Framework to select the top ML classifier for robust seizure detection and prediction: A comparison-based study using multiple preictal time and feature sets

Rajlakshmi Borthakur*, Amit Kumar, Upendra Kumar Jena, Dibya Jyoti Borthakur
TerraBlue XT, email*: raji@teblux.com, www.teblux.com
Submitted: 28 April, 2021

Abstract

We present a machine learning framework aimed at increasing performance seizure detection systems. Despite the diversity of ML-based available, the methodology to select an optimal classifier or ensemble model for seizure research is not commonly known. Our study is aimed at bridging this gap by showing use of a statistically guided machine learning framework that has delivered reliable on-field performance.

We empirically examined performances of 11 different classification algorithms, with additional 13 variants, executing 300 machine learning models against 6 fixed preictal windows (ranging from 5 minutes to 60 minutes) and 3 different sets of features as part of our study. Using a base of 55 subjects, of which we used 43 subjects to train and validate multiple ensemble models and classifiers, we achieved up to 98.3% F1 score, with 94.8% chance level performance on our data, and a False Positive Rate (FPR) of 0.0156/hr. We applied the top ensembles and classifier variants on 12 completely new subjects, unseen by the models to test their on-field efficacy. One of our models detected all 12 seizures on the unseen data and predicted 11 of them before electrographic onset. Average prediction time of models across all subjects was 31.56 minutes. At an individual level the earliest prediction was recorded at 176 minutes, which is the highest known seizure prediction time recorded in literature. Our analysis showed that a 30 minute window could be the most ideal preictal window for training seizure models.

The significance of our work is that we are reporting the highest F1 score of 98.3% on seizure data so far recorded with sample size of higher than 10. We are able to show high performance of seizure models using wearables data, detecting seizures at par with neurologist verified EEG onset time, and predicting seizures up to 176 minutes before onset, which is significantly higher than previously reported scores. Our work pays important foundations for using wearables for actual monitoring of PWEs in out-of-hospital settings, which might be an important step in ensuring their safety and quality of life.

Keywords: seizure detection, seizure prediction, epilepsy wearables, machine learning for epilepsy

1 Introduction

Epilepsy is one of the most common neurological disorders, affecting an estimated 50 million people worldwide.1 Uncertainty about when a seizure will strike is the most debilitating aspect of living with epilepsy.2 Seizure frequency3, seizure severity, stigma, fear, comorbidities like the presence of cognitive or psychiatric disorders adversely affects the quality of life of person’s with epilepsy (PWE)4 and their families.5 From a clinician’s perspective, proper detection and prediction of seizures would enhance efficacy of treatment, help optimize drug dosing and minimize drug side effects.6 The need to enhance the quality of lives of PWEs, and to reduce avoidable deaths7 have led to growing interest in predictive, ambulatory, wearable and machine learning (ML)/artificial intelligence (AI)-based epilepsy management technologies.

Various ML and data mining algorithms that perform regression, clustering, visualization and classification of seizure data have been described in the context of seizures. Detecting seizures is mostly a classification problem as the goal is to distinguish an ‘ictal’ class (group) from preictal and non-seizure classes, or to identify the time at which a class showed ictal properties.
Though substantial work has been done in detection of seizures using machine learning, there hasn’t been adequate comparisons of multiple classifiers, preictal time windows and nature of datasets, nor discussions on their interactions and validity of scoring methods. A review of the prominent literature on epilepsy gave the following indications:

- The most common ML classifiers reported is literature are Random Forest, Support Vector Machines (SVM), Decision Tree, Extreme Learning Machine (ELM), back-propagation neural network (BPNN), Gaussian Mixture Model (GMM), Multi-Layer Perceptron (MLP), Linear Discriminant Analysis (LDA), AdaBoost and so on. While a range of algorithms have been used by researchers, it is generally not made clear as to why a particular classifier or combination of classifiers is preferable than the rest. The choice of classifier, especially for medical datasets should ideally depend on its utility in solving the research problem.

- The time preceding a seizure is referred to as the preictal state. For some seizures, the epileptogenic transient changes develop very close to the onset, while for others it appears several minutes to hours prior to onset. In ML-based seizure studies, arbitrary, fixed periods of time before seizure onset is denoted as the preictal period, which has been reported from 2 minutes to 240 minutes. The preictal class annotation on seizure datasets should ideally be done based on statistical guidance, instead of arbitrary reasons as this class holds important evolutionary clues about an incoming seizure.

- The reporting of performance of classifiers is often confusing. Several of the seizure models have reported their results in terms of accuracy which is considered an unreliable measure, especially in unbalanced datasets. In most of the epilepsy or even healthcare-related datasets, it is common for the adverse event to belong to a minority class, while the non-event related data forms the majority class. In such scenarios, the accuracy measure tends to make errors in predicting the right results. Only a handful of the seizure-related ML algorithms have reported their findings using other, more stronger evaluation metric for imbalanced datasets.

Our work brings together classifier and preictal time selection with phase-wise performance metrics to show how proper interactions of these elements can lead to consistently reliable classifier performance.

1.1 Contribution

The main contributions of our paper are:

- Providing a step-by-step process to identify the most appropriate classifiers, preictal time and feature sets for the study of seizures, using a methodology that could be used across domains and devices.

- Showing extensive use of statistical methods to select and eliminate classifiers in a phase-wise manner across the research timeline.

- Showing that it is possible to achieve not only high subject-specific scores but also a consistently high overall seizure detection scores using an optimal combination of classifiers, preictal time and feature sets.

- Providing on-field evidence of results from multiple classifiers and ensembles, with the ability to predict seizure up to 176 minutes prior to EEG onset.

To the best of our knowledge, we are the first to report the best F1 performance on seizure data and predict seizure up to 176 minutes in advance.
2 Related work

A recent review of seizure detection with machine learning using EEG data\(^{47}\) shows classifier performance ranges from 56.23% to 86% in terms of the F measure and up to 100% in accuracy, sensitivity and specificity.\(^{48\ 49\ 50\ 51\ 52\ 53\ 54}\) Seizure prediction windows have been reported between 1 sec\(^{55}\) to up to 2 hours.\(^{56}\) Cardiac monitoring of seizures have been reported by using electrocardiogram (ECG or EKG), Holter monitors and photoplethysmography (PPG). Using ECG as the base for their seizure study, Malarvili and Meshah reported a sensitivity of 85.7% and a specificity of 84.6%.\(^{57}\) Varon et al. achieved a positive predictive value (PPV) of 86.2% and a sensitivity of 100% for partial seizures, and a PPV of 84.3% and a sensitivity of 93.1% for generalized seizures,\(^{58}\) while Fujiwara described a sensitivity of 91% and false positive rate of 0.7/h− using a multivariate statistical process for epileptic seizure prediction.\(^{59}\)

Wearable devices using Accelerometry (ACM),\(^{60\ 61\ 62}\) Electrocardiogram (EKG),\(^{63\ 64}\) Surface Electromyography (sEMG),\(^{65\ 66}\) Electrodermal Activity (EDA),\(^{67\ 68}\) skin temperature,\(^{69}\) and respiration analysis\(^{70}\) have been used to detect seizure activities. Publications on seizure monitoring with wearable devices have reported higher accuracy, but limited precision and F1 scores in seizure detection.\(^{71\ 72\ 73\ 74}\) Kurada et al. showed that some models had a sensitivity of 100%-94% in detecting seizures, but with precision of 0.02%-35%, and F1 score of 0.03%-51%.\(^{75}\) The low performance of ML/AI algorithms in seizure prediction and detection using wearables depend on various factors, including data acquisition techniques, usability of the equipment, and the quality of data used for experiments.\(^{76}\)

3 Framework for seizure detection and prediction

The larger context of our exploratory analysis was to identify factors that impacted performance of ML algorithms in the context of wearables in seizure research. The specific objective of the current study was to determine which classifier, preictal time and feature set combination provided best seizure detection performance at par with EEG-identified onset time. Additionally, we investigated the impact of predetermined preictal windows on classifier performance, to understand which preictal window gave the best result.

We chose ML-based algorithms over deep learning (DL) for our work as DLs are often difficult to interpret. Interpretability, consistency and scalability were the main criteria for our classifier selection.

3.1 Subject enrolment

We enrolled subjects admitted to the video-EEG monitoring unit at the National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore, India, after receiving approval from the Institutional Ethics Committee. The study was conducted between 2017 and 2020, on subjects aged 15-65 years. All subjects signed informed consent form prior to being enrolled. We recorded data of 106 subjects with 202 total seizures. Out of 202, 53 were pseudo seizures, 34 seizures had artefacts and distortions, and 60 seizures were in close proximity with another seizure (within 2 hours of each other). We considered 55 seizures for our work, where 43 seizures were used for model building and 12 seizures used for unseen, unbiased testing of the trained models.

3.2 Data acquisition

We focussed on the activities of the Autonomic Nervous System (ANS), as it provides reliable clues about seizures.\(^{77\ 78\ 79\ 80\ 81\ 82}\) We used a multimodal-sensing methodology focusing on Blood Volume Pulse (BVP) and Electrodermal Activity (EDA) to track the activities of the ANS. We extracted data using devices from BioSignal Plux and Shimmer Sensing, along with a wearable prototype designed as per our proprietary specifications. Data was sampled at 100 Hz for both EDA and BVP, which were were extracted from the distal and proximal phalanges of the non-dominant hand. The subjects were also monitored simultaneously with video EEG (vEEG).
3.3 Data pre-processing
We analyzed continuous, non-overlapping per 10 millisecond segments of data. During pre-processing, data was converted from raw form to their respective units, filtering and smoothening was done, along with timestamp insertion and artefact detection and removal. Both EDA and BVP data were filtered using a low band pass, 2\textsuperscript{nd} order Butterworth filter.

