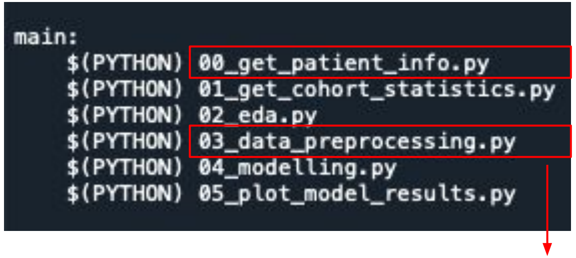
1 INTRODUCTION

SCOPE: Analysis of the publicly available PTB ECG dataset.

CODE STRUCTURE: The analysis is done in 5 separate steps, each assigned to a different script. Two additional utility scripts are used (config.py & utils.py). To launch the analysis, simply call the Makefile on the terminal.

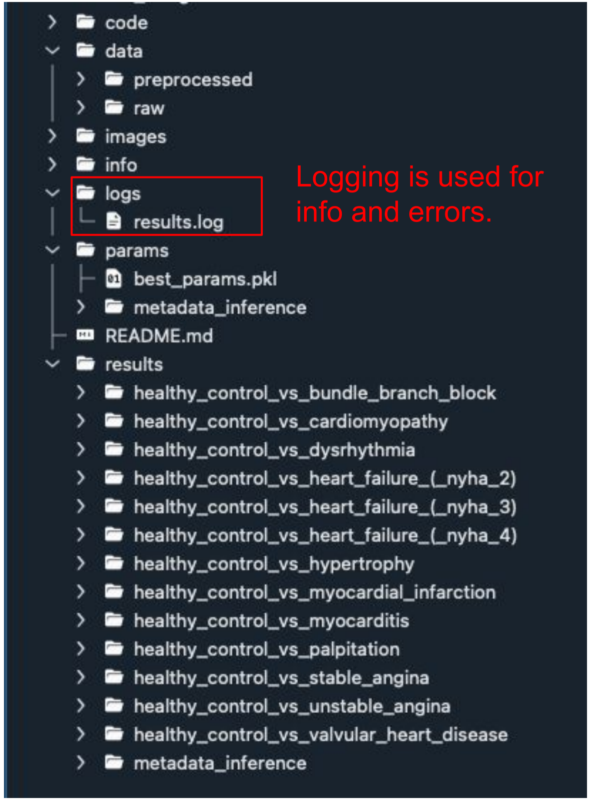
The code uses the sklearnex optimization module to speed up all Scikit-learn operations.



```
main:
	$(PYTHON) 00_get_patient_info.py
	$(PYTHON) 01_get_cohort_statistics.py
	$(PYTHON) 02_eda.py
	$(PYTHON) 03_data_preprocessing.py
	$(PYTHON) 04_modelling.py
	$(PYTHON) 05_plot_model_results.py
```

Fully parallelized at the patient level.

PROJECT STRUCTURE

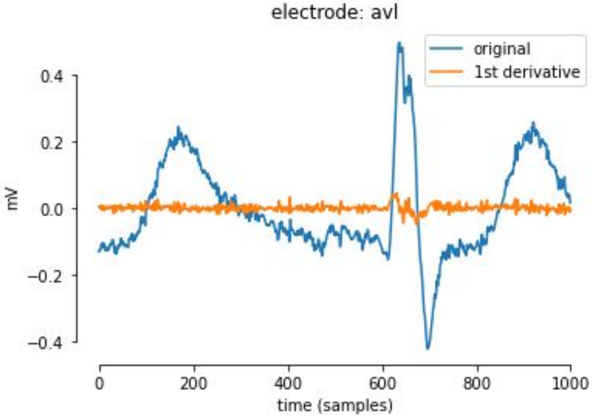


```
> code
v data
    > preprocessed
    > raw
> images
> info
v logs
    └ results.log
v params
    ├ best_params.pkl
    > metadata_inference
├ README.md
v results
    > healthy_control_vs_bundle_branch_block
    > healthy_control_vs_cardiomyopathy
    > healthy_control_vs_dysrhythmia
    > healthy_control_vs_heart_failure_(_nyha_2)
    > healthy_control_vs_heart_failure_(_nyha_3)
    > healthy_control_vs_heart_failure_(_nyha_4)
    > healthy_control_vs_hypertrophy
    > healthy_control_vs_myocardial_infarction
    > healthy_control_vs_myocarditis
    > healthy_control_vs_palpitation
    > healthy_control_vs_stable_angina
    > healthy_control_vs_unstable_angina
    > healthy_control_vs_valvular_heart_disease
    > metadata_inference
```

