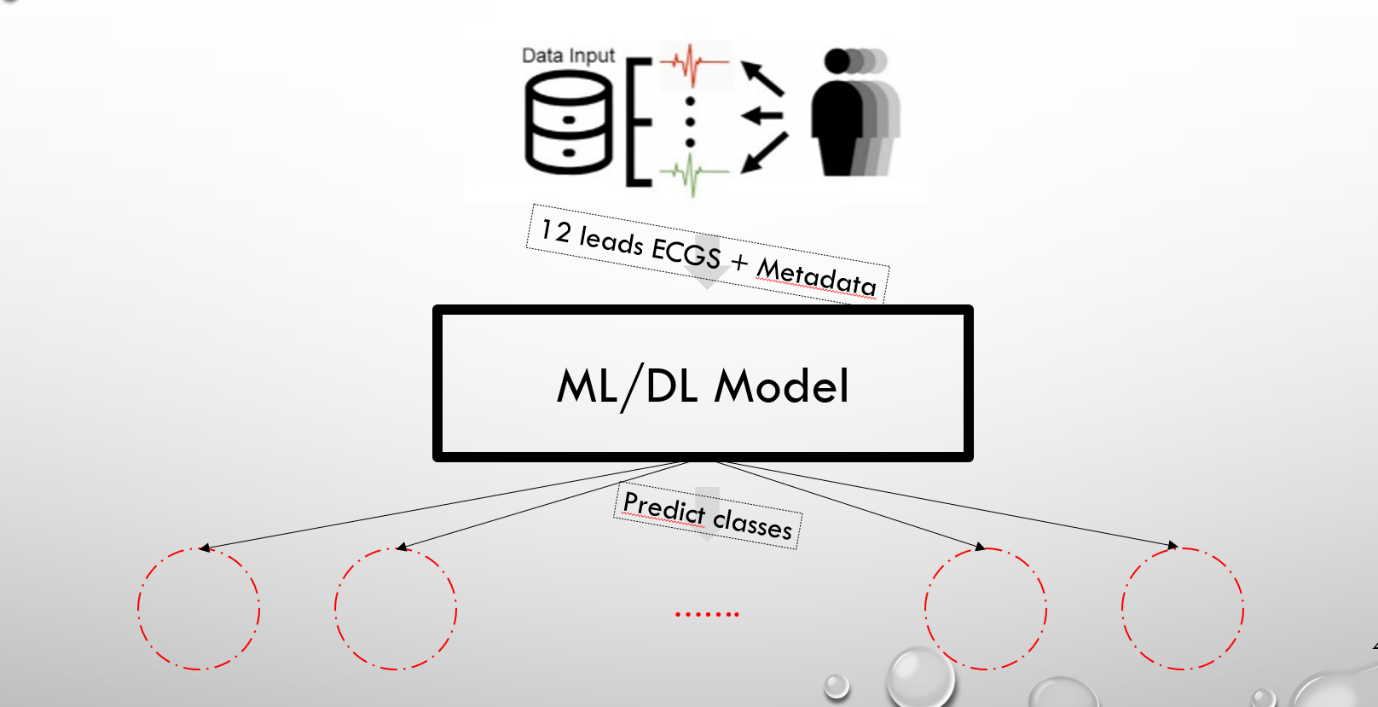
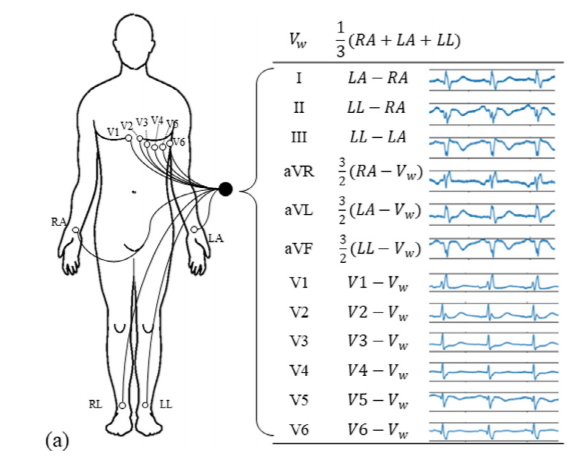
**AIM LAB REPORT : PhysioNet Challenge**

**Classification of 12-leads ECGs**

**Organisation of the Challenge:**

* 400+ teams from all around the world competing
* Duration of the Official Phase: 1st July – 23rd August
* Current Leaderboard: Best Team with a score of 0.666

