Learning data quality in EEG

Rossi Cecilia
dept. of Information Engineering
University of Padova
Padova, Italy
cecilia.rossi.3@studenti.unipd.it

Canderle Filippo
dept. of Information Engineering
University of Padova
Padova, Italy
filippo.canderle@studenti.unipd.it

***Abstract*—Electroencephalography (EEG) is widely employed in medicine to understand brain activity. Nowadays, EEG signals are also widely used in combination with machine learning methods to automatically distinguish between healthy and pathological subjects.**
However, EEG signal acquisition is often disrupted by noise and artifacts, leading to inaccurate predictions in classification tasks. This study aims to implement two different pipelines involving two support vector machines (SVM) models to achieve a multiclass classification of the EEG signals into artificial, healthy, and pathological categories.
In the first pipeline, we initially distinguished between clean and artifactual signals, and subsequently classified the clean signals into healthy and pathological subcategories.
In the second pipeline, instead, we initially classified the signals into healthy and abnormal. Within the abnormal category, we used another SVM to further discriminate between pathological and artifactual signals.
We trained and evaluatee the efficiency of the two processing pipelines using the TUAB and TUAR datasets, from the Temple University open source EEG resources. Finally, we draw conclusions based on our findings.

***Index Terms*—EEG, SVM, Multimodal sensing**

I. BACKGROUND

Electroencephalography (EEG) is a non-invasive recording technique that measures the electrical activity of the brain using electrodes placed on the scalp. EEG offers numerous advantages for both scientific research and practical applications. Firstly, EEG provides high temporal resolution, allowing the analysis of changes in brain activity over time with precision by millisecond. This makes EEG ideal for studying cognitive tasks and rapidly occurring brain processes. Moreover, EEG is particularly sensitive to dynamic changes in brain activity, enabling accurate detection of transient variations. This makes it a valuable tool for examining attention, perception, language, and motor control.
Another significant advantage of EEG is its non-invasiveness, meaning it poses minimal risks to research participants. As a result, EEG is safe and can be used on healthy individuals as well as patients.
A common research area is the classification of brain patterns. By employing machine learning algorithms, patterns of brain activity recorded from EEG can be analyzed and classified to identify specific mental states or recognize particular events. For instance, EEG can be used to classify patients as healthy or pathological, as well as implementing models to accurately distinguish between specific disorders.
Regarding the machine learning techniques used with EEG, there are several commonly adopted approaches. Some of the most common machine learning models used in this field include support vector machines (SVM), random forests, decision trees, as well as some deep learning solutions such artificial neural networks (ANN), convolutional neural networks (CNN) and recurrent neural networks (RNN). These models can be adapted to analyze EEG data, extract relevant features and classify patterns of brain activity.
A common problem in EEG analysis is the presence of artifacts, which can be caused by factors such as improper electrode placement, environmental noise, or other interference. To mitigate the negative impact of such artifacts on machine learning models, it is possible to consider implementing a dedicated layer for the detection of abnormal EEG signals, aimed at filtering or correcting contaminated data before it is used for model training. This is normally carried out by expert neurologist who, by only relying on the visual inspection of data and their experience, have to discriminate between artifactual and normal paths. This process is time-consuming, repetitive and could be affected by personal interpretation factors and mistakes. For this reason, the implementation of accurate automatic detection methods able to identify the presence of abnormal patterns in the EEG path and, furthermore, to understand if those irregularities are related to physiological/clinical factors or interference, could be fundamental for future applications.

II. MATERIALS AND METHODS

A. Dataset

The TUAR and TUAB datasets are extensive collections of electroencephalography (EEG) recordings from patients with abnormal neurological conditions, sourced from the Temple University Hospital EEG Corpus. Specifically, the TUAR dataset consists of abnormal EEG recordings, that encompasses a range of neurological disorders such as epilepsy, stroke, brain tumors, and other pathologies that can affect brain activity.
While, the TUAB dataset is a subset comprising abnormal EEG recordings with various pathological conditions, including epilepsy and other conditions characterized by alterations in brain electrical activity.
These datasets are widely utilized in the scientific community