3.4 Feature engineering
We extracted various features for our experiments, but used only the relevant ones based on Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA). In the final form, we extracted 3 sets of features from the BVP and the EDA signals for our study. Set A consisted of 39 basic features as described in literature,\textsuperscript{83, 84} Set B consisted of 66 features that are a combination of Set A with a few statistical features. Set B has been further described as supplementary material. Set C contained 116 features derived by combining Set A and B features, along with relevant features from Tsfresh.\textsuperscript{85}

3.5 Dataset creation with preictal periods
We used 18 datasets, with each made up of a combination of a feature set and a preictal time. The preictal periods used for dataset creation were 5, 10, 15, 30, 45 and 60 minutes.

During label annotation, the period prior to the recorded electrographic onset of a seizure was labelled as ‘Preictal.’ The data prior to the designated preictal period was labelled ‘Non Seizure’, while the seizure onset time and 2 minute after it was labelled as ‘Ictal’. Each dataset was trained and tested using the 11 classifiers and the 13 variations, providing unique performance scores in 3 phases of the study.

3.6 Experimental phases
To arrive at the best combination of classifiers, preictal time and feature set, we performed our work in 3 phases:

- In phase 1, we experimented with 11 different classifiers, using a randomly selected feature set to pick the top 3 classifiers. We compared their performance against all 6 preictal time periods. The goal was to rank classifiers based on the consistency of performance across preictal windows and to reduce computational load in further work by dismissing poor performers right at the beginning.

- In phase 2, we used only the top 3 classifiers and 13 variants to evaluate performance over 3 different feature sets and 6 preictal time periods. This phase enabled us to confirm the best classifiers and pairs and find the most suitable feature set, preictal time and combinations.

- In phase 3, we used the findings from phase 2 to create multiple ensemble models, and tested their performance on the best suited dataset as well as on data of 12 individual subjects, unseen by the models.

3.6.1 Phase 1 – Identifying top classifiers for further exploration
The best known ML classification algorithms use probabilistic methods, rule-based learners, linear models such as neural networks and support vector machines, decision trees and instance-based learners.\textsuperscript{86} As a first step, we selected 11 such popular classifiers and used them on 6 datasets, each made up by combining Set B and a specific preictal time, resulting in 66 model executions.

The classifiers that we chose are Logistic Regression (LR), Stochastic Gradient Descent (SGD), k-Nearest Neighbour (KNN), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), AdaBoost (ADB), Extreme Gradient Boost (XGB), Light GBM (LGBM) and Artificial Neural Network (ANN).
3.6.1.1 Performance evaluation of phase 1

For all the 3 phases, we measured performance of the models using 2 rules:

- a classifier’s seizure detection performance was measured by its ability to detect the ictal class at par with a neurologist validated EEG onset time of seizure, and
- predictive performance was measured through a rule denoting that all predicted ictal segments must be preceded by the preictal class for a minimum of 5 minutes for the prediction criteria to be met. Though it is understandable that the preictal period can be as low as 1 minute to a few seconds in real scenarios, we formulated the rule to bring about rigidity in the prediction mechanism to reduce false alarms.

For performance evaluation, we split the datasets in a ratio of 80:20 training and validation sets. We used 5-fold cross validation, where the training data was randomly divided into 5 equal subsets. For every iteration, a single subset was used as the validation set, while the remaining 4 subsets were used for training, with the process repeating 5 times till all the sets got trained. This method is superior to Leave-One-Out Cross-Validation (LOOCV), where a model is iteratively fitted on the whole training set. LOOCV is also prone to high variance and over fitting. 5-fold cross validation addresses the drawbacks of LOOCV by using the k-th portion of data to validate, enabling the model to train on a larger number of observations.

We used optimization techniques like Grid Search to find the best combination of hyper parameters in an individual classifier. Post tuning, we repeated the 5-fold cross validation with the recommended hyper parameters and obtained multiple performance metrics for the training and validation sets.

3.6.1.2 Statistical evaluation of phase 1

We performed statistical evaluation of results in each phase, using F1 score as our main performance metric, though we did calculate other statistics. We used MCC to evaluate better than chance performance as MCC is considered a more reliable performance metrics than even the F measure.\(^{97}\) et al. In the final phase, we additionally applied the Youden’s index, positive and negative likelihood ratios, discriminant power, Area Under Curve (AUC) and False Positive Rate (FPR).

The formulas used are as follows, where TP refers to True Positive, TN is True Negative, FP is False Positive and FN is False Negative:

- F1 score: F1 is a weighted average of precision and recall. It can be calculated as \(\frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}\)
- Precision: Also called the positive predictive value (PPV), it represents the fraction of actual seizures among those instances that have been classified as seizures. The formula is \(\frac{\text{TP}}{\text{TP} + \text{FP}}\)
- Recall (Sensitivity): This is the proportion of actual seizures that are correctly classified as such with formula \(\frac{\text{TP}}{\text{TP} + \text{FN}}\)
- Matthews correlation coefficient (MCC): MCC is a method to evaluate chance level performance. It produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset. Formula is \(\text{MCC} = \frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}^{0.5}\)
- Youden’s Index: Youden’s index \(\gamma\) evaluates the algorithm’s ability to avoid failure. It equally weights the algorithm’s performance on positive and negative samples, with the formula \(\gamma = \text{sensitivity} - (1 - \text{specificity})\)
• Likelihood: The relation between the likelihood of two algorithms A and B determines which algorithm is preferable and in which situation. Formula used are $\rho^+ = \text{sensitivity} / 1 - \text{specificity}$ and $\rho^- = 1 - \text{sensitivity} / \text{specificity}$

• Discriminant Power (DP): DP evaluates how well an algorithm distinguishes between positive and negative examples. Formula used is $\text{DP} = \sqrt{3/\pi} \times (\log X + \log Y)$

• Area Under Curve (AUV): The AUC corresponds to the arithmetic mean of sensitivity and specificity values of each class. Formula is $\text{(TNR+TPR)}/2$

• False Positive Rate (FPR): Refers to the conditional probability of a positive test result given an event that was not present. Formula is $1-\text{TNR}$

### 3.6.1.2.1 Performance scores in phase 1

The following table shows the individual F1 scores of each of the 11 classifiers on the 6 unique datasets. The highest score of 97.6% F1 came from LGB on the 30 minute dataset.

**Table 1: F1 scores of 11 classifiers across 6 datasets**

| # | Dataset* | LR | SGD | KNN | DT | NB | SVM | RF | ADB | XGB | LGBM | ANN |
|---|----------|----|-----|-----|----|----|-----|----|-----|-----|------|-----|
| 1 | SETB_5m  | 0.447 | 0.381 | 0.943 | 0.932 | 0.386 | 0.898 | 0.957 | 0.948 | 0.956 | 0.948 | 0.933 |
| 2 | SETB_10m | 0.453 | 0.356 | 0.937 | 0.938 | 0.422 | 0.857 | 0.970 | 0.940 | 0.970 | 0.949 | 0.945 |
| 3 | SETB_15m | 0.477 | 0.415 | 0.937 | 0.936 | 0.395 | 0.878 | 0.971 | 0.939 | 0.969 | 0.9624 | 0.948 |
| 4 | SETB_30m | 0.477 | 0.415 | 0.937 | 0.923 | 0.395 | 0.878 | 0.967 | 0.946 | 0.969 | 0.9768 | 0.948 |
| 5 | SETB_45m | 0.445 | 0.440 | 0.495 | 0.497 | 0.331 | 0.532 | 0.576 | 0.528 | 0.543 | 0.5592 | 0.574 |
| 6 | SETB_60m | 0.505 | 0.483 | 0.455 | 0.443 | 0.271 | 0.498 | 0.518 | 0.484 | 0.527 | 0.5354 | 0.553 |

*The 6 datasets are made up by combining feature set B features with 6 distinct preictal time period.

The below table shows the range and mean scores of each of the 11 classifiers, averaged across the datasets. As can be seen, LGB, RF and XGB scored consistently better than the rest on the 6 datasets.