**Aim of the Challenge:**

**Type of Data Input: 12-leads ECGs + Metadata (Age, Sex):**

**Detailed DataBase Composition:**

* 127 different pathologies
* 6 different Databases

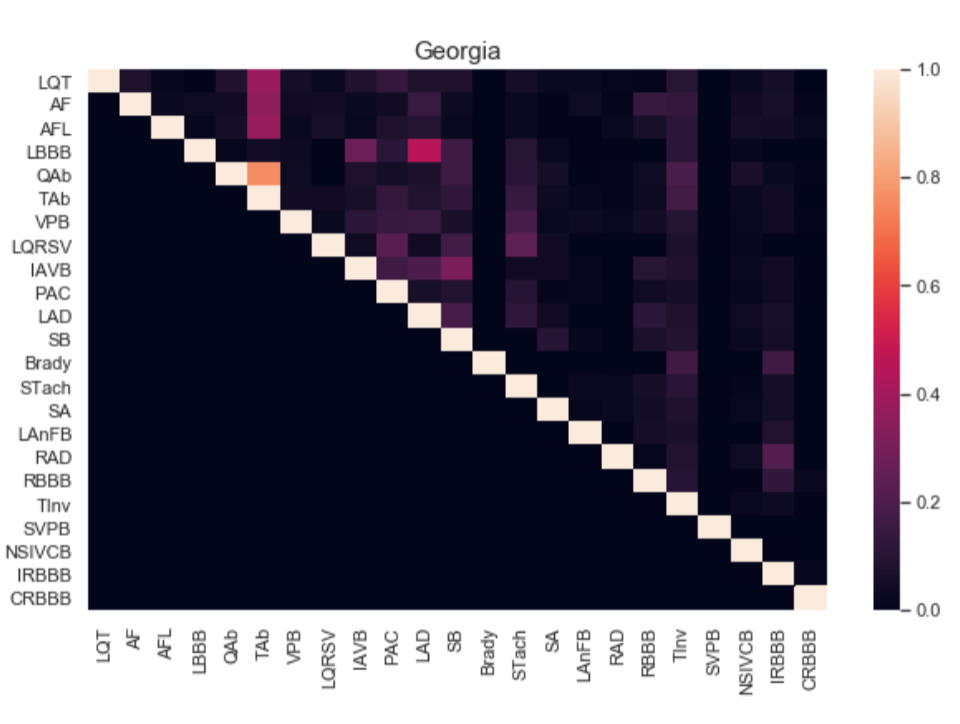
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Number of examples | Sample Rate (Hz) | Provenance | Multi-Labelled examples | Mean Length of an example |
| CPSCA | 6800 | 500 | China | Yes | 10s |
| CPSCB | 3400 | 500 | China | Yes | 10s |
| Georgia | 10200 | 500 | US | Yes | 15s |
| PTB | 500 | **1000** | Europe | Yes | 15s |
| PTBXL | 21000 | 500 | Europe | Yes | 15s |
| St Pet | 74 | **257** | Russia | Yes | **30min** |

|  |  |
| --- | --- |
| **Pathologies** | **Number of Examples**  **In the Competition DataBase** |
| AF | 2345 |
| AFL | 308 |
| Brady | 259 |
| IAVB | 1318 |
| IRBBB | 1221 |
| LAnFB | 1254 |
| LAD | 2126 |
| LBBB | 982 |
| LPR | 338 |
| LQRSV | 526 |
| LQT | 1090 |
| NSIVCB | 897 |
| PR | 299 |
| PAC | 1337 |
| PVC | 552 |
| QAb | 824 |
| RAD | 403 |
| RBBB | 2018 |
| SA | 1087 |
| SB | 1606 |
| SNR | 12019 |
| STach | 1555 |
| TAb | 1865 |
| Tinv | 832 |

**Low number of examples**

Total Number of Examples: ~40, 000 12 leads ECG (biggest open database)

**Multi-Labelled Examples, Georgia Database:**

**Remark:** We have seen no obvious consistency in the pathology co-occurences between DataBases. This is therefore **more complicated to insert prior knowledg**e in the simultaneous apparition of pathologies to our methods.

**Scoring Metrics:**

In our DataBases, there are roughly 120 pathologies. However, not all of them are scored: there are only **27 scored pathologies**. Among these 27 pathologies, there are groups of pathologies that are considered exactly the same by the scoring metrics:

* PVC and VPB
* PAC and SVPB
* CRBBB and RBBB

This means that misclassifying CRBBB as RBBB for example is not harmful at all, there is exactly no difference.

Therefore, there are **24 different scored pathologies (we re-labelled VPB as PVC, SVPB as PAC and CRBBB as RBBB).**

**Choice of Training**

Which pathologies do we insert on our training set ? Only the scored pathologies ? All the pathologies ?

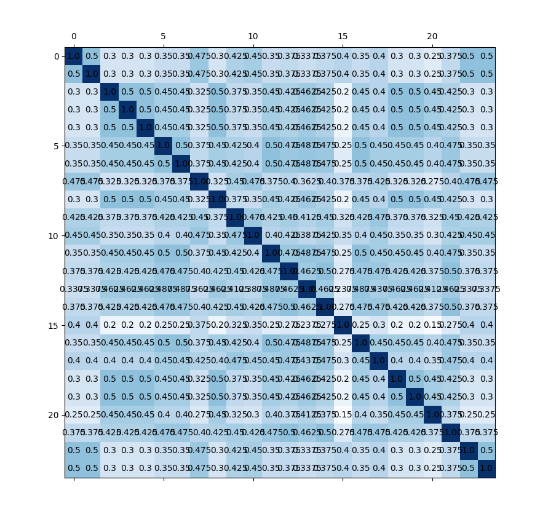
We chose not to insert the unscored pathologies in our training phase.

Indeed, the output of our classifier on ‘other’ pathologies does not count: **the classification on an unscored pathology is not considered in the scoring metrics.**