for the development and evaluation of EEG analysis algorithms, including those based on machine learning approaches, for classification, artifact detection, and diagnosis of neurological conditions.
For this study, a significantly reduced and pre-processed version of the datasets was used. The signals from the original datasets were segmented into two-second windows, as is commonly done with this type of data, resulting in segments of 500 samples, acquired using a sampling frequency of 250 Hz. Each dataset included EEG recordings from 200 subjects, with 21 channels for the TUAB dataset and 23 channels for the TUAR dataset. As both datasets shared 21 common channels, the two additional channels in the TUAR dataset were discarded, and the datasets were merged into a single final dataset.
Those datasets differed principaly on the label associated to them. In fact, in the TUAR dataset, more emphasis was given to the presence of artifacts, related to external sources of noise or physiological interferences, with respect to clear paths. Instead, in the TUAB dataset, the windows were labelled according to the healthy or pathological/abnormal appearance. The final dataset used for training and testing our SVM models had dimensions of 400 (subjects) x 500 (samples per subject) x 21 (channels).
This methodology allowed a focused analysis while ensuring consistency and compatibility between the datasets, enabling robust evaluation of the proposed pipelines.

B. Research Methods

We can divide our research work into three main phases: initial data preprocessing, implementation of the first SVM cascade, and subsequently the second SVM cascade.

a) Pre-processing: The raw data was loaded from two HDF5 format files. The two datasets, TUAB and TUAR, as explained in the ”Dataset” section, originally had dimensions of 400x500x21 and 400x500x23, respectively. To make the two datasets compatible, two channels were removed from the second dataset. Each sample in the two datasets has a corresponding label, that could be from one of the following:

- TUAR: clean (1) and artifactual (0)
- TUAB: normal (1) and abnormal (0)

Also, by quickly checking the data, we noticed that the patients IDs in the two datasets were different, so that there was not a correspondence between the patients in two collections, that could have led to unbalanced in the following phases of our analysis. Therefore, we can consider our total dataset as composed of two distinct sources of information, each with a binary label of 0-1. To make the two datasets communicate with each other, we considered the ”clean” class from TUAR to be equivalent to the ”normal” class from TUAB, assuming that all the clean signals are also normal (so not pathological) and viceversa.
With this modification, along with the fusion of the two original datasets, we transitioned to a labeling of our samples where we have three possible classes: artefactual, healthy, and pathological (abnormal).
Finally, to transition from a 3D dataset to a 2D dataset, we transformed the set from 400x500x21 to one with dimensions (400x21) x 500. This resulted in 8,400 samples of length 500, where we consider acquisitions from the same subject but different channels as completely distinct samples. This modification, empirically tested, will lead to improved performance in the implementation of the two pipelines. After the preprocessing, the final dataset was divided among the classes as follows:

- class 1 (Clean Normal): 4200
- class 2 (Artifact): 2100
- class 3 (Abnormal): 2100

With this splitting completed, the data were ready to be fed into the two pipelines.