**Table 2: Range and mean performance scores of 11 classifiers**

| # | Classifier Variants | F1 (M) | F1 (M) | R (M) | R (M) | R (M) | R (M) | P Range | P Range | Acc Range | Acc (M) | MCC Range | MCC (M) |
|---|----------------------|--------|--------|-------|-------|-------|-------|---------|---------|-----------|---------|-----------|---------|
| 1 | LGB                  | 0.97-  | 0.97-  | 0.810 | 0.99-  | 0.837 | 0.97-  | 0.918   | 0.93-     | 0.823     |          |         |
| 2 | RF                   | 0.97-  | 0.96-  | 0.819 | 0.98-  | 0.837 | 0.97-  | 0.920   | 0.92-     | 0.827     |          |         |
| 3 | XGB                  | 0.97-  | 0.96-  | 0.817 | 0.98-  | 0.831 | 0.96-  | 0.914   | 0.92-     | 0.817     |          |         |
| 4 | ADB                  | 0.94-  | 0.92-  | 0.781 | 0.99-  | 0.819 | 0.97-  | 0.892   | 0.912-    | 0.767     |          |         |
| 5 | ANN                  | 0.94-  | 0.93-  | 0.807 | 0.96-  | 0.830 | 0.97-  | 0.892   | 0.896-    | 0.767     |          |         |
| 6 | KNN                  | 0.94-  | 0.91-  | 0.769 | 0.98-  | 0.810 | 0.97-  | 0.874   | 0.898-    | 0.738     |          |         |
| 7 | DT                   | 0.93-  | 0.95-  | 0.785 | 0.93-  | 0.774 | 0.94-  | 0.870   | 0.897-    | 0.726     |          |         |
| 8 | SVM                  | 0.89-  | 0.88-  | 0.744 | 0.96-  | 0.780 | 0.96-  | 0.869   | 0.824-    | 0.706     |          |         |
| 9 | LR                   | 0.50-  | 0.49-  | 0.460 | 0.72-  | 0.587 | 0.86-  | 0.729   | 0.57-     | 0.273     |          |         |
| 10| SGD                  | 0.48-  | 0.63-  | 0.406 | 0.67-  | 0.593 | 0.86-  | 0.699   | 0.48-     | 0.198     |          |         |
| 11| NB                   | 0.42-  | 0.35-  | 0.412 | 0.46-  | 0.434 | 0.70-  | 0.604   | 0.201-    | 0.150     |          |         |

*Name is Classifier name, Range is minimum and maximum limits, M refers to the mean, R is recall, P is Precision, Acc is Accuracy and MCC is Mathews correlation coefficient.

We performed statistical analysis on the F1 scores reported above using the benchmarks for multiple machine learning comparisons by Demsar.

---

6
The variations we used are see the performance of the models in both balanced and unbalanced versions of the datasets. In class imbalance, the algorithm

Preictal
Of the clas

In our dataset

The decision to use
variation made up of a combination of individual scores alone, as detailed statistical investigation may show a completely different ranking order.

3.6.2 Phase 2 – Top classifier, feature set, preictal time and combination selection

In the next phase, we further evaluated the performance of RF, XGB and LGB using 18 datasets made up of a combination of 3 sets of features and 6 preictal time periods. We used a total of 13 variations of the selected classifiers.

The decision to use different variations of the classifiers were made to address class imbalance in our datasets. Depending on the feature set and the preictal window chosen, the distribution of the classes would vary in each dataset, with data in the Ictal class ranging from 01%-4%, Preictal class between 09%-44% and Non-Seizure class between 53% - 86%. When there is class imbalance, the algorithms become more biased towards predicting the majority class, as they do not have enough data to learn the patterns in the minority class. Hence, it is essential to see the performance of the models in both balanced and unbalanced versions of the datasets. The variations we used are as follows:

- Synthetic Minority Oversampling Technique (SMOTE) blends under-sampling of the majority class with a special form of over-sampling of the minority class. The minority class is over-sampled by creating “synthetic” examples, by operating in the ‘feature space’ rather than the ‘data space’.

Based on the normality and the homogeneity tests, we selected the Friedman’s test for determining the confidence intervals of the central tendency. After rejecting the null hypothesis of the Friedman’s test (p=0.05), we inferred differences between populations are significant, if the difference of the mean rank is greater than the critical distance (CD) = 4.904 of the Nemenyi test. The below table shows the mean rank, the median as central tendency, the median absolute deviation (MAD) from the median as measure for the variance, the confidence interval (CI) of the median, the effect size in comparison to the highest ranking approach and the magnitude.

### Table 4: Results of statistical evaluation in phase 1

| # | Classifier | Mean Rank | Median | MAD | CI (lower) | CI (Upper) | Effect | Magnitude |
|---|-----------|-----------|--------|-----|------------|------------|--------|-----------|
| 1 | RF        | 1.91      | 0.96   | 0.01| 0.51       | 0.97       | 0      | Negligible|
| 2 | XGB       | 2.41      | 0.96   | 0.01| 0.52       | 0.97       | -0.03  | Negligible|
| 3 | LGB       | 2.66      | 0.94   | 0.03| 0.53       | 0.97       | 0.57   | Medium    |
| 4 | ANN       | 3.50      | 0.93   | 0.01| 0.55       | 0.94       | 1.80   | Large     |
| 5 | ADB       | 5.16      | 0.94   | 0.01| 0.48       | 0.94       | 1.84   | Large     |
| 6 | KNN       | 6.83      | 0.93   | 0.00| 0.45       | 0.94       | 2.61   | Large     |
| 7 | SVM       | 7.16      | 0.86   | 0.03| 0.49       | 0.89       | 4.08   | Large     |
| 8 | DT        | 7.33      | 0.92   | 0.01| 0.44       | 0.93       | 2.55   | Large     |
| 9 | LR        | 8.33      | 0.46   | 0.02| 0.44       | 0.50       | 26.89  | Large     |
| 10| SGD       | 10.00     | 0.41   | 0.04| 0.35       | 0.48       | 16.90  | Large     |
| 11| NB        | 10.66     | 0.39   | 0.02| 0.27       | 0.42       | 27.23  | Large     |

As shown in Table 4, though LGB was the best scorer with 97.6% F1, RF showed better ability to generate consistent scores across the 6 datasets. The above analysis shows RF (Mean Rank of 1.91), XGB (Mean Rank of 2.41) and LGB (Mean Rank of 2.66) to be the top ranked performers and thus qualified for further investigation. The worst performing classifiers are NB, followed by SGD and LR. The data presented above shows us the importance of not deciding on classifiers based on individual scores alone, as detailed statistical investigation may show a completely different ranking order.

### Table 3: Statistical evaluation of 11 classifiers

| # | Population | Sample | P value alpha | Normality | Homoscedastic | Omnibus | Post hoc |
|---|------------|--------|---------------|-----------|---------------|---------|----------|
| 1 | 11 classifiers | 6 paired samples | L32E-07 | False (Shapiro-Wilk, with Bonferroni correction) | True | Friedman | Nemenyi |
- 'balanced': During hyper-parameter tuning we specify the class_weight as 'class_weight = balanced'. This mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n_samples / (n_classes * np.bincount(y)).
- 'balanced sub-sample': During hyper-parameter tuning we specify the class_weight as 'class_weight = balanced_subsample.' This mode is the same as 'balanced' except that weights are computed based on the bootstrap sample for every tree grown.
- Class weights: By assigning higher weights to the minority class, we can influence the training outcome by heavily penalizing the misclassification made on the minority class. This, thus, forces the algorithms to be less biased towards the majority class. We experimented with weights between 4-30% for the Ictal class and determined optimal weights using a Python script.

3.6.2.1 Performance evaluation of phase 2

For performance evaluation, the datasets were again split into 80:20 ratio of training and validation sets. We performed 5-fold cross validations, tuned hyper parameters and repeated 5-fold cross validations tuning to find the best scores. The table below shows the F1 scores of each of the different variants of the classifiers.

| Table 5: F1 scores of each classifier against every dataset |
|-------------|-------|------|------|------|------|------|------|------|------|
|            | #     | Dataset | RF   | RF_B | RF_BS | RF_S | RF_W | XGB  | XGB_B | XGB_S |
| 1          | 1     | SETA_5m | 0.943 | 0.943 | 0.943 | 0.934 | 0.958 | 0.953 | 0.938 | 0.959 |
| 2          | 2     | SETA_10m| 0.953 | 0.943 | 0.944 | 0.941 | 0.964 | 0.967 | 0.922 | 0.961 |
| 3          | 3     | SETA_15m| 0.953 | 0.935 | 0.935 | 0.960 | 0.957 | 0.962 | 0.940 | 0.970 |
| 4          | 4     | SETA_30m| 0.953 | 0.956 | 0.956 | 0.940 | 0.975 | 0.971 | 0.932 | 0.973 |
| 5          | 5     | SETA_45m| 0.953 | 0.542 | 0.542 | 0.535 | 0.535 | 0.535 | 0.535 | 0.535 |
| 6          | 6     | SETA_60m| 0.953 | 0.516 | 0.516 | 0.519 | 0.526 | 0.533 | 0.544 | 0.544 |
| 7          | 7     | SETB_5m | 0.943 | 0.949 | 0.949 | 0.939 | 0.959 | 0.950 | 0.958 | 0.953 |
| 8          | 8     | SETB_10m| 0.950 | 0.947 | 0.947 | 0.941 | 0.971 | 0.950 | 0.965 | 0.971 |
| 9          | 9     | SETB_15m| 0.965 | 0.940 | 0.945 | 0.946 | 0.970 | 0.960 | 0.945 | 0.970 |
| 10         | 10    | SETB_30m| 0.953 | 0.971 | 0.962 | 0.956 | 0.981 | 0.975 | 0.954 | 0.979 |
| 11         | 11    | SETB_45m| 0.561 | 0.563 | 0.556 | 0.559 | 0.573 | 0.566 | 0.564 | 0.573 |
| 12         | 12    | SETB_60m| 0.511 | 0.529 | 0.515 | 0.522 | 0.541 | 0.540 | 0.534 | 0.544 |
| 13         | 13    | SETC_5m | 0.904 | 0.900 | 0.909 | 0.906 | 0.909 | 0.909 | 0.890 | 0.902 |
| 14         | 14    | SETC_10m| 0.914 | 0.907 | 0.908 | 0.912 | 0.920 | 0.912 | 0.919 | 0.920 |
| 15         | 15    | SETC_15m| 0.912 | 0.909 | 0.907 | 0.910 | 0.918 | 0.919 | 0.907 | 0.923 |
| 16         | 16    | SETC_30m| 0.903 | 0.913 | 0.917 | 0.919 | 0.920 | 0.921 | 0.919 | 0.921 |
| 17         | 17    | SETC_45m| 0.575 | 0.579 | 0.580 | 0.633 | 0.591 | 0.585 | 0.543 | 0.583 |
| 18         | 18    | SETC_60m| 0.524 | 0.535 | 0.535 | 0.604 | 0.541 | 0.555 | 0.564 | 0.563 |

RF, XGB and LGB are the unbalanced versions of RF, XGB and LGB, 'B' is balanced, 'BS' is balanced sub-sample and 'S' is SMOTE. XGB and LGB have built in sub-sampling.

