## Experimental Design

A second step common to any AI problem is the definition of the experimental design for the training and testing of models. This experimental design should include: a strategy for splitting data, the metrics for the evaluation of models and the statistics for their comparison. Figure 1 shows the general pipeline for the experimental design.

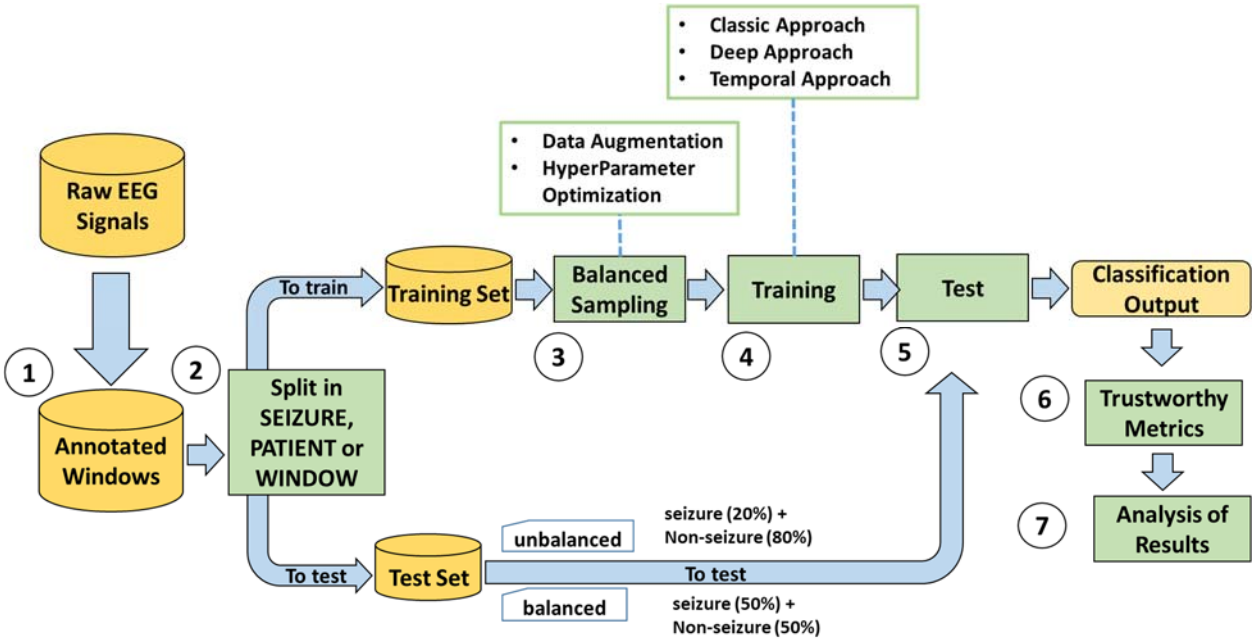


*Figure 1.Experimental Design General Flow Chart*

The splitting of data in training and test is directly associated to the personalization level of the model and the assessment of its generalization. Data is split according to what sampling unit we consider for our population in terms of the individuals that we will assess with our metrics to compute statistics. This sampling unit or individual does not need to be the same that we consider as input to the model. For instance, if the goal was to diagnose epilepsy, the input to our model would be temporal windows of recordings of a person, but the sample for evaluation of the goodness of detection of epilepsy would be the person, not the window.

In our case, data can be split according to 3 sampling levels: temporal windows seizure episode or patient. Regardless of the sampling level, test sets can be balanced and contain the same number of normal and seizure windows or unbalanced according to the real proportion between normal and seizure windows. The latter is a more realistic scenario, more suitable to assess the impact of false positives and trustworthiness of models.

A sampling at window level, would split all input windows in train and test using a kfold cross-validation approach. This split provides the lowest assessment of generalization capabilities for seizure detection because it does not guarantee that the test and train windows do not belong to the same seizure episode of a given subject. In other words, the test set is not independent from the training samples and, thus, it is very close to evaluating with the same set of training samples. Therefore, it sets an upper bound of top performance.

A sampling at seizure episode would evaluate whether a model can detect new seizures in a given set of patients. This design would evaluate models personalized for a given set of subjects. In this case, the k-fold would be at seizure episode, so that the test set would be given by windows belonging to a seizure and normal intervals not considered in the training set. Seizure selection can be done for each patient and normal episodes for testing can be selected from the normal recordings, thus leaving seizure recordings for training.

For a sampling at patient level, test windows should all belong to a subject excluded from the training. That is the kfold splitting is done at patient level. This is the most generalist splitting given that the model is a population one and the test assesses its performance in a new subject never seen during training.

Metrics should guarantee trustworthiness of models in the sense that they can detect any bias towards any of the classes (especially the majority one). Global accuracy should be discarded in case of unbalanced test sets. More suitable metrics are confusion matrices, precision, recall and metrics related to ROC curves (like AUC or F1-score).

Finally, statistics could consider several levels of aggregation of data in order to analyse the sources of variability: aggregation at patient level to discard outliers and aggregation at population level.

The **goal** is to define a complete experimental design and test it in a simple model based on classic descriptors of EEG seizures and a classic classifier (like SVM). This model also sets a baseline classic approach to be compared to deep approaches.