Definition of an Annotated DataSet

The first step of any AI problem is to set the inputs and outputs of the AI method. The inputs are the data/variables extracted from the raw data of the data base that the AI method will receive, while the output is the information/variables that the AI method predicts/issues after processing the input data and that should be annotated to constitute a ground truth (GT) for training and testing methods.

In our case, the raw data are 24 EEG recordings from the CHB-MIT-Epilepsy public data base. Each recording is from a single subject and lasts hours. Given that seizures are punctual episodes that last seconds, in order to feed data into the model and the model be able to classify such data as seizure and non-seizure, the continuous signal is segmented into small time window. So input data for the model is of shape E x T, where E is the number of EEG channels and T is the window length (number of seconds times sampling frequency). A simple approach to segment EEG signals is to slide a window over the continuous signals and crop small sequence data. Time windows of 1, 2, 4, 5 seconds are commonly used for epilepsy detection and recognition in the literature.

Regarding data annotation for the generation of GT, the CHB-MIT-Epilepsy recordings include information about the second a seizure starts and ends. These intervals should be annotated as seizures (class 1). For the remaining intervals between seizures, one should take into account that windows close to a seizure might present a transitional pattern different from the one found during seizures and normal intervals. It follows that they should be either discarded or treated differently (for instance as different classes). Figure 1 illustrates the partition of a signal into the different intervals: normal (non-seizure, preictal (interval previous to a seizure), seizure and potsictal (after a seizure).

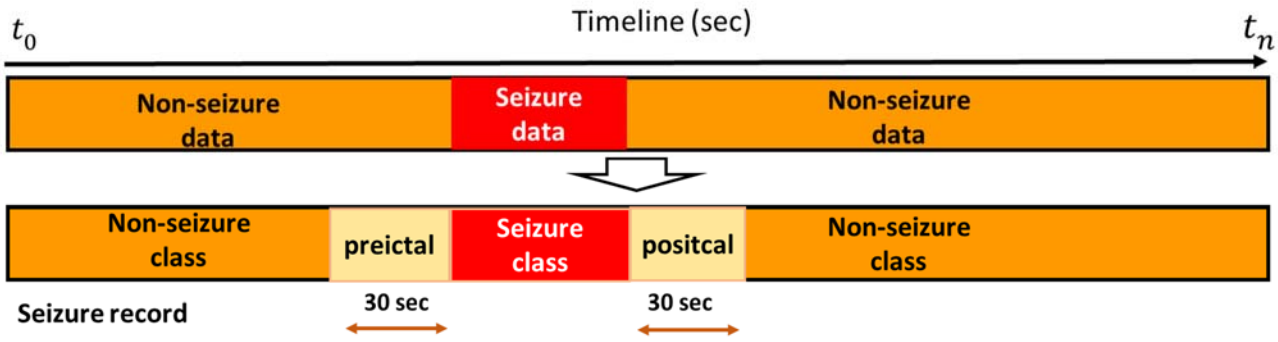

Figure 1. Data Annotation

In case of a classification problem, a strategy for managing data unbalancing during training is mandatory. Strategies can include data augmentation of the minority class, use of weighted losses, down-sampling strategies. Data augmentation to increase the number of seizures before to make the selection of the non-seizure class can be done for instance, considering window overlap only for this class as illustrated in Figure 2. Weighted losses give more importance to minority classes by rescaling the loss of each class inversely proportional to the class frequency. Finally, down-sampling strategies carefully select the EEG seizure part and randomly select EEG non-seizure segments to construct a balanced dataset [17].

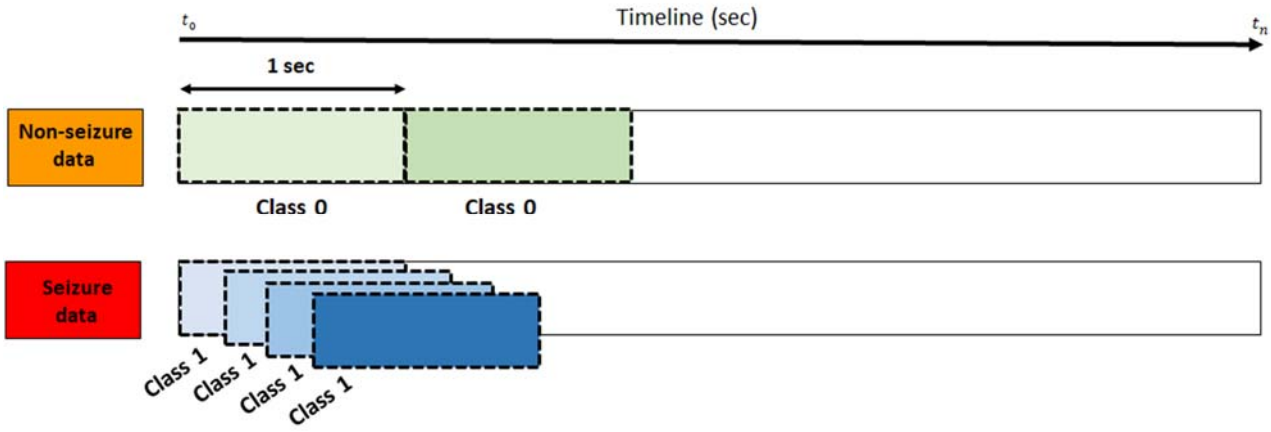

Figure 2. Data Augmentation using window overlap

A selection of recordings ensuring good compromise between training samples and computational time is recommended.

The **goal** of this step is to prepare input and output data for the AI methods: cut signals in temporal windows, filter data if needed, perform data augmentation and defining, for each window, a GT annotation.