In the above table, the highest individual F1 score was 98.1% from RF(weighted), followed by 97.9% from XGB(weighted) and LGB (unbalanced) and LGB (weighted) at 97.7% F1. All these performances were recorded on the feature Set B and 30 minute preictal dataset (row 10). The lowest F1 score was given by RF (unbalanced) on Set A, 60 minute dataset (row 6) with 50.1%. The table highlights how individual F1 scores can be quite dissimilar between variations of the same classifier, and unless probed it is not evident. It also shows the importance of evaluating balanced and unbalanced versions of the datasets, as depending on the classifier variant used, different results could be generated. The following table shows the ranges and mean values of F1, recall, precision, accuracy and MCC for the 13 classifier variants on the 18 datasets.

| Table 6: Range and mean performance scores of phase 2 classifier variants |
|-------------|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| #           | Classifier Variants | RF     | F1 (M) | F1 (N) | R     | R (M) | R (N) | P     | P (M) | P (N) | Acc   | Acc (M) | Acc (N) | MCC   | MCC (M) | MCC (N) |
| 1           | XGB_W          | 0.98-54 | 0.803  | 0.812  | 0.98-54 | 0.98-54 | 0.869  | 0.98-54 | 0.926  | 0.94-63 | 0.83  |
| 2           | RF_W           | 0.98-54 | 0.803  | 0.812  | 0.98-54 | 0.98-54 | 0.869  | 0.98-54 | 0.926  | 0.94-63 | 0.83  |
| 3           | LGBM_W         | 0.97-54 | 0.803  | 0.812  | 0.98-54 | 0.98-54 | 0.869  | 0.98-54 | 0.926  | 0.94-63 | 0.83  |
From Table 8, it is clear that Table 8, it is clear that LGB (balanced) has got the best mean rank, and hence can be considered better than the rest of the classifier variants when the 5 performance metrics (F1, Precision, Recall, Accuracy, and Matthews Coefficient) are evaluated. It can be seen that though RF (Weighted) with 98.1% F1 had a higher individual score, it came second when ranges and mean scores of the various performance metrics are considered.

Further analysis needs to be carried out to determine the top ranked classifier, feature set, and preictal time based on the data in Table 6.

3.6.2.2 Statistical evaluation of phase 2

3.6.2.2.1 Top classifiers for seizure data

Evaluation on the entire range of performance metrics revealed 3 different variations of LGB (balanced, weighted and SMOTE) to be the top 3 mean ranked classifiers.

**Table 7: Statistical evaluation of classifiers**

| # | Classifier Variants | F1 Range | F1 (M) | R Range | R (M) | P Range | P (M) | Acc Range | Acc (M) | MCC Range | MCC (M) |
|---|---------------------|----------|--------|---------|--------|---------|--------|-----------|----------|-----------|--------|
| 4 | LGB | 0.97-0.53 | 0.013 | 0.97-0.53 | 0.794 | 0.99-0.53 | 0.838 | 0.97-0.80 | 0.922 | 0.93-0.60 | 0.82 |
| 5 | LGB_B | 0.97-0.53 | 0.014 | 0.97-0.53 | 0.8 | 0.98-0.53 | 0.833 | 0.97-0.79 | 0.919 | 0.93-0.59 | 0.82 |
| 6 | XGB | 0.97-0.53 | 0.016 | 0.97-0.53 | 0.801 | 0.99-0.53 | 0.837 | 0.97-0.79 | 0.922 | 0.93-0.59 | 0.82 |
| 7 | XGB_B | 0.97-0.53 | 0.017 | 0.97-0.53 | 0.803 | 0.98-0.53 | 0.836 | 0.97-0.79 | 0.922 | 0.93-0.59 | 0.83 |
| 8 | RF | 0.97-0.51 | 0.004 | 0.95-0.51 | 0.782 | 0.99-0.52 | 0.835 | 0.97-0.78 | 0.914 | 0.95-0.54 | 0.80 |
| 9 | RF | 0.95-0.50 | 0.004 | 0.95-0.49 | 0.786 | 0.99-0.51 | 0.83 | 0.97-0.76 | 0.911 | 0.92-0.50 | 0.80 |
| 10 | XGB_S | 0.96-0.53 | 0.009 | 0.95-0.53 | 0.795 | 0.98-0.53 | 0.828 | 0.97-0.79 | 0.917 | 0.92-0.59 | 0.81 |
| 11 | LGB_S | 0.96-0.54 | 0.010 | 0.95-0.53 | 0.797 | 0.99-0.54 | 0.829 | 0.98-0.80 | 0.921 | 0.94-0.61 | 0.82 |
| 12 | RF_BS | 0.96-0.51 | 0.003 | 0.95-0.51 | 0.781 | 0.99-0.52 | 0.833 | 0.97-0.78 | 0.912 | 0.93-0.54 | 0.80 |
| 13 | RF_S | 0.96-0.51 | 0.010 | 0.95-0.51 | 0.792 | 0.97-0.52 | 0.833 | 0.97-0.77 | 0.911 | 0.92-0.54 | 0.80 |

*Range is minimum and maximum limits, M refers to the mean, R is recall, P is Precision, Acc is Accuracy and MCC is Matthews correlation coefficient.

Post rejecting the null hypothesis of the Friedman’s test, we found differences between populations are significant, if the difference of the mean rank is greater than the critical distance (CD) =0.5524 of the Nemenyi test.

**Table 8: Results of statistical evaluation of classifiers**

| # | Classifier | Mean Rank | Median | MAD | CI (lower) | CI (Upper) | Effect | Magnitude |
|---|------------|-----------|--------|-----|------------|------------|--------|-----------|
| 1 | LGB_B | 2.25 | 0.92 | 0.06 | 0.54 | 0.98 | 0.06 | Negligible |
| 2 | LGB_W | 3.44 | 0.92 | 0.06 | 0.54 | 0.97 | 0.06 | Negligible |
| 3 | LGB_S | 4.36 | 0.91 | 0.07 | 0.52 | 0.98 | 0.08 | Negligible |
| 4 | XGB_S | 5.36 | 0.92 | 0.07 | 0.53 | 0.97 | 0.06 | Negligible |
| 5 | RF_W | 6.22 | 0.92 | 0.06 | 0.53 | 0.97 | 0.09 | Negligible |
| 6 | LGB | 6.86 | 0.91 | 0.06 | 0.53 | 0.97 | 0.11 | Negligible |
| 7 | XGB_W | 7.11 | 0.91 | 0.04 | 0.54 | 0.96 | 0.11 | Negligible |
| 8 | XGB_B | 7.47 | 0.91 | 0.05 | 0.53 | 0.97 | 0.11 | Negligible |
| 9 | XGB | 8.58 | 0.91 | 0.04 | 0.53 | 0.96 | 0.11 | Negligible |
| 10 | RF_S | 9.30 | 0.91 | 0.04 | 0.51 | 0.96 | 0.18 | Negligible |
| 11 | RF | 9.58 | 0.91 | 0.06 | 0.50 | 0.96 | 0.19 | Negligible |
| 12 | RF_BS | 10.11 | 0.91 | 0.05 | 0.51 | 0.96 | 0.22 | Small |
| 13 | RF_B | 10.33 | 0.91 | 0.05 | 0.51 | 0.97 | 0.25 | Small |
Precision, Recall, Accuracy and MCC) are considered. It must be noted that in phase 2, we determined RF_W as the best individual scorer, XGB_W as the best scorer in terms of range of performance, while LGB_B is the best mean ranked performer. This fact would not have been known had additional step-wise calculations were not performed.

### 3.6.2.2.2 Best feature set selection

We analyzed the performance metrics of Phase 2 to derive the best performing feature set. The statistical analysis on sets was done as follows:

| Table 9: Statistical evaluation of feature sets |
|-----------------------------------------------|
| # | Population | Sample | P value alpha | Normality | Homoscedastic | Omnibus | Post hoc |
|---|------------|--------|---------------|-----------|---------------|---------|----------|
| 1 | 3 feature sets | 5 performance scores (F1, Recall, Precision, Accuracy and MCC) | 0.2465 | False (Shapiro-Wilk, with Bonferoni correction) | True | Friedman | Nemenyi |

Post rejecting the null hypothesis of the Friedman’s test, we found differences between populations are significant, if the difference of the mean rank is greater than the critical distance (CD) =0.5524 of the Nemenyi test.

| Table 10: Results of statistical evaluation of feature sets |
|-------------------------------------------------------------|
| # | Classifier | Mean Rank | Median | MAD | CI (lower) | CI (Upper) | Effect | Magnitude |
|---|-----------|------------|--------|-----|------------|------------|--------|-----------|
| 1 | B | 1.6 | 0.834 | 0.01 | 0.81 | 0.91 | 0 | Negligible |
| 2 | C | 1.8 | 0.831 | 0.05 | 0.77 | 0.92 | -0.33 | Small |
| 3 | A | 2.6 | 0.813 | 0.01 | 0.80 | 0.90 | 0.60 | Medium |

Out of the 3 feature sets, Set B (with 66 features) has got the best mean rank, and hence can be considered better than the rest. Set C with 116 features and Set A with 39 features came second and third respectively. This goes to show that it is essential to experiment with different feature sets and understand the contributions of each feature in enhancing the results, before deciding on the ultimate set of features to use.