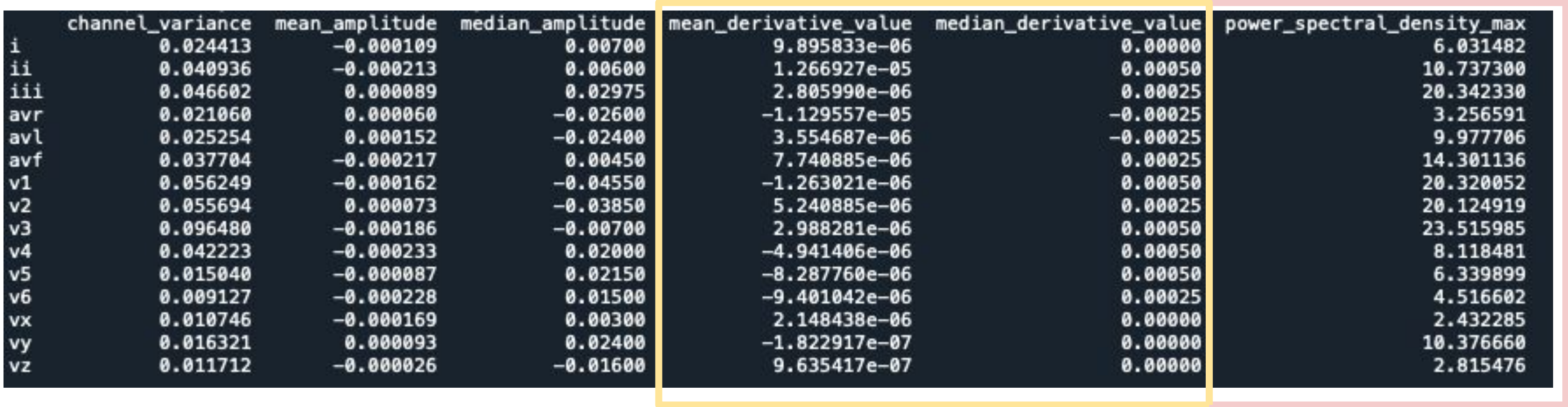
Logging is used for info and errors.

2 FEATURE EXTRACTION

For each patient, and each recording, a series of 6 features were extracted for every sensor.



The 1st derivative is calculated as a measurement of abrupt changes in the signal. Essentially, a way to identify the “R” peaks without presupposing any threshold.



```
     channel_variance  mean_amplitude  median_amplitude  mean_derivative_value  median_derivative_value  power_spectral_density_max
i            0.024413       -0.000109           0.00700           9.895833e-06                  0.00000                    6.031482
ii           0.040936       -0.000213           0.00600           1.266927e-05                  0.00050                   10.737300
iii          0.046602        0.000089           0.02975           2.805990e-06                  0.00025                   20.342330
avr          0.021060        0.000060          -0.02600          -1.129557e-05                 -0.00025                    3.256591
avl          0.025254        0.000152          -0.02400           3.554687e-06                 -0.00025                    9.977706
avf          0.037704       -0.000217           0.00450           7.740885e-06                  0.00025                   14.301136
v1           0.056249       -0.000162          -0.04550          -1.263021e-06                  0.00050                   20.320052
v2           0.055694        0.000073          -0.03850           5.240885e-06                  0.00025                   20.124919
v3           0.096480       -0.000186          -0.00700           2.988281e-06                  0.00050                   23.515985
v4           0.042223       -0.000233           0.02000          -4.941406e-06                  0.00050                    8.118481
v5           0.015040       -0.000087           0.02150          -8.287760e-06                  0.00050                    6.339899
v6           0.009127       -0.000228           0.01500          -9.401042e-06                  0.00025                    4.516602
vx           0.010746       -0.000169           0.00300           2.148438e-06                  0.00000                    2.432285
vy           0.016321        0.000093           0.02400          -1.822917e-07                  0.00000                   10.376660
vz           0.011712       -0.000026          -0.01600           9.635417e-07                  0.00000                    2.815476
```

