng ZH, Chung FL, Wang G, et al. Recognition of epileptic EEG signals using a novel multiview tsk fuzzy system[J]. *IEEE Trans Fuzzy Syst,* 2017, 25(1): 3–20.

[15] Golmohammadi M, Torbati AHHN, de Diego SL, et al. Automatic analysis of EEGs using big data and hybrid deep learning architectures[J]. *Front Hum Neurosci,* 2019, 13: 76.

[16] Shah V, von Weltin E, Lopez S, et al. The temple university hospital seizure detection corpus[J]. *Front Neuroinform,* 2018, 12: 83.

[17] Obeid I, Picone J. The temple university hospital EEG data corpus[J]. *Front Neurosci,* 2016, 10: 196.

[18] López S, Gross A, Yang S, et al. An analysis of two common reference points for EEGs[C]//Proceedings of 2016 IEEE Signal Processing in Medicine and Biology Symposium. Philadelphia, PA, USA: IEEE, 2016: 1–5.

[19] Ferrell S, Mathew V, Ahsan T, et al. The temple university hospital EEG corpus: electrode location and channel labels (Report)[R]. Philadelphia, Pennsylvania, USA: The Neural Engineering Data Consortium, Temple University, 2019.

[20] Berg AT, Berkovic SF, Brodie MJ, et al. Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE commission on classification and terminology, 2005-2009[J]. *Epilepsia,* 2010, 51(4): 676–685.

[21] Bao FS, Liu X, Zhang C. PyEEG: an open source python module for EEG/MEG feature extraction[J]. *Comput Intell Neurosci,* 2011, 406391.

[22] Miller R. Theory of the normal waking EEG: from single neurones to waveforms in the alpha, beta and gamma frequency ranges[J]. *Int J Psychophysiol,* 2007, 64(1): 18–23.

[23] Wilson SB, Scheuer ML, Plummer C, et al. Seizure detection: correlation of human experts[J]. *Clin Neurophysiol,* 2003, 114(11): 2156–2164.

[24] Duncan JS. Chapter 15-occipital and parietal lobe epilepsies [EB/OL]. ILAE Lecturer, Fifteenth Teaching Weekend on Epilepsy. [2019-03-01]. https://www.epilepsysociety.org.uk/lecture-notes-0#.XHj_6YhKguU.

[25] Diehl B, Duncan JS. Chapter 13-temporal lobe epilepsy [EB/OL]. ILAE Lecturer, Fifteenth Teaching Weekend on Epilepsy. [2019-03-01]. https://www.epilepsysociety.org.uk/lecture-notes-0#.XHj_6YhKguU.

[26] [2019-03-01]. https://www.templehealth.org/services/neurosciences/patient-are/programs/epilepsy.

[27] Shah V, Golmohammadi M, Ziyabari S, et al. Optimizing channel selection for seizure detection[C]//Proceedings of 2017 IEEE Signal Processing in Medicine and Biology Symposium. Philadelphia, PA, USA: IEEE, 2017: 1–5.

[28] Ramantani G, Maillard L, Koessler L. Correlation of invasive EEG and scalp EEG[J]. *Seizure,* 2016, 41: 196–200.

[29] Khoa TQ, Ha VQ, Toi VV. Higuchi fractal properties of onset epilepsy electroencephalogram[J]. *Comput Math Methods Med,* 2012, 2012: 461426.

[30] Zhang YD, Yang SH, Liu Y, et al. Integration of 24 feature types to accurately detect and predict seizures using scalp EEG signals[J]. *Sensors,* 2018, 18(5): 1372.