**However, the classification of a scored pathology is considered.**

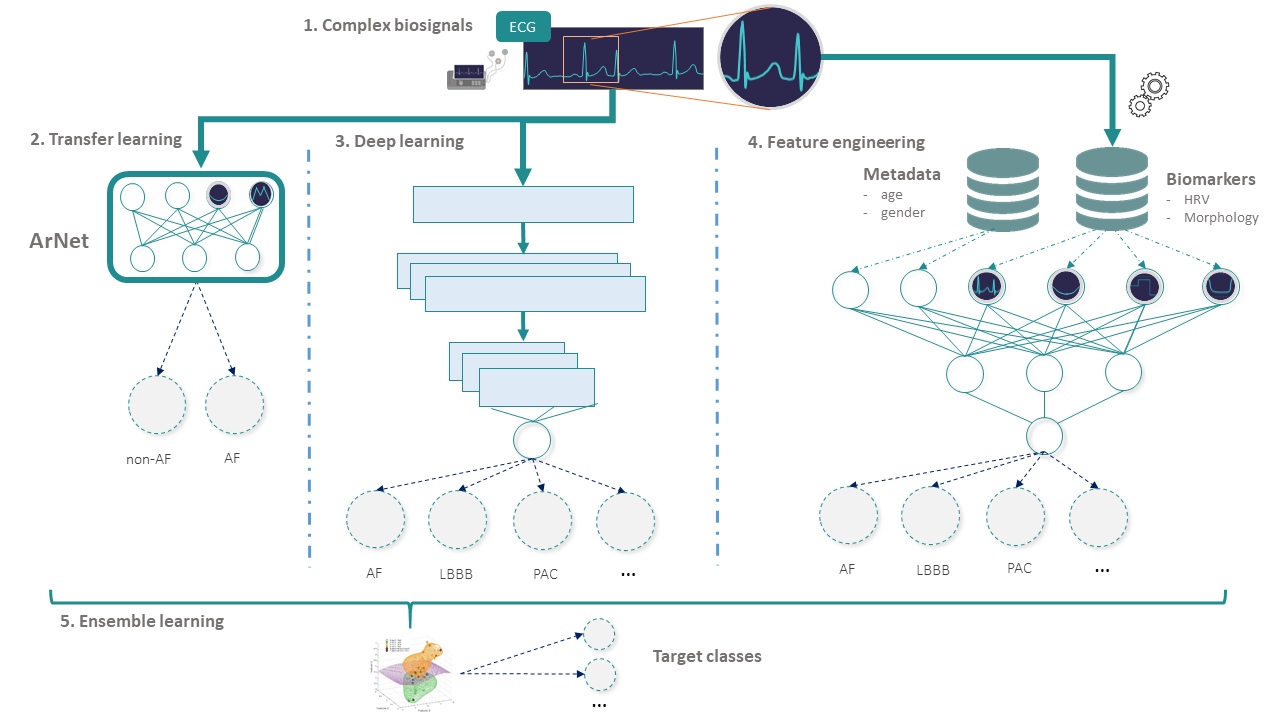
Therefore, we do not want our classifier to learn the existence of other pathologies in case it would classify a scored pathology as an unscored one: **we only consider examples with scored pathologies. This reduces our training DataBase down to ~36, 000 ECGs.**

**How to compute ?**

****

This is the weight matrix, which is weighting the misclassification between pathologies: some misclassifications are **more serious** than others: different from the unofficial fbeta score. This scoring metrics tries to reproduce the reality: some pathologies are often misclassified (because one is a subpathology of the other). Therefore, if our algorithm makes the same mistake, it’s “okay”.

How to compute? Basically, you compute the confusion matrix of your model and then you sum the scalar products columnwise with the columns of the weight matrix (therefore: high value in the weight matrix means that the misclassification between both pathologies is not very serious).

**Methods:**

David

Jeremy

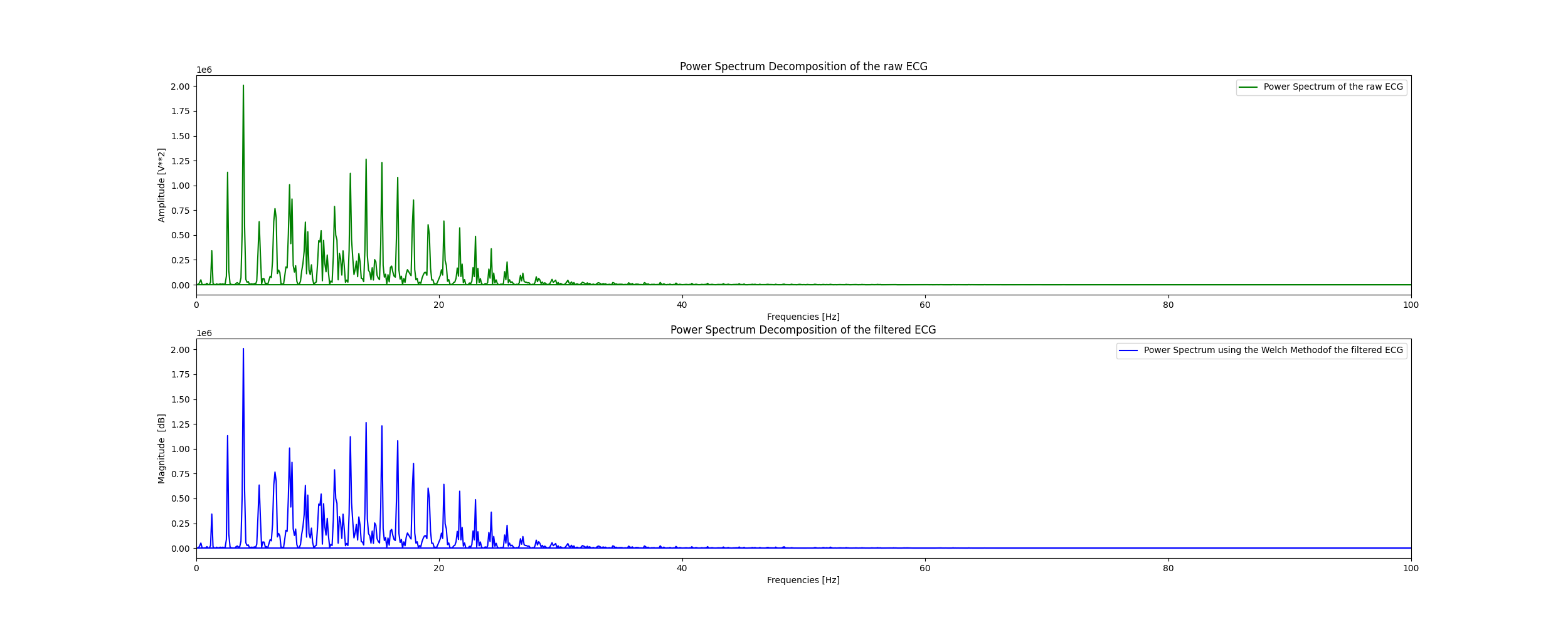
?