### 3.6.2.2.3 Most suitable preictal time

In the statistical evaluation of phase 2 performance scores, the 30 minute preictal window emerged as a better performer compared to the other 5 periods. The analysis to identify the best preictal time was done as follows:

| Table 11: Statistical evaluation of preictal time |
|-------------------------------------------------|
| # | Population | Sample | P value alpha | Normality | Homoscedastic | Omnibus | Post hoc |
|---|------------|--------|---------------|-----------|---------------|---------|----------|
| 1 | 6-preictal time | 5 performance scores (F1, Recall, Precision, Accuracy and MCC) | 0.0020 | True (Shapiro-Wilk, with Bonferoni correction) | False | Friedman | Nemenyi |

We rejected the null hypothesis of the Friedman test, and concluded that differences between populations are significant, if the difference of the mean rank is greater than the critical distance (CD) =0.5524 of the Nemenyi test. The table below shows the ranking of the preictal time, with 30 minutes at the top.
When 30 minute preictal period was chosen, all the classifiers in both phase 1 and phase 2 gave their best scores. It can be seen that the scores of all the classifiers increased steadily from the 5th to the 30th minutes, but dropped drastically in the 45th and 60th minutes. Our data shows that preictal windows in the range of 10 to 30 minutes is a better choice for researchers than windows above 30 minutes. Our findings are also in line with Texeira et al., who compared 4 different preictal time (10, 20, 30 and 40) and reported that 30.47 min was the most suitable for a patient-specific best predictor. Bandarabadi et al. tested the same 4 preictal times, reporting the preferred average as 33.7 min.  

3.6.2.2.4 Best combination of classifier, set and preictal time

In the above three sections, we derived the top ranked classifier variants, feature set and preictal period to do seizure experiments with. Our next goal is to determine the top combination of classifier variant, feature and preictal time using statistical methods.

### Table 12: Results of statistical evaluation of preictal time

| # | Preictal Time | Mean Rank | Median | MAD | CI (lower) | CI (Upper) | Effect | Magnitude |
|---|---------------|-----------|--------|-----|------------|------------|--------|-----------|
| 1 | 30 min        | 2         | 0.947  | 0.01| 0.90       | 0.98       | 0      | Negligible|
| 2 | 10 min        | 2.2       | 0.940  | 0.03| 0.87       | 1.01       | 0.25   | Small     |
| 3 | 15 min        | 2.8       | 0.938  | 0.02| 0.88       | 0.99       | 0.37   | Small     |
| 4 | 5 min         | 3         | 0.938  | 0.03| 0.85       | 1.02       | 0.30   | Small     |
| 5 | 45 min        | 5         | 0.648  | 0.11| 0.38       | 0.90       | 3.50   | Large     |
| 6 | 60 min        | 6         | 0.611  | 0.11| 0.35       | 0.86       | 4.03   | Large     |

The statistical analysis was conducted for 13 populations with 5 paired samples. The family-wise significance level of the tests is alpha=0.050. We failed to reject the null hypothesis that the population is normal. We applied Bartlett's test for homogeneity and failed to reject the null hypothesis that the data is homoscedastic. As we have more than two populations and all populations are normal and homoscedastic, we use repeated measures ANOVA as omnibus test to determine if there are any significant differences between the mean values of the populations. We used the post-hoc Tukey HSD test to infer which differences are significant.

### Table 13: Statistical evaluation of best combination

| # | Population | Sample | P value alpha | Normality | Homo-scedastic | Omnibus | Post hoc |
|---|------------|--------|---------------|-----------|----------------|---------|----------|
| 1 | 13 classifier | 5 set of performance scores (FL, Recall, Precision, Accuracy and MCC) | 7.954833837315617e-13 | True | True | Anova | Tukey HSD |

### Table 14: Results of statistical evaluation of 13 classifier on Set B and 30 minutes preictal time

| Classifier | Mean Rank | Mean | STD | CI (lower) | CI (Upper) | Effect | Magnitude |
|------------|-----------|------|-----|------------|------------|--------|-----------|
| 1 RF W+30 min+B | 1.6 | 0.973 | 0.01 | 0.95 | 0.99 | 0.0 | Negligible |
| 2 XGB W+30 min+B | 2.6 | 0.970 | 0.015 | 0.95 | 0.99 | 0.20 | Small |
| 3 RF+B+30min+B | 3.2 | 0.969 | 0.012 | 0.94 | 0.98 | 0.33 | Small |
| 4 LGB W+30 min+B | 3.9 | 0.967 | 0.017 | 0.94 | 0.98 | 0.37 | Small |
| 5 LG+B+30 min+B | 4.3 | 0.967 | 0.017 | 0.94 | 0.98 | 0.38 | Small |
| 6 XGB+B+30 min+B | 5.9 | 0.964 | 0.018 | 0.94 | 0.98 | 0.54 | Medium |
| 7 RF B+30 min+B | 7.1 | 0.957 | 0.016 | 0.93 | 0.97 | 1.08 | Large |
| 8 RF+30 min+B | 8.7 | 0.951 | 0.018 | 0.93 | 0.97 | 1.35 | Large |
| 9 XGB+30 min+B | 9.1 | 0.951 | 0.019 | 0.93 | 0.97 | 1.34 | Large |
| 10 LGB+30 min+B | 9.5 | 0.953 | 0.024 | 0.93 | 0.97 | 1.03 | Large |
| 11 RF+S+30 min+B | 10.4 | 0.948 | 0.019 | 0.92 | 0.96 | 1.47 | Large |
| 12 XGB+S+30 min+B | 11.7 | 0.945 | 0.021 | 0.92 | 0.96 | 1.56 | Large |
| 13 LGB+S+30 min+B | 13 | 0.935 | 0.018 | 0.91 | 0.95 | 2.39 | Large |

Since our analysis showed 30 minutes at the optimal preictal period and Set B as the recommended set, we had to identify which classifier variant gave best performance against the recommended set and time. The results show that when specifically grouped with the top ranked feature set and time, RF (weighted) is actually the best performing classifier variant,
beating LGB variants which emerged as the top 3 individual classifiers as described in Table 8. This analysis shows us that we are very likely to get better results if we use the ranked classifiers or make ensembles out of them.

3.6.3 Phase 3 – Ensembling

As a next step, we proceeded to create ensemble models with the top ranked combinations shown on Table 14. The goal was to evaluate if ensembling will give even higher scores compared to top individual classifier scores. It is not mandatory to create ensembles, and we are free to work directly with individual high-performing classifiers.

3.6.3.1 Statistical evaluation of phase 3

Using a Python script, we generated multiple combinations of ensembles from the top 4 classifier variants and combined their results with the results of the top 4 classifier variants.

The best performer was Ensemble 1 (E1) made up of the weighted and balanced variants of RF, XGB and LGB, with an F1 score of 98.3%, precision of 98.1%, recall of 98.3.5%, and MCC of 95.4%. E1 also had the highest AUC of 98.3% and lowest FPR of 0.0156/hr. The performance of E1 was greater than the top individual performance of 98.1 F1 from RF. W seen in phase 2, indicating that ensembling is in fact a useful medium to increase performance scores.

### Table 15: Performance scores of ensembles versus independent variants

| # | Model                | F1   | Recall | Precision | Accuracy | MCC   | AUC   | FPR  |
|---|----------------------|------|--------|-----------|----------|-------|-------|------|
| 1 | E1 (Weight + Balance)| 0.9831 | 0.9831 | 0.9765   | 0.9546   | 0.9837 | 0.0156 |
| 2 | E2 (Weights only)    | 0.9810 | 0.9814 | 0.9806   | 0.9736   | 0.9489 | 0.9821 | 0.0172 |
| 3 | RF W (Weighted)      | 0.9810 | 0.9814 | 0.9806   | 0.9736   | 0.9489 | 0.9821 | 0.0172 |
| 4 | XGB W (Weighted)     | 0.9789 | 0.9797 | 0.9782   | 0.9706   | 0.9434 | 0.9804 | 0.0188 |
| 5 | LGB, W (Weighted)    | 0.9727 | 0.9746 | 0.9711   | 0.9618   | 0.9269 | 0.9755 | 0.0236 |
| 6 | RF, B (Balanced)     | 0.9705 | 0.9581 | 0.9839   | 0.9765   | 0.9544 | 0.9709 | 0.0163 |

As an additional check for classifier performance, we calculated the Youden's index, positive and negative likelihood ratios and the discriminant power to get better insights about the nature of the various classifiers.