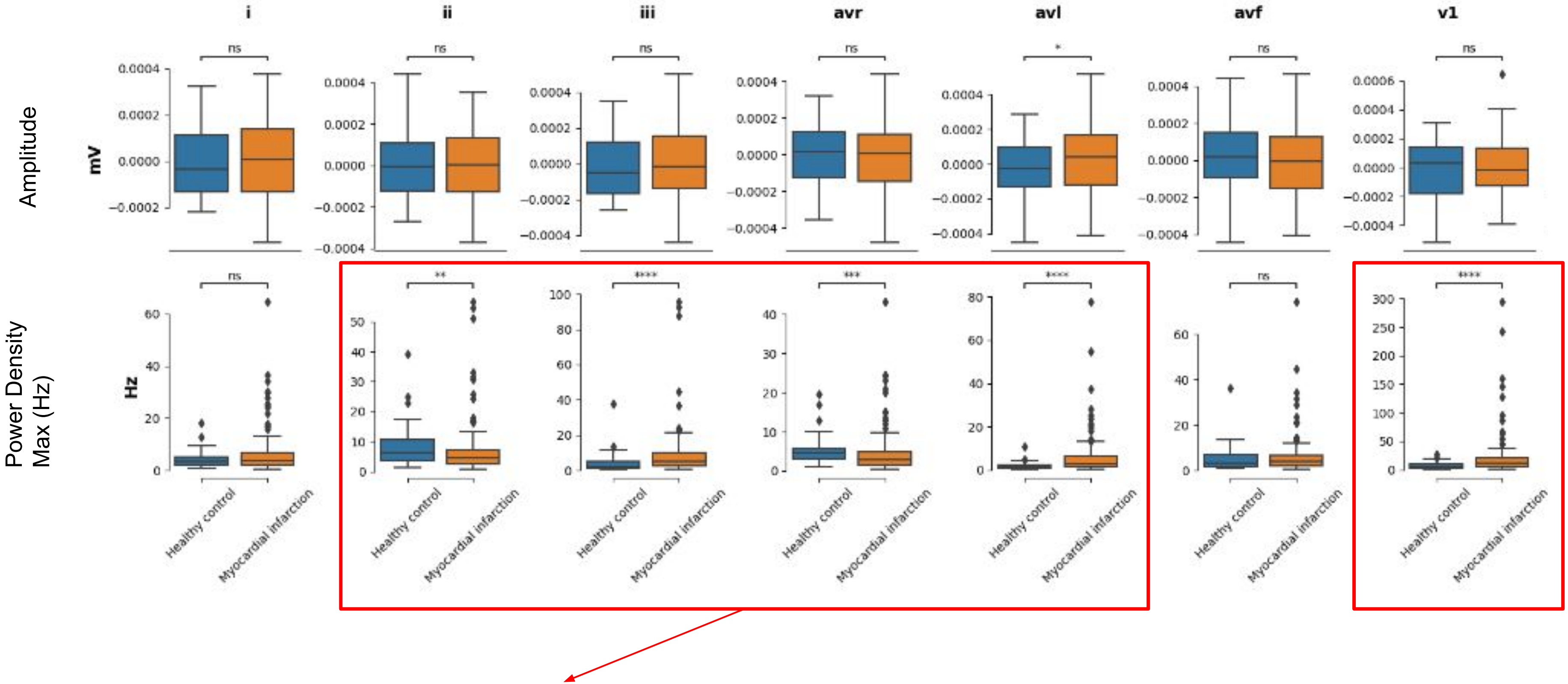
Maximum of power spectral density estimated using Welch's method (in Hz).