**Deep Learning Approach:**

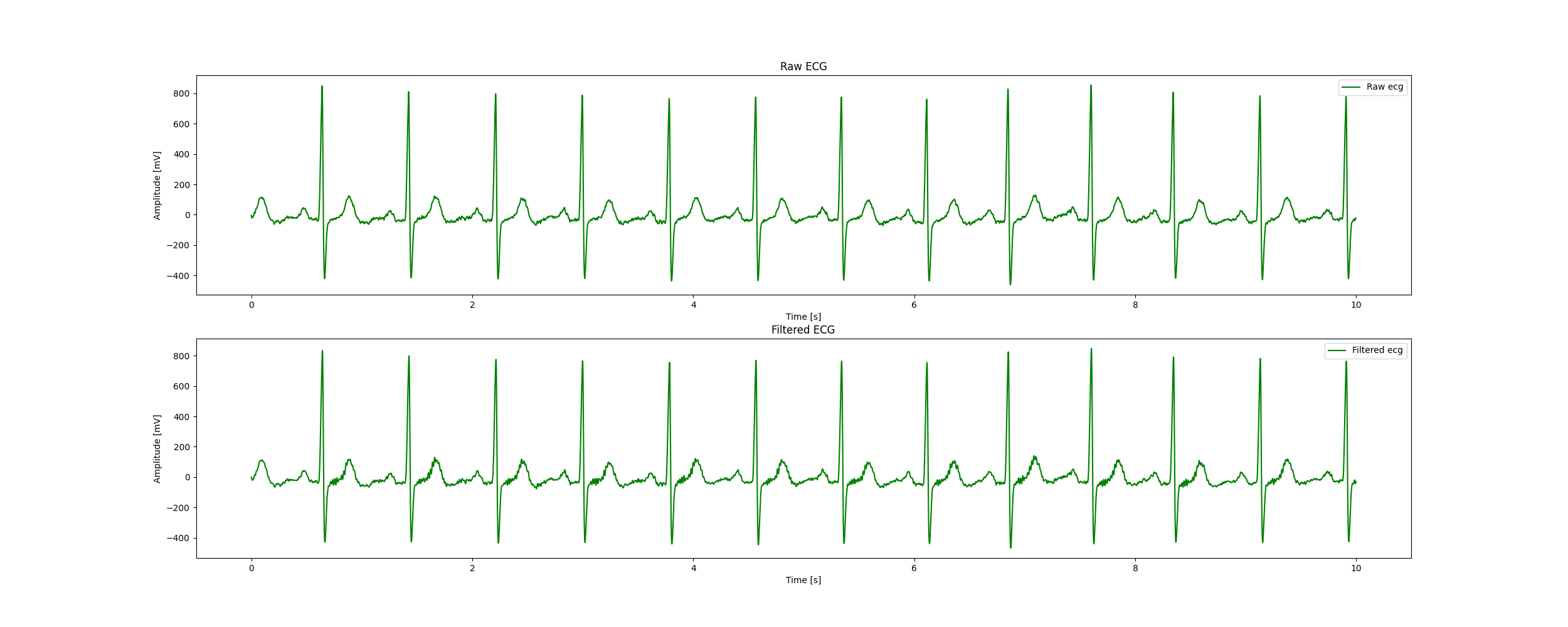
**Feature Engineering:**

**Signal Processing:**

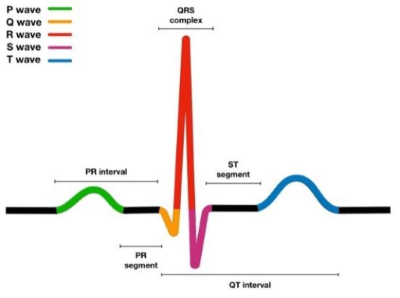
**Filter:** bandpass filter [0.05; 100] + notch [58; 62] + notch [48; 52]

**Frequency Domain : Power Spectrum Decomposition of the Raw and Filtered ECG**

We do not see any visual differences between both ecgs. Maybe our representation is wrong (I have also tried the log-domain representation, and did not have much better results).

**Time domain:**

**Feature Extraction:**

****

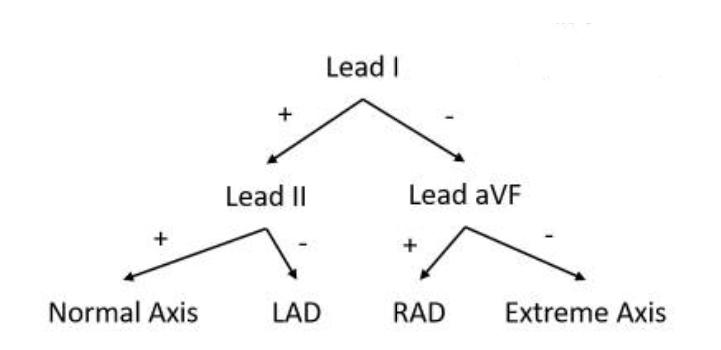
**Aim:** Extract relevant features in the ecg replicating the cardiologist approach.