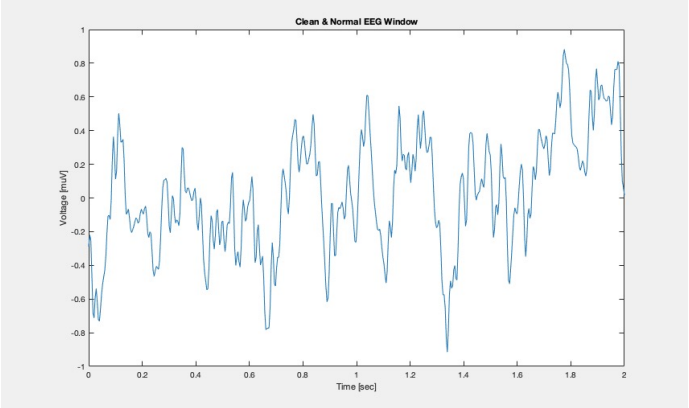

Fig. 1. An example of clean-normal sample

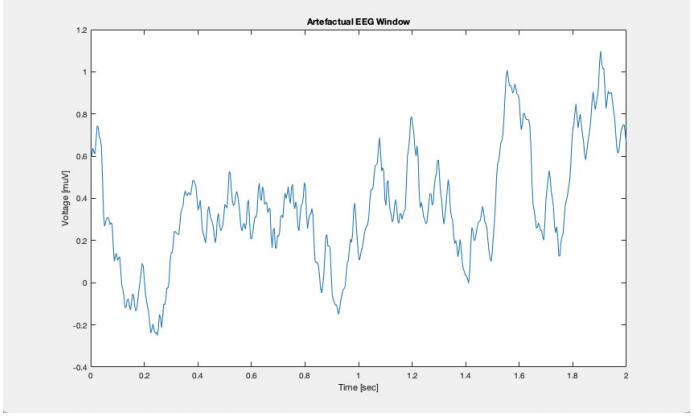

Fig. 2. An example of artifactual sample

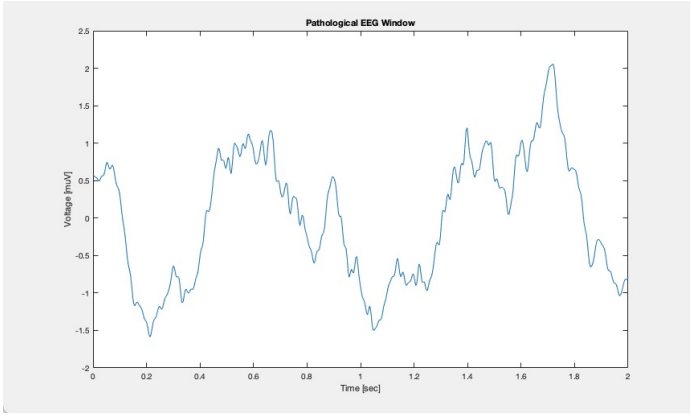

Fig. 3. An example of pathological sample

b) First Pipeline: In the first pipeline, our initial objective was to classify between healthy patients and patients exhibiting abnormal behavior, which could have been either an artifact or a pathological condition. In the second step, we then evaluated all the "abnormal" EEG samples and classified them as either pathological EEG patterns or signals distorted by the presence of artifacts.

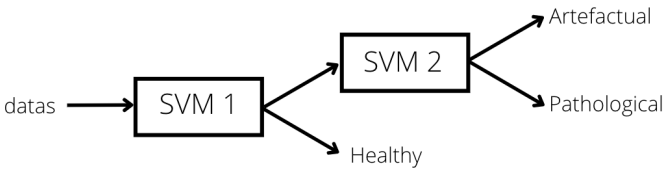

Fig. 4. First Pipeline Scheme

c) Second Pipeline: In the second pipeline, at the beginning our goal was to differentiate between EEG traces that contained artifacts and clean signals, which could have belonged to either the "healthy" or "pathological" categories. Then, a second SVM (Support Vector Machine), taking as inputs all the samples that were classified as clean in the first stage, provided a more accurate prediction, classifying them as either "healthy" or "pathological".

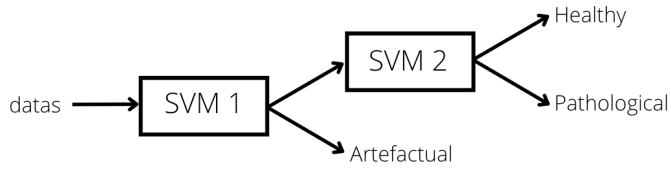

Fig. 5. Second Pipeline Scheme

Both pipelines were trained on 75% of the entire dataset, while keeping the remaining 25 % as the final test set, with the goal of determining which one performed better on the overall results.
In summary, the question we wanted to answer was:

- is it better to first focus on discriminating between a "healthy" trace and an "abnormal" trace, and then focus on what the abnormal trace is related to, *or*
- is it better to discarding as soon as possible the traces clearly related to noise or artifacts and later, discriminate the physiological,remaining ones between healthy and pathological?

III. RESULTS

At the end of both the pipeline, the Abnormal samples are labelled as 0, so they refer to the negative. The abnormal, pathological EEG traces are discriminated with respect to the Artifactual windows (First Pipeline) and the Clean&Healthy paths (Second Pipeline).

This finding underlines how the final aim of the work was to understand which of the two pipelines, in the end, enabled us to better identify the pathological EEG recordings.