* These operations took place for all recordings of a given patient. The corresponding script (“03_data_preprocessing.py”) is parallelized at the patient level.

3 EXPLORATORY DATA ANALYSIS

* A full list of comparisons is available at the “images” directory.

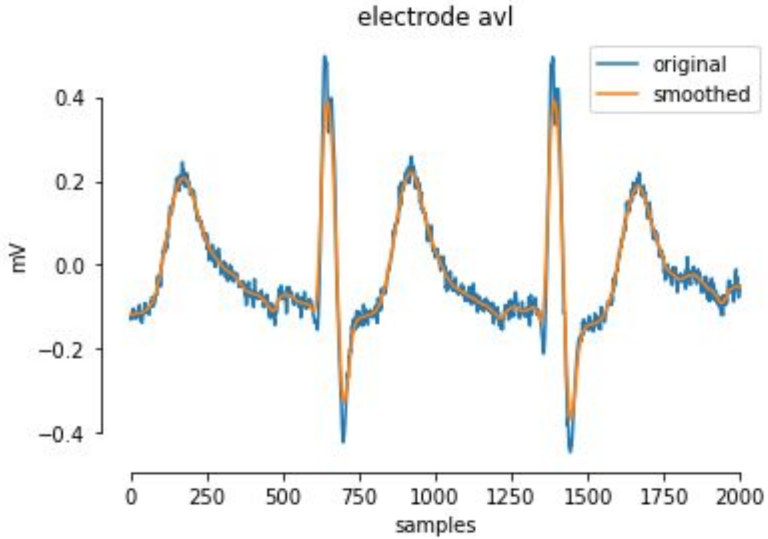
Healthy control VS Myocardial Infarction



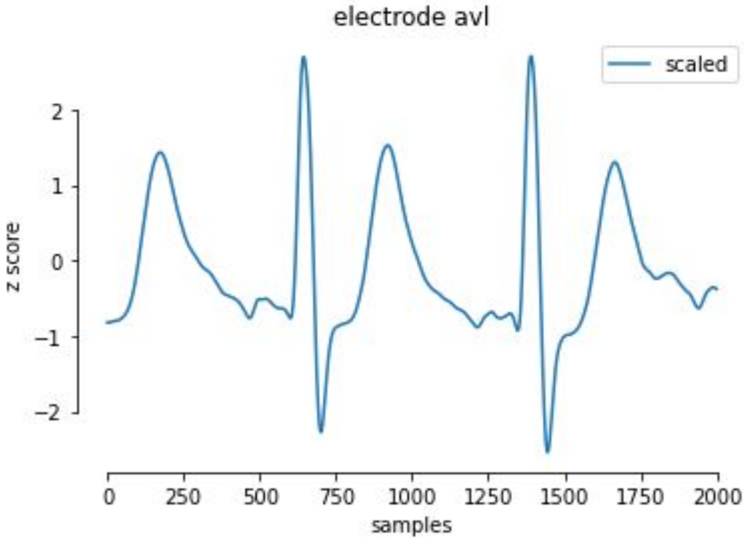
We observe a statistically significant difference in multiple electrodes (p<0.05; Mann-Whitney test, corrected with Bonferroni corrections for multiple comparisons) for the power density feature, but not this amplitude. This type of analysis can help us build more interpretable models or identify limitations of existing ones.

3 PREPROCESSING

(i) The time series were smoothed with a 10ms Gaussian Kernel.