### Table 16: YI, PLR, NLR and DP

| # | Ensemble          | Youden’s Index(γ) | Positive Likelihood Ratio(ρ+) | Negative Likelihood Ratio (ρ−) | Discriminant Power(DP) |
|---|-------------------|-------------------|--------------------------------|---------------------------------|-------------------------|
| 1 | E1 (Weight Balance) | 1.0               | 50.54                          | 0.00                            | 1.77                    |
| 2 | E2 (Weights only) | 1.0               | 40.43                          | 0.00                            | 1.71                    |
| 3 | RF, W (Weighted) | 1.0               | 40.43                          | 0.00                            | 1.71                    |
| 4 | XGB, W (Weighted) | 1.0               | 33.69                          | 0.00                            | 1.67                    |
| 5 | LGB, W (Weighted) | 1.0               | 22.46                          | 0.00                            | 1.57                    |
| 6 | RF, B (Balanced) | 0.92              | 67.39                          | 0.08                            | 1.84                    |

As per the Youden’s Index, except for RF, B(balanced), all the ensembles and independent classifier variants have good ability to avoid failures when it comes to the Ictal class. RF, B has some chances of failure in identifying the Ictal class, but higher values of γ in the Preictal and Non Seizure classes indicate comparatively better abilities in handling these classes.

A higher positive likelihood (ρ+) indicates a better probability to identify a particular condition. As per the (ρ+), all classifiers and ensembles have ‘infinitely’ superior ability to locate the positives (true and false positives) in the Ictal class. RF, B has the best ability to locate the
positives in the Non Seizure condition, while E1 has better abilities to identify the Preictal condition. In terms of (ρ−), all ensembles and classifier variants, except RF_B have good abilities to identify the negatives (true and false negatives) in the Ictal class. RF_B is also the best performer when it comes to locating negatives in the Non Seizure and Preictal classes. RF_B again has better abilities to discriminate among the 3 classes based on higher values of Discriminant Power, which is actually in disagreement with the findings of the Youden’s Index regarding the Ictal class. Thus, we expect to see good performance from all the models on the field, except for RF_B, which might throw some surprises.

3.6.3.2 Testing on unseen data

The statistical analysis in the previous section has given us a hint about how the different classifiers may behave in real world testing. To be qualified for use as a clinical decision making tool, it is imperative for all results to be tested on real world data or such data that is previously unseen by the models. Hence, to test the soundness of the findings, we tested the performance of the classifier variants and ensembles on 12 actual patient’s data. Table 18 shows the findings:

Table 17: Performance on real world data

| Subject | E1 | E2 | RF_B | RF_W | LGB_W | XGB_W |
|---------|----|----|------|------|-------|-------|
| Test Subject | D P Min. | D P Min. | D P Min. | D P Min. | D P Min. | D P Min. |
| TS1 | Y Y 26 | Y Y 26 | N N - | Y Y 83 | Y Y 25 | Y Y 26 |
| TS2 | Y N 0 | N N 0 | N N - | N N 0 | Y Y 0 | N N 0 |
| TS3 | Y Y 7 | Y N 0 | N N - | Y N 0 | Y N 5 | Y N 0 |
| TS4 | N N 0 | Y Y 8 | N N - | N N 0 | Y Y 0 | N N 0 |
| TS5 | Y Y 27 | Y N 0 | N N - | Y Y 5 | Y N 23 | N N 0 |
| TS6 | Y Y 9 | Y Y 27 | N N - | Y Y 90 | Y Y 9 | Y Y 9 |
| TS7 | Y Y 24 | Y Y 9 | N N - | Y Y 9 | Y Y 24 | Y Y 24 |
| TS8 | N N 0 | Y Y 24 | N N - | N N 0 | Y Y 2 | Y N 5 |
| TS9 | Y Y 31 | Y Y 13 | N N - | Y Y 13 | Y Y 30 | Y Y 29 |
| TS10 | N N 0 | Y Y 30 | N N - | Y Y 25 | Y Y 176 | N N 0 |
| TS11 | N N 0 | Y Y 33 | N N - | Y Y 30 | Y Y 33 | N N 0 |
| TS12 | Y Y 151 | Y Y 151 | N N - | Y Y 151 | Y Y 7 | Y Y 146 |

| Count -> | 9 7 35.4 | 11 9 32.1 | 0 0 | 9 8 50.7 | 12 10 37.1 | 7 5 34.14 |

*TS is Test Subject, ‘D’ is Detection, ‘P’ is Prediction and ‘Min’ refers to the time to clinical seizure onset or the predicted preictal time. ‘Y’ represents yes and inform whether the model has been able to correctly identify a seizure or predict the preictal time. N represents a model's inability to either detect or predict a seizure.*

E1 with 98.3% F1 was the top performer as per the phase 3 analysis, however, in real world testing it was able to correctly detect only 9 out of 12 seizures, and predict only 7, with an average prediction time of 35.4 minutes. E2 with 98.1% F1, however, was able to detect 11 out of 12 seizures, though it had an F1 of 98.1% F1. RF_W had identical performance score as E2, but it was able to detect only 9 out of 12 seizures and predict 8, though it had the best average predictive time of 50.7 minutes among all the different models. However, the best performer on real world data was LGB_W, which was able to detect all the seizures and predicted 10 of them. Interestingly, LGB_W was a comparatively mediocre scorer in phase 3, similar to XGB_W. The worst performer was RF_B, which was not able to identify any one of the seizures. The hint about RF_B's performance can been had from Youden' Index and the negative likelihood ratio that indicated that RF_B had lesser abilities to identify the Ictal class compared to the others. In terms of predictive abilities, the highest individual prediction time was 176 minutes given by LGB_W, while the best overall prediction time across models was given by RF_W. The mean prediction time of all the 6 models combined is 31.56 minutes.

This final step of testing is critical to understand the real abilities of different models on the field. Depending on the requirement of either detecting a seizure or predicting it, a single model or a combination of several could be used for further study, validation and real life implementation. Unless the last two steps (phase 3) and real world testing was carried out, we would not have about the true abilities of our models. Our methodology thus enables
decision makers to choose a model that meets their clinical and research requirements, balancing between detection and prediction needs.

3.7 Post experiment analysis

Through our work, we are able to show how realistic, predictive and high performing seizure models could be created by following a phase-wise approach. We were guided by statistical insights at every stage of our work, leading us to understand the inner workings of the various classifiers, algorithms and models we used.

We derived the following insights by analysing the results of our experiments.

- Phase-wise selection of classifiers based on robust statistical methods yield more reliable results in unseen data during testing. We were able to determine the classifiers that were more likely to give better results, depending on the nature of data. We recommend that researchers experiment with different classifiers and their variations, eliminating redundant ones methodically and choosing ones that are best qualified to represent their data in a consistent manner.

- Seizure models built using 30 minutes prior to seizure onset as the preictal time provided the best results across phase 1, 2 and 3 classifiers. As the fixed preictal time increased from 5 to 30 minutes, the performance of the classifiers increased, but dropped drastically in the 45th and 60th minutes. In fact, the classifiers made more errors in distinguishing the ictal and Non seizure classes in the 45th and 60th minute datasets, and Preictal and Non seizure class between 5 and 15 minutes. During the 30th minute, the top classifiers perfectly identified the ictal class, but made comparatively lesser errors in distinguishing the Preictal and Non seizure class.

  **Figure 1: Classifier performance against preictal time**

- Simple features provided best results. Features that tap into the frequency attributes of the EDA and BVP signals gave better results compared to complex features requiring higher computational power. Though we calculated multiple features (over 700 from Tsfresh alone) for each signal, such features were shown to be redundant during analysis of feature importance. While for us 66 features gave better results, researchers can arrive at their own number based on feature performances.

- Tree-based classification models were preferable for our seizure data. Our work shows that due to the non-linearity and complex relationship between dependent and independent variables in seizure data, tree-based models like RF, XGB and LGB outperformed other methods. Unlike others, these algorithms are non-parametric and can efficiently deal with large datasets without imposing a complicated parametric structure. As shown through
our multiple statistical analysis of classifiers, these 3 classifiers consistently perform better than the rest, irrespective of the preictal windows used.

3.8 Future work

Our work shows promise in laying the foundation for a machine learning framework for seizure research with wearables. The validity of our finding needs to be tested using more heterogeneous seizure data and replicated rigorously in both clinical and field settings. Our future work will be focused on refining our understanding of seizures using traditional equipment like EEG and ECG along with wearables, creating dynamic methods to identify the preictal period using hybrid ML/AI methodologies, predicting seizures in realtime, verifying the quality of data sampled using varying sampling rates, studying the quality of data collected from multiple sites, analysing the impact of various anti-epilepsy drugs and extending our ML/AI methodologies to other aspects of seizure research including Sudden Death in Epilepsy (SUDEP).

4 Conclusion

Our work is an important step in bringing structure to ML-based seizure research. We show that a step-wise classifier selection mechanism based on the project’s or researchers’ end goals will lead to realistic performance scores, with better than chance results. The limitation of our study is that it was done on data collected from a controlled-hospital environment, with a limited number of subjects. More studies need to be conducted to prove the effectiveness of the methodologies described here. Automated seizure detection has the potential to change the life of people suffering from seizures, by enabling faster access to treatment and care. We hope our work will help others to arrive at the right strategies to adopt while exploring seizures in the context of making lives better of people with epilepsy.

5 Acknowledgement

We thank the Ethical Committee of the National Institute of Mental Health and Neurosciences (NIMHANS), India for enabling us to collect data for our study. Sincere thanks to Dr. Sanjib Sinha, Dr. M Ravindranad Chowdary and Dr. Raghavendra K of the Neurology department of NIMHANS, along with the entire staff and technicians for their kind support and co-operation for our work. We are especially indebted to Dr. Raghavendra K for reviewing our methodologies, matching the wearable-derived seizure onset time with EEG onset time, and giving us valuable suggestions. We are thankful to the Physiology department, especially Dr. Sathyaprabha T. for her continuous support and for enabling us access to the NIMHANS ANS Lab for calibration of our wearable against standard equipment. We are indebted to Dr. Natwar Sharma, Dr. Shivakumar Shamarao and Dr. Pramod Krishnan for their guidance and support in building the technology. We are indebted to the entire team at Biosignals Plux for supporting us with prototypes, without which we would not have been able to complete our work. We are grateful to the Govt. of India, Departments of Biotechnology, BIRAC, C-CAMP, Bangalore Bioinnovation Center and the Govt. of Karnataka for financially supporting our work through grants. Finally, our heartfelt gratitude to all employees, past and present of TerraBlue XT, for being part of the research.

6 Conflict of interest

Rajlakshmi Borthakur is the founder of TerraBlue XT and the innovator of TJay, a wearable device for epilepsy management. TerraBlue XT has received grants from the Department of Biotechnology (DBT), Govt. of India and the Govt. of Karnataka, India. The company has also received funds from two GoI-supported equity funds. Amit Kumar and Upendra Jena are both Data Scientists, while Dibyajyoti Borthakur is the Chief Technology Officer (CTO) at TerraBlue XT.
7 References

1. https://www.who.int/news-room/fact-sheets/detail/epilepsy#
2. Dumanis, S.B., French, J.A., Bernard, C., Worrell, G.A., Fureman, B.E., 2017. Seizure forecasting from idea to reality. Outcomes of the my seizure gauge epilepsy innovation Institute workshop. eNeuro 4 (6).
3. Singh, Prem & Pandey, Achyut. (2017). Quality of life in epilepsy. International Journal of Research in Medical Sciences. 5. 10.18203/2320-6012.ijrms20170024.
4. Sanyal D., Roy MK. Understanding stigma in epilepsy. Neuroscience Today. 2004;8:136-9.
5. Jacoby A. Epilepsy and the quality of everyday life. Findings from a study of people with well-controlled epilepsy. Soc Sci Med 1992;34(6):657–66.
6. M. O. Baud, J. K. Kleen, E. A. Mirro, J. C. Andrechak, D. King Stephens, E. F. Chang, and V. R. Rao. Multi-day rhythms modulate seizure risk in epilepsy. Nat. Commun., 9(1):88, 01 2018. 19.
7. Ryvlin, P, Ciumas, C, Wisniewski, I, Beniczky, S. Wearable devices for sudden unexpected death in epilepsy prevention. Epilepsia. 2018; 59 (S1): 61– 66. https://doi.org/10.1111/epi.14054
8. Donos C, Dümpelmann M, Schulze-Bonhage A. Early seizure detection algorithm based on intracranial EEG and random forest classification. Int J Neur Syst. 2015;25(05):1550023.
9. Orellana MP, Cerqueira F (2016) Personalized epilepsy seizure detection using random forest classification over one-dimension transformed EEG data. bioRxiv, 070300
10. Birjandtalab J, Pouryan MB, Cogan D, Nourani M, Harvey J. Automated seizure detection using limited-channel eeg and non-linear dimension reduction. Comput Biol Med. 2017;82:49–58.
11. Mursalin M, Zhang Y, Chen Y, Chawla NV. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. Neurocomputing. 2017;241:204–214.
12. Truong ND, Kuhlmann L, Bonyadi MR, Yang J, Faulks A, Kavehei O (2017) Supervised learning in automatic selection for epileptic seizure detection. Exp Syst Appl
13. Tzimourta KD, Tzallas AT, Giannakeas N, Astrakas LG, Tsalikakis DG, Angelidis P, Tsiouras MG. A robust methodology for classification of epileptic seizures in eeg signals. Health Technol. 2019;9(2):135–142
14. Wang X, Gong G, Li N, Qiu S. Detection analysis of epileptic EEG using a novel random forest model combined with grid search optimization. Front Human Neurosci. 2019;13:12.
15. Alickovic E, Kevric J, Subasi A. Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. Biomed Sign Process Contr. 2018;39:94–102.
16. Hussein R, Elgendi M, Wang ZJ, Ward RK. Robust detection of epileptic seizures based on l1-penalized robust regression of EEG signals. Exp Syst Appl. 2018;104:153–167.
17. Sharma RR, Varshney P, Pachori RB, Vishwakarma SK. Automated system for epileptic EEG detection using iterative filtering. IEEE Sens Lett. 2018;2(4):1–4.
18 Shoeb A, Guttag J (2010) Application of machine learning to epileptic seizure detection. In: 2010 the 27th International Conference on Machine learning, Haifa, Israel

19 Zhang Y, Zhang Y, Wang J, Zheng X. Comparison of classification methods on EEG signals based on wavelet packet decomposition. Neural Comput Appl. 2014;26(5):1217–1225. doi: 10.1007/s00521-014-1786-7.

20 Ahmad MA, Khan NA, Majeed W (2014) Computer assisted analysis system of electroencephalogram for diagnosing epilepsy. In: Pattern Recognition (ICPR), 2014 22nd International Conference On, pp. 3386–3391. IEEE

21 Amin HU, Malik AS, Ahmad RF, Badruddin N, Kamel N, Hussain M, Chooi W-T. Feature extraction and classification for eeg signals using wavelet transform and machine learning techniques. Austr Phys Eng Sci Med. 2015;38(1):139–149.

22 Zabihi M, Kiranyaz S, Ince T, Gabbouj M (2013) Patient-specific epileptic seizure detection in long-term eeg recording in paediatric patients with intractable seizures

23 Fergus P, Hussain A, Hignett D, Al-Jumeily D, Abdel-Aziz K, Hamdan H. A machine learning system for automated whole-brain seizure detection. Appl Comput Inform. 2016;12(1):70–89.

24 Chen D, Wan S, Xiang J, Bao FS. A high-performance seizure detection algorithm based on discrete wavelet transform (dwt) and eeg. PLoS ONE. 2017;12(3):0173138.

25 Satapathy SK, Jagadev AK, Dehuri S. Weighted majority voting based ensemble of classifiers using different machine learning techniques for classification of eeg signal to detect epileptic seizure. Informatica. 2017;41(1):99.

26 Subasi A, Kevric J, Canbaz MA. Epileptic seizure detection using hybrid machine learning methods. Neural Comput Appl. 2019;31(1):317–325.

27 Al Ghayab HR, Li Y, Siuly S, Abdulla S. Epileptic seizures detection in eegs blending frequency domain with information gain technique. Soft Comput. 2019;23(1):227–239.

28 Selvakumari RS, Mahalakshmi M, Prashelee P. Patient-specific seizure detection method using hybrid classifier with optimized electrodes. J Med Syst. 2019;43(5):121

29 Chen S, Zhang X, Chen L, Yang Z. Automatic diagnosis of epileptic seizure in electroencephalography signals using nonlinear dynamics features. IEEE Access. 2019;7:61046–61056.

30 Fasil O, Rajesh R. Time-domain exponential energy for epileptic eeg signal classification. Neurosci Lett. 2019;694:1–8.

31 Parvez MZ, Paul M. Epileptic seizure detection by analyzing eeg signals using different transformation techniques. Neurocomputing. 2014;145:190–200.

32 Guo L, Rivero D, Dorado J, Rabunal JR, Pazos A. Automatic epileptic seizure detection in eegs based on line length feature and artificial neural networks. J Neurosci Methods. 2010;191(1):101–109.

33 Birjandtalab J, Jarmale VN, Nourani M, Harvey J (2018) Imbalance learning using neural networks for seizure detection. In: 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), pp. 1–4. IEEE
34 Tzallas AT, Tsipouras MG. Fotiadis DI (2007) Automatic seizure detection based on time-frequency analysis and artificial neural networks. Comput Intell Neurosci. 2007;2007:80510.

35 Bandarabadi, Mojtaba, et al. "On the proper selection of preictal period for seizure prediction." Epilepsy & Behavior 46 (2015): 158-166.

36 Mirowski P, Madhavan D, LeCun Y, Kuzniecky R. Classification of patterns of EEG synchronization for seizure prediction. Clin Neurophysiol 2009;120:1927–40.

37 Williamson JR, Bliss DW, Browne DW, Narayanan JT. Seizure prediction using EEG spatiotemporal correlation structure. Epilepsy Behav 2012;25:230–8.

38 Park Y, Luo L, Parhi KK, Netoff T. Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. Epilepsia 2011;52:1761–70

39 Rasekhi J, Mollaei MRK, Bandarabadi M, Teixeira CA, Dourado A. Preprocessing effects of 22 linear univariate features on the performance of seizure prediction methods. J Neurosci Methods 2013;217:9–16.

40 Moghim N, Corne DW. Predicting epileptic seizures in advance. PLoS One 2014;9: e99334

41 Teixeira CA, Direito B, Bandarabadi M, Le Van Quyen M, Valderrama M, Schelter B, et al. Epileptic seizure predictors based on computational intelligence techniques: a comparative study with 278 patients. Comput Methods Programs Biomed 2014; 114:324–36

42 Valderrama M, Alvarado C, Nikolopoulos S, Martinerie J, Adam C, Navarro V, et al. Identifying an increased risk of epileptic seizures using a multi-feature EEG–ECG classification. Biomed Signal Process Control 2012;7:237–44.

43 Bandarabadi M, Teixeira CA, Rasekhi J, Dourado A. Epileptic seizure prediction using relative spectral power features. Clin Neurophysiol 2015;126:237–48

44 Mormann F, Kreuz T, Riecke C, Andrzejak RG, Kraskov A, David P, et al. On the predictability of epileptic seizures. Clin Neurophysiol 2005;116:569–87.

45 Rasekhi J, Karami-Mollaei M, Bandarabadi M, Teixeira C, Dourado A. Epileptic seizure prediction based on ratio and differential linear univariate features. J Med Signals Sensors 2015;5:1–11.

46 Gu, Qiong, Li Zhu, and Zhihua Cai. "Evaluation measures of the classification performance of imbalanced data sets." International symposium on intelligence computation and applications. Springer, Berlin, Heidelberg, 2009.

47 Siddiqui, Mohammad Khubeb, et al. "A review of epileptic seizure detection using machine learning classifiers." Brain informatics 7 (2020): 1-18.

48 Acharya, U Rajendra & Hagiwara, Yuki & Adeli, Hojjat. (2018). Automated Seizure Prediction. Epilepsy & behavior : E&B. 88. 10.1016/j.yebeh.2018.09.030.

49 Ghaderyan P, Abbasi A, Sedaaghi MH. An efficient seizure prediction method using KNN-based undersampling and linear frequency measures. Journal of Neuroscience Methods 2014;232:134–142

50 Vahabi Z, Amirfattahi R, Shayegh F, Ghassemi F. Online epileptic seizure prediction using wavelet-based bi-phase correlation of electrical signals tomography. International Journal of Neural Systems 2015;25:1550028

51 Zhang Z, Parhi KK. Low-complexity seizure prediction from iEEG/sEEG using spectral power and ratios of spectral power. IEEE Transactions on Biomedical Circuits and Systems 2016;10:693-706.
52 Boonyakitanont, Poomipat, et al. "A review of feature extraction and performance evaluation in epileptic seizure detection using EEG." Biomedical Signal Processing and Control 57 (2020): 101702.

53 Paul, Yash. "Various epileptic seizure detection techniques using biomedical signals: a review." Brain informatics 5.2 (2018): 1-19.

54 Tzallas, Alexandros T., et al. "Automated epileptic seizure detection methods: a review study." Epilepsy-histological, electroencephalographic and psychological aspects (2012): 75-98.

55 Cho D, Min B, Kim J, Lee B. EGG-based prediction of epileptic seizures using phase synchronization elicited from noise-assisted multivariate empirical mode decomposition. IEEE Transactions on Neural Systems and Rehabilitation Engineering 2017;25:1309-1318

56 Tsiouris KM, Pezoulas V, Zervakis M, Konitsiotis S, Koutsiouris DD, Fotiadis DI. A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. Computers in Biology and Medicine 2018;99:24-37

57 Malarvili, M. Balakrishnan, and Mostefa Mesbah. "Newborn seizure detection based on heart rate variability." IEEE transactions on biomedical engineering 56.11 (2009): 2594-2603.

58 Varon C, Caicedo A, Jansen K, Lagae L, Van Huffel S. Detection of epileptic seizures from single lead ECG by means of phase rectified signal averaging. Conf Proc IEEE Eng Med Biol Soc. 2014;2014:3789-3790. doi:10.1109/EMBC.2014.6944448

59 Fujiwara, Koichi, et al. "Epileptic seizure prediction based on multivariate statistical process control of heart rate variability features." IEEE Transactions on Biomedical Engineering 63.6 (2015): 1321-1332.

60 Beniczky S, Polster T, Kjaer TW, Hjalgrim H. Detection of generalized tonic clonic seizures by a wireless wrist accelerometer: a prospective, multicentre study. Epilepsia 2013;54(4):e58–61.

61 Lockman J, Fisher RS, Olson DM. Detection of seizure-like movements using a wrist accelerometer. Epilepsy Behav 2011;20(4):638–41.

62 Jallon P, Bonnet S, Antonakios M, Guillemaud R. Detection system of motor epileptic seizures through motion analysis with 3D accelerometers. Conf Proc IEEE Eng Med Biol Soc 2009;2009:2466–9.

63 Jansen K, Varon C, Van Huffel S, Lagae L. Peri-ictal ECG changes in childhood epilepsy: implications for detection systems. Epilepsy Behav 2013;29(1): 72–6.

64 Romero I, Grundlehner B, Penders J. Robust beat detector for ambulatory cardiac monitoring. Conf Proc IEEE Eng Med Biol Soc 2009;2009:950–3.

65 Larsen SN, Conradsen I, Beniczky S, Sorensen HB. Detection of tonic epileptic seizures based on surface electromyography. Conf Proc IEEE Eng Med Biol Soc 2014;2014:942–5.

66 Conradsen I, Beniczky S, Wolf P, Jennum P, Sorensen HB. Evaluation of novel algorithm embedded in a wearable sEMG device for seizure detection. Conf Proc IEEE Eng Med Biol Soc 2012;2012:2048–51.

67 Poh MZ, Swenson NC, Picard RW. A wearable sensor for unobtrusive: long term assessment of electrodermal activity. IEEE Trans Biomed Eng 2010;57(5):1243–52.

68 Poh MZ, Loddenkemper T, Swenson NC, Goyal S, Madsen JR, Picard RW. Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor. Conf Proc IEEE Eng Med Biol Soc 2010;2010:4415–8.

69 Kim B, Nogueira AB, Thome-Souza S, Kapur K, Klehm J, Jackson M, et al. Diurnal and nocturnal patterns of autonomic neurophysiological measurements are related to timing of seizures. In: Poster presented at: CNS 2015. Child Neurology Society annual meeting; 2015.
Singh K, Katz ES, Zarowski M, Loddenkemper T, Llewellyn N, Manganaro S, et al. Cardiopulmonary complications during pediatric seizures: a prelude to understanding SUDEP. Epilepsia 2013;54(6):1083–91.

Heldberg, B.; Kautz, T.; Leutheuser, H.; Hopfengartner, R.; Kasper, B.; Eskofier, B. Using wearable sensors for semiology-independent seizure detection—Towards ambulatory monitoring of epilepsy. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 5593–5596.

Al-Bakri, A.; Villamar, M.; Haddix, C.; Bensalem-Owen, M.; Sunderam, S. Noninvasive seizure prediction using autonomic measurements in patients with refractory epilepsy. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 2422–2425.

Vandecasteele, K.; De Cooman, T.; Gu, Y.; Cleeren, E.; Claes, K.; Paesschens, W.; Hu_el, S.; Hunyadi, B. Automated Epileptic Seizure Detection Based onWearable ECG and PPG in a Hospital Environment. Sensors 2017, 17, 2338.

Regalia, G.; Onorati, F.; Lai, M.; Caborni, C.; Picard, R. Multimodal wrist-worn devices for seizure detection and advancing research: Focus on the Empatica wristbands. Epilepsy Res. 2019, 153, 79–82.

Kurada, Abhinav V., et al. "Seizure detection devices for use in antiseizure medication clinical trials: A systematic review." Seizure 66 (2019): 61-69.

Siirtola, Pekka, et al. "Wearable sensors to detect seizures and attacks-challenges of data gathering."

Freeman R, Schachter SC (1995) Autonomic epilepsy. Semin Neurol 15:158–166

Freeman R (2006) Cardiovascular manifestations of autonomic epilepsy. Clin Auton Res 16:12–17. https://doi.org/10.1007/s10286-006-0278-y

Liporace JD, Sperling MR (1997) Simple autonomic seizures. In: Engel Jr, Pedley TA (eds) Epilepsy: a comprehensive textbook. Lippincott-Raven Publishers, Philadelphia, pp 549–555

Reeves AL (1997) Autonomic activity in epilepsy: diagnostic considerations and implications. J Epilepsy 10:111–116

Lesser RP, Lüders H, Dinner DS, Morris HHI (1987) Simple partial seizures. In: Lüders H, Lesser RP (eds) Epilepsy: electroclinical syndromes. Springer, London, pp 223–278

Baumgartner C, Lurger S, Leutmezer F (2001) Autonomic symptoms during epileptic seizures. Epileptic Disord 3:103–116

Braithwaite, Jason J., et al. "A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments." Psychophysiology 49.1 (2013): 1017-1034.

Force, T. (1996). Standards of measurement, physiological interpretation and clinical use. Task force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Circulation, 93(5), 1043-1065.

https://tsfresh.com/

Tanwani, Ajay Kumar, et al. "Guidelines to select machine learning scheme for classification of biomedical datasets." European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Springer, Berlin, Heidelberg, 2009.
87 Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." BMC genomics 21.1 (2020): 1-13.

88 Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." The Journal of Machine Learning Research 7 (2006): 1-30.

89 Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

90 Teixeira, César Alexandre, et al. "Epileptic seizure predictors based on computational intelligence techniques: A comparative study with 278 patients." Computer methods and programs in biomedicine 114.3 (2014): 324-336.

91 Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz. "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." Australasian joint conference on artificial intelligence. Springer, Berlin, Heidelberg, 2006.

92 Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27(2):130-135. doi: 10.11919/j.issn.1002-0829.215044