The performance of the two classifiers was computed in terms of training and test error, as much as training and test accuracy. Furthermore, in both cases the number of true positive, true negative, false positive and false negative samples was computed on both the training and test set.
The accuracy equation used to evaluate performances of our models is given by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Before presenting you the results we would also want to remind you that the error of a classifier is given by:

$$\text{Error} = 1 - \text{Accuracy}$$

a) Pipeline 1: In the first pipeline, the results after training show an accuracy of **0.8597** during the training phase, which decreases to a value of **0.7010** in the test. The confusion matrices for both the training and testing phases, expressed in terms of False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN), are provided below:

TABLE I
CONFUSION MATRIX OF THE FIRST PIPELINE - TRAIN

	Negative	Positive
True	1525	**797**
False	272	**107**

TABLE II
CONFUSION MATRIX OF THE FIRST PIPELINE - TEST

	Negative	Positive
True	498	**250**
False	276	**43**

It's important to remember that, in this case, the negative samples are related to the pathological EEG paths, while the positive ones are associated to the artifactual traces, so to situations in which the signal is affected by external interferences and noise independent from its health condition. Both cases are related to the clear presence of abnormalities in the EEG

path and the good results obtained by applying this automatic detection method emphasise how, after discriminate between clear and unusual patterns, we later focus on the origin of the signal could be a successful approach to solve the problem.

In fact, on the train we have a very low error of type I (False Positives), but an higher error of type II (False Negative). Of course, this could be undesired, since it would always be better to realizing that an healthy subject if not affected by any disease by further analysis, that missing a diagnosis of a really ill patient.

b) Pipeline 2: In the second pipeline, the results after training show an accuracy of **0.8374** during the training phase, which decreases to a value of **0.7659**. It may seem that this second pipeline has poorer performance during the training phase compared to the first one. However, it demonstrates better generalization during the testing phase, as evidenced by the reported confusion matrices.

TABLE III
CONFUSION MATRIX OF THE SECOND PIPELINE - TRAIN

	Negative	Positive
True	4128	**743**
False	821	**125**

TABLE IV
CONFUSION MATRIX OF THE SECOND PIPELINE - TEST

	Negative	Positive
True	975	**219**
False	308	**58**

Also in this case, both in the training and test set, we can appreciate a lower error of type I (False Postive) with respect to the error of type II (False Negative), which could be not ideal.

We would like to specify that in the first classification using the second pipeline, there is a slight data imbalance, as logically the first class contains approximately twice as many samples as the other class. However, it was not possible to resolve this issue as the data remains unchanged, and we have observed that it does not result in over-fitting.

Finally, we would like to discuss the impact of pre-processing on the final results. Initially, we considered considering only 400 samples (the same number of the different patients), characterized by a number of features equal to 500x21. So every sample was described by a single vector obtained by concatenating all the 2seconds windows acquired by every EEG channel. But, we soon realized that by giving as an input this dataset to out SVM models, even if the computation took a much lower computational time, because of the high dimensionality it resulted in overfitting on the training data, with very bad performances on the test set. For this reason, we considered a different approach, in order to reduce the data dimensionality and increase the number of samples. This idea, as described in the pre-processing part of the methods, proved an actual improvement in the overall classifiers performance.

IV. DISCUSSION

The aim of this study was to investigate a real clinical scenario in which a physician examines an EEG trace in two phases, implicitly applying a two-step multiclassification process. In our automated models designed to simulate this pipeline, we observed that the best performances are achieved by the second pipeline. This work aims to demonstrate how it is more advantageous to firstly distinguish between artifact signals and non-artifact signals, and only then proceed to classify between healthy patients and pathological patients.

One challenge we encountered in this study was working with a limited amount of data. Due to this constraint, we opted to employ a traditional machine learning approach, such as SVMs, instead of utilizing deep learning solutions, which would likely have yielded better performance if we had a significantly larger dataset.

V. CODE

The implementation, done in matlab, can be found at this link:
https://github.com/FilippoCanderle/Learning-data-quality-in-EEG

REFERENCES

[1] Giulia Cisotto Member, IEEE, Alessio Zanga , Joanna Chlebus, Italo Zoppis, Sara Manzoni, and Urszula Markowska-Kaczma Comparison of Attention-based Deep Learning Models for EEG Classification