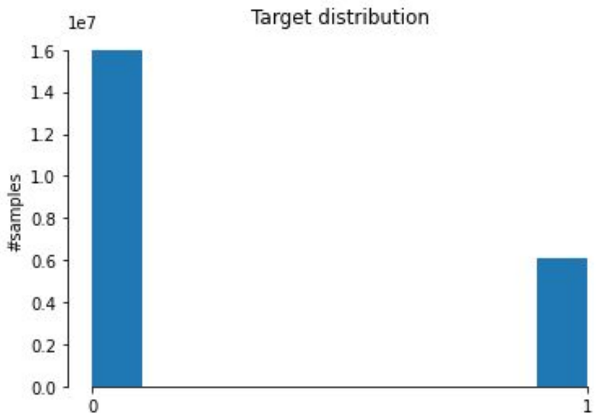


(ii) The smoothed signal was scaled to zero mean and unit variance.



* These operations took place for all recordings of a given patient. The corresponding script (“03_data_preprocessing.py”) is parallelized at the patient level. The data are stored in the “preprocessed” directory,

4 MODELLING

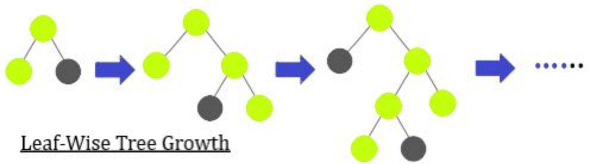


The target distribution is skewed. This indicates that I cannot use metrics such as F1 or accuracy to evaluate our model, AND that I must use a STRATIFIED k-Fold approach to avoid overfitting.

To computationally constrain the task, I transformed the classification to a univariate binary classification problem, essentially contrasting each population (e.g: “myocarditis”) against data from the healthy control group.

As a first approach, I labelled each sample of the time-series based on the population where the patient belonged (e.g: all samples from healthy patients were labelled as 0 and all others as 1). Thus, this a time-resolved classification approach.

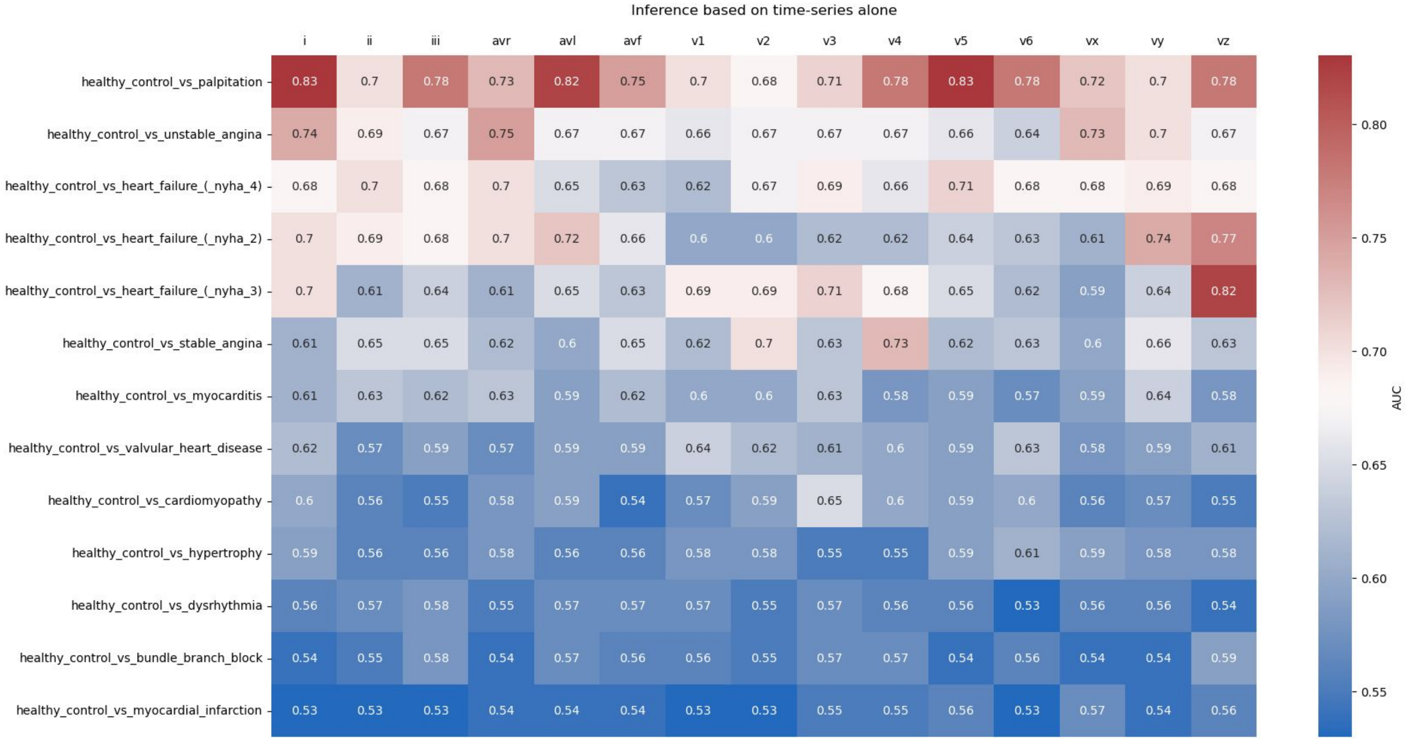
LightGBM



Leaf-Wise Tree Growth

I chose the Light Gradient Boosting Machine classifier, which a lighter version of XGBoost, currently the state-of-the-art in ML classification.

5 RESULTS



Inference based on time-series alone

	i	ii	iii	avr	avl	avf	v1	v2	v3	v4	v5	v6	vx	vy	vz
healthy_control_vs_palpitation	0.83	0.7	0.78	0.73	0.82	0.75	0.7	0.68	0.71	0.78	0.83	0.78	0.72	0.7	0.78
healthy_control_vs_unstable_angina	0.74	0.69	0.67	0.75	0.67	0.67	0.66	0.67	0.67	0.67	0.66	0.64	0.73	0.7	0.67
healthy_control_vs_heart_failure_(_nyha_4)	0.68	0.7	0.68	0.7	0.65	0.63	0.62	0.67	0.69	0.66	0.71	0.68	0.68	0.69	0.68
healthy_control_vs_heart_failure_(_nyha_2)	0.7	0.69	0.68	0.7	0.72	0.66	0.6	0.6	0.62	0.62	0.64	0.63	0.61	0.74	0.77
healthy_control_vs_heart_failure_(_nyha_3)	0.7	0.61	0.64	0.61	0.65	0.63	0.69	0.69	0.71	0.68	0.65	0.62	0.59	0.64	0.82
healthy_control_vs_stable_angina	0.61	0.65	0.65	0.62	0.6	0.65	0.62	0.7	0.63	0.73	0.62	0.63	0.6	0.66	0.63
healthy_control_vs_myocarditis	0.61	0.63	0.62	0.63	0.59	0.62	0.6	0.6	0.63	0.58	0.59	0.57	0.59	0.64	0.58
healthy_control_vs_valvular_heart_disease	0.62	0.57	0.59	0.57	0.59	0.59	0.64	0.62	0.61	0.6	0.59	0.63	0.58	0.59	0.61
healthy_control_vs_cardiomyopathy	0.6	0.56	0.55	0.58	0.59	0.54	0.57	0.59	0.65	0.6	0.59	0.6	0.56	0.57	0.55
healthy_control_vs_hypertrophy	0.59	0.56	0.56	0.58	0.56	0.56	0.58	0.58	0.55	0.55	0.59	0.61	0.59	0.58	0.58
healthy_control_vs_dysrhythmia	0.56	0.57	0.58	0.55	0.57	0.57	0.57	0.55	0.57	0.56	0.56	0.53	0.56	0.56	0.54
healthy_control_vs_bundle_branch_block	0.54	0.55	0.58	0.54	0.57	0.56	0.56	0.55	0.57	0.57	0.54	0.56	0.54	0.54	0.59
healthy_control_vs_myocardial_infarction	0.53	0.53	0.53	0.54	0.54	0.54	0.53	0.53	0.55	0.55	0.56	0.53	0.57	0.54	0.56

The model performs very well on distinguishing patients suffering from palpitation against healthy controls. The lowest performance is achieved for the myocardial infarction patients. Interestingly, not all electrodes provide the same predictive power. Additionally, even though we observed differences in the frequency domain for the “myocardial infarction” patients, the time-resolved model remains at chance level across all electrodes. This indicates that this simple approach is not sufficient and the model more likely underfits due to the high temporal bias. More elaborated models are required to achieve high classification performance.

6 NEXT STEPS

Explainable ECG classification with saliency maps (e.g: Le Nguyen, Thach, and Georgiana Ifrim. "A short tutorial for time series classification and explanation with MrSQM." *Software Impacts* 11 (2022): 100197.). In principle, the saliency maps should agree with the clinical practice.

(i)

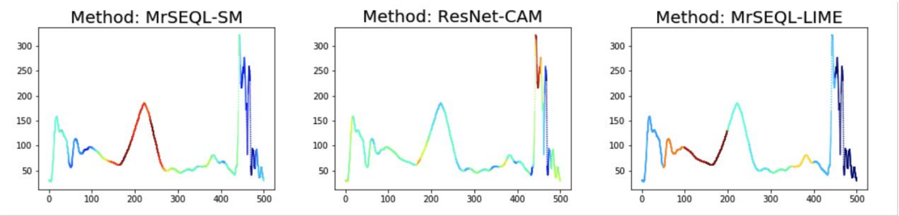
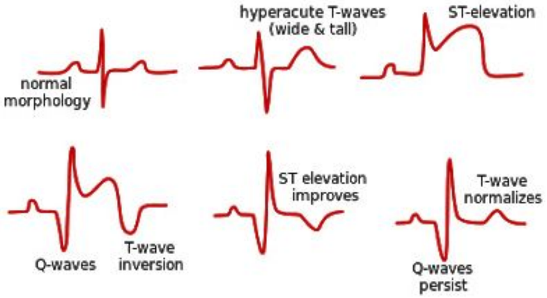


Fig. 1: Saliency map explanations for a motion time series obtained using different explanation methods. In this figure, the most discriminative parts are colored in deep red and the most non-discriminative parts are colored in deep blue.



(ii) Models more suited to handle temporal data (LSTMS)

(iii) “Epoching”. The data should be cropped around the “R” peak. Each segment is called an epoch. Then, different epochs will be averaged for each class. A time-resolved classifier can then be run on these average epochs. This will increase the statistical power of the analysis, and, importantly, will provide a time resolved picture on the segments that contribute the most to the difference (essentially building a saliency map but in a far less complicated way).

7 What I would study next to enhance my skills.

(i) @COURSERA

Browse > Data Science > Machine Learning

Offered By

University of Glasgow

Informed Clinical Decision Making using Deep Learning Specialization

Apply Deep Learning in Electronic Health Records. Understand the road path from data mining of clinical databases to clinical decision support systems

Fani Deligianni

(ii) PAPERS:

1. Faouzi, Johann. “Time Series Classification: A review of Algorithms and Implementations.” *Machine Learning (Emerging Trends and Applications)* (2022).

2. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. *Data mining and knowledge discovery*, *33*(4), 917-963.

3. Gupta, Varun, et al. “A review of different ECG classification/detection techniques for improved medical applications.” *International Journal of System Assurance Engineering and Management* (2022): 1-15.

4. Weimann, Kuba, and Tim OF Conrad. “Transfer learning for ECG classification.” *Scientific reports* 11.1 (2021): 1-12.