**One Case Example: RAD: Right Axis Deviaton**

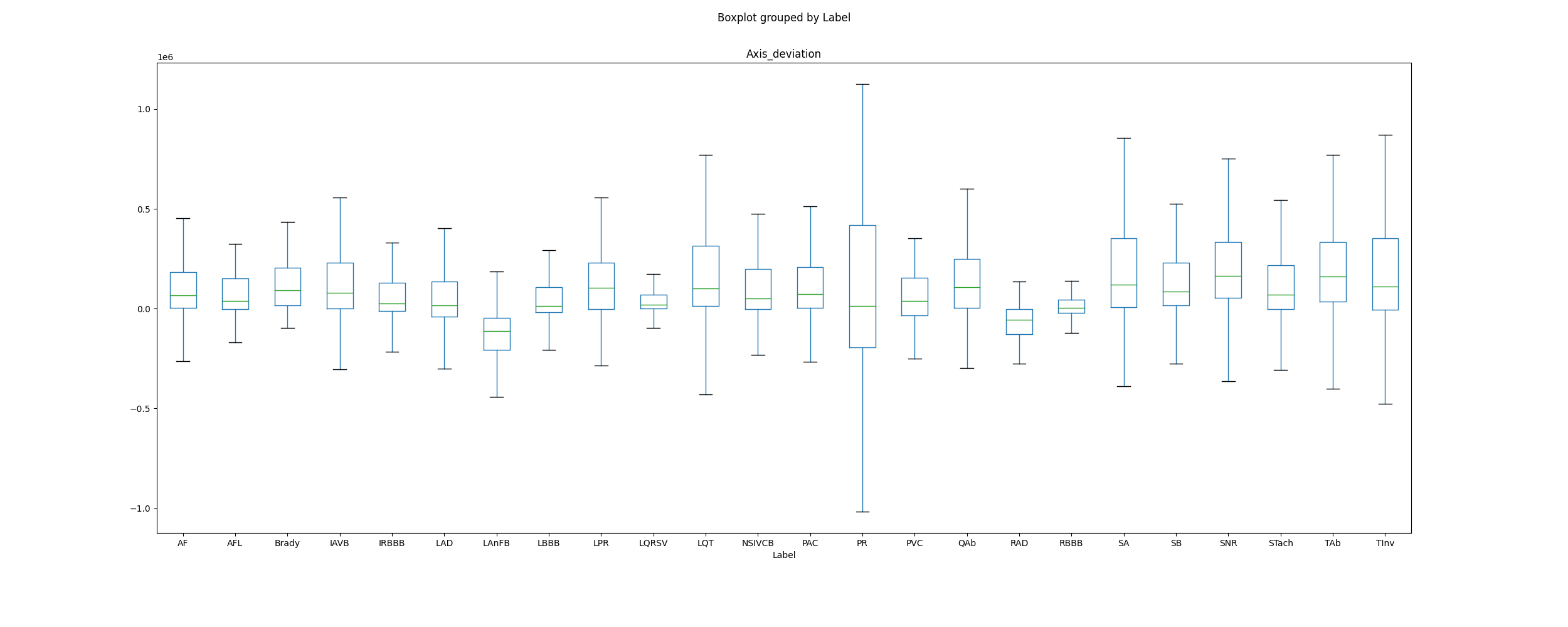
In electrocardiography, left axis deviation (LAD) is a condition wherein the mean electrical axis of ventricular contraction of the heart lies in a frontal plane direction between −30° and −90°. This is reflected by a QRS complex positive in lead I and negative in leads aVF and II. (Source: [wikipedia](https://en.wikipedia.org/wiki/Left_axis_deviation#:~:text=In%20electrocardiography%2C%20left%20axis%20deviation,in%20leads%20aVF%20and%20II.)).

The condition of LAD is usually defined by a QRS electrical axis and an age (the sane QRS electrical axis varies with the patient’s age).

Method in order to determine the nature of the QRS axis (source: [*ncbi*](https://www.ncbi.nlm.nih.gov/books/NBK470532/)):



We are going to extract these features for the classification of LAD and RAD. If our classification method is robust, we should perform relatively well because this decision tree structure particularly fits our Random Forest Classifier. The features I will extract for my Classifier are the value of net QRS deflection on leads I, II and aVF and their signs: 6 new features.

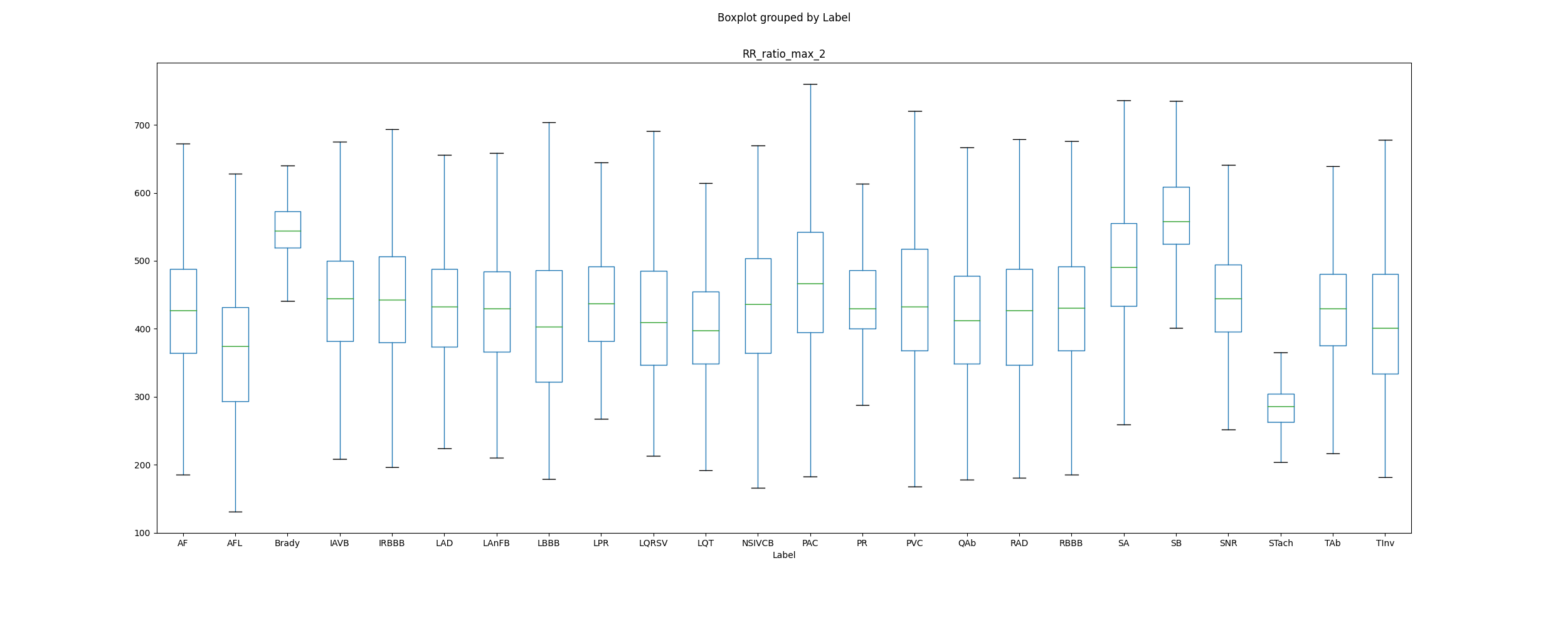
Here is the boxplot of: (net\_QRS\_deflection\_lead\_I)\* (net\_QRS\_deflection\_lead\_X) where X = II if the first sign is positive and X = aVF else. Therefore, we should have a negative sign for this feature for the conditions LAD and RAD, with different causes for LAD and RAD.

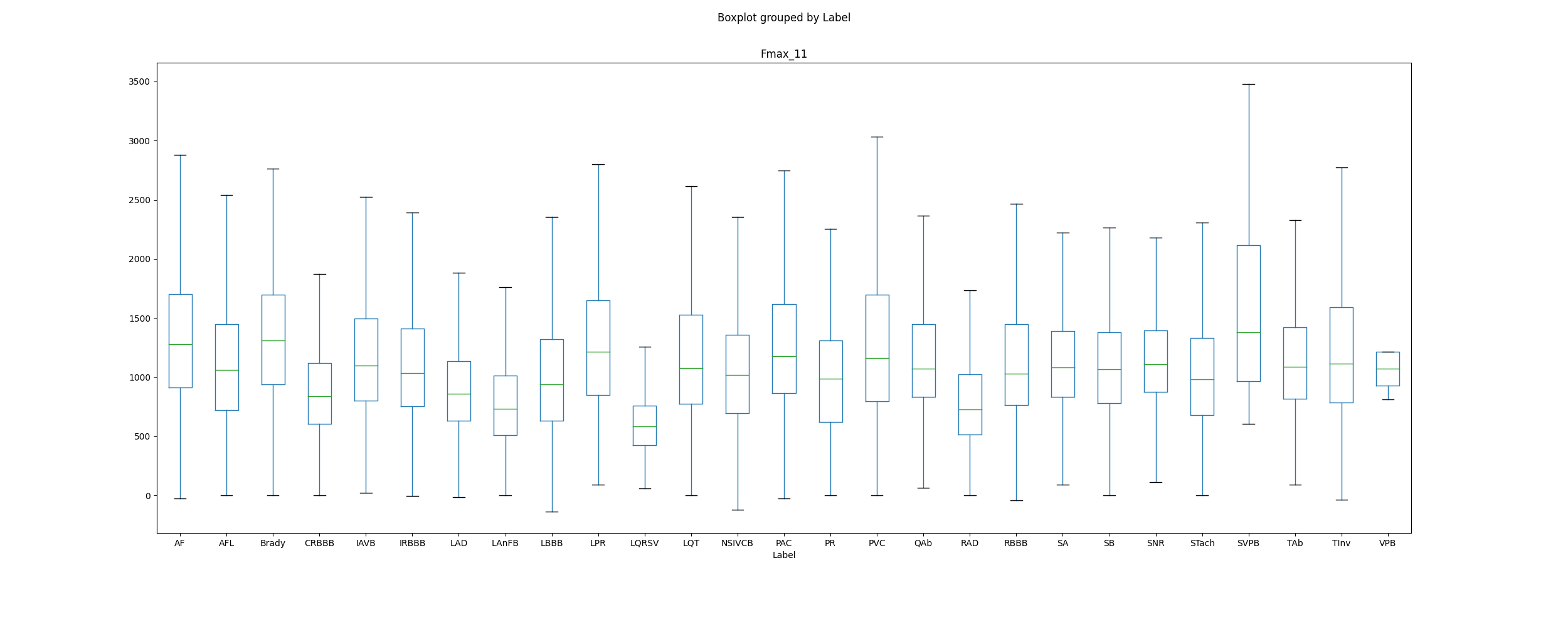
We can clearly see that this new feature helps to differentiate LAD, LAnFB and RAD from the other pathologies (LAnFB and LAD are the two lowest boxplots): it will help our classification. The values of the LAnFB are differentiated in the boxplot because an ECG characteristics of LAnFB is LAD.

For now, I did not yet extracted the features for some pathologies (to be done before the end of the week). Here is a recapitulative table of the features I extracted specifically for every patohlogy.

|  |  |  |
| --- | --- | --- |
| **Pathologies** | **Number of Examples**  **In the Competition DataBase** | **Number of Features extracted specifically for this pathology per lead** |
| AF | 2345 | 16 |
| AFL | 308 | 16 (the ones from AF) |
| Brady | 259 | 1 |
| IAVB | 1318 | 3 |
| IRBBB | 1221 | 0 |
| LAnFB | 1254 | 0 |
| LAD | 2126 | 0 |
| LBBB | 982 | 20 |
| LPR | 338 | 6 (3 AVB, 3 challenge) |
| LQRSV | 526 | 5 |
| LQT | 1090 | 5 (challenge) |
| NSIVCB | 897 | 0 |
| PR | 299 | 0 |
| PAC | 1337 | 12 |
| PVC | 552 | 26 |
| QAb | 824 | 0 |
| RAD | 403 | 0 |
| RBBB | 2018 | 20 (the ones from LBBB) |
| SA | 1087 | 0 |
| SB | 1606 | 0 |
| SNR | 12019 | 0 |
| STach | 1555 | 0 |
| TAb | 1865 | 0 |
| Tinv | 832 | 0 |

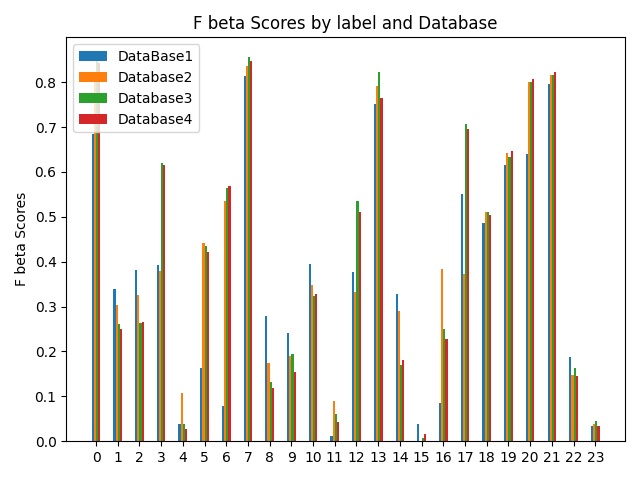
**Remark:** Some features extracted for some specific pathologies help to classify other pathologies.

**SA:**

**LQRSV:**

**ML Strategy:**

* **Data Standardization: Min-Max scaler**
* **Multi-Label classification: MultiLabel Binarizer + OneVsRest XGBoost Classification**
* **Fixed Train/Test split across the team**
* **No Hyperparameter Tuning (**default parameters include class weighting**)**

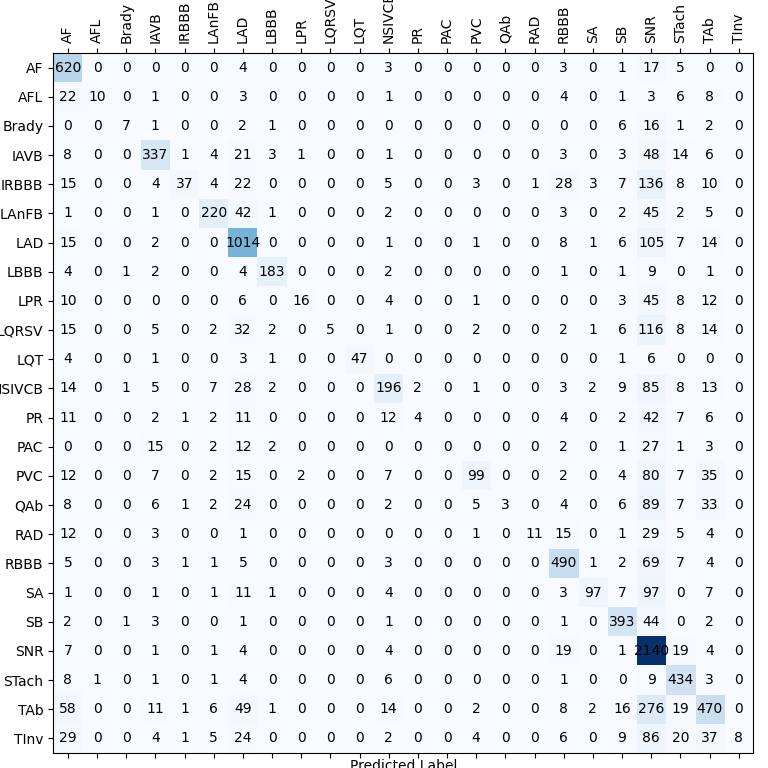
**Experiments using Increasing DataBases:**

**Results:**

**Fbeta scores (included former scores for Single Label Classification)**

|  |  |  |
| --- | --- | --- |
| **Pathologies** | **Fbeta score**  **Multi-Label Classification** | **Fbeta score**  **Single-Label Classification** |
| AF | 0.91 | 0.84 |
| AFL | 0.20 | 0.25 |
| Brady | 0.20 | 0.26 |
| IAVB | 0.72 | 0.61 |
| IRBBB | 0.14 | 0.02 |
| LAnFB | 0.60 | 0.42 |
| LAD | 0.72 | 0.57 |
| LBBB | 0.80 | 0.85 |
| LPR | 0.17 | 0.11 |
| LQRSV | 0.03 | 0.15 |
| LQT | 0.75 | 0.33 |
| NSIVCB | 0.60 | 0.04 |
| PR | 0.03 | 0.51 |
| PAC | 0.0 | 0.76 |
| PVC | 0.42 | 0.18 |
| QAb | 0.01 | 0.01 |
| RAD | 0.13 | 0.22 |
| RBBB | 0.80 | 0.70 |
| SA | 0.44 | 0.50 |
| SB | 0.80 | 0.64 |
| SNR | 0.80 | 0.80 |
| STach | 0.90 | 0.82 |
| TAb | 0.46 | 0.14 |
| Tinv | 0.04 | 0.03 |

**Confusion Matrix:**



**Competition score: 0.62**

**Main Source of Mistakes:** Misclassification of some examples as mainly represented classes (LAD/AF/**SNR**).

**Data Augmentation:**

4 Databases:

**Paper:** *A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients, JianweiZheng, JianmingZhang, Sidy Danioko, HaiYao, HangyuanGuo Cyril Rakovski*

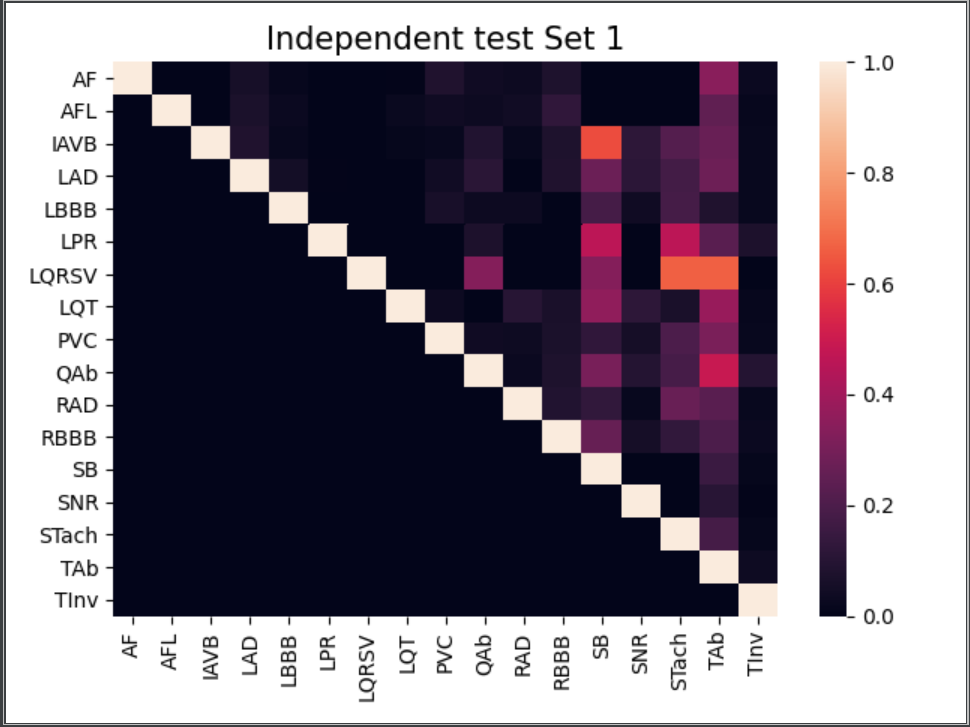
**Raw description:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Provenance** | **Number of Examples** | **Mean length of an example** | **Previous filtering (code availability)** | **Multi-Label** | **Sample Rate** |
| **China** | **10646** | **10s** | **Yes** | **Yes** | **500** |

**DataBase composition:**

|  |  |
| --- | --- |
| SB | 3836 |
| TAb | 1869 |
| SNR | 1826 |
| AF | 1754 |
| STach | 1532 |
| RBBB | 437 |
| AFL | 441 |
| LAD | 380 |
| PVC | 306 |
| PAC | 277 |
| IAVB | 246 |
| QAb | 233 |
| RAD | 221 |
| TInv | 157 |
| LBBB | 93 |
| LQT | 57 |
| LPR | 13 |
| LQRSV | 3 |

**Pathologies Co-occurences**

****

**Paper:** *A 12-Lead ECG database to identify origins of idiopathic ventricular arrhythmia containing 334 patients JianweiZheng, Guohua Fu, KyleAnderson, HuiminChu Cyril Rakovski*

**Raw description:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Provenance** | **Number of Examples** | **Mean length of an example** | **Previous filtering (code availability if yes)** | **Multi-Label** | **Sample Rate** |
| **China** | **334** |  | **Yes/Yes** | **No** | **2000** |

**DataBase composition:**

|  |  |
| --- | --- |
| PVC | 325 |
| VT (Ventricular tachycardia) = ‘other’ | 9 |

**Paper:** *Lobachevsky University Electrocardiography Database*

**Raw description:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Provenance** | **Number of Examples** | **Mean length of an example** | **Previous filtering (code availability if yes)** | **Multi-Label** | **Sample Rate** |
| **Russia** | **243** | **10s** | **No** | **No** | **500** |

**DataBase composition:**

|  |  |
| --- | --- |
| SNR | 143 |
| STach | 4 |
| SB | 25 |
| LAD | 66 |
| RAD | 3 |
| IAVB | 10 |
| IRBBB | 29 |
| RBBB | 4 |
| LBBB | 4 |
| NSIVCB | 4 |

**Paper:** *Automatic diagnosis of the 12-lead ECG using a deep neural network*

**Raw description:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Provenance** | **Number of Examples** | **Mean length of an example** | **Previous filtering (code availability if yes)** | **Multi-Label** | **Sample Rate** |
| **Brazil** | **200** | **10s** | **No** | **No** | **300-600Hz** |

**DataBase composition:**

|  |  |
| --- | --- |
| IAVB | 28 |
| RBBB | 34 |
| LBBB | 30 |
| SB | 16 |
| AF | 13 |

**Summary Data Augmentation**

|  |  |
| --- | --- |
| SB | 3861 |
| TAb | 1869 |
| SNR | 1969 |
| AF | 1767 |
| STach | 1536 |
| RBBB | 475 |
| AFL | 441 |
| LAD | 446 |
| PVC | 631 |
| PAC | 277 |
| IAVB | 284 |
| QAb | 233 |
| RAD | 224 |
| TInv | 157 |
| LBBB | 123 |
| LQT | 57 |
| LPR | 13 |
| LQRSV | 3 |
| IRBBB | 29 |
| NSIVCB | 4 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pathologies** | **Number of Examples**  **In the Competition DataBase** | **Number of Features extracted specifically for this pathology per lead** | **Fbeta score**  **Multi-Label Classification** | **Fbeta score**  **Single-Label Classification** | **Opportunities for Data Augmentation** |
| AF | 2345 | 16 | 0.91 | 0.84 | 1767 |
| AFL | 308 | 16 (the ones from AF) | 0.20 | 0.25 | 441 |
| Brady | 259 | 1 | 0.20 | 0.26 |  |
| IAVB | 1318 | 3 | 0.72 | 0.61 | 284 |
| IRBBB | 1221 | 0 | 0.14 | 0.02 | 29 |
| LAnFB | 1254 | 0 | 0.60 | 0.42 |  |
| LAD | 2126 | 0 | 0.72 | 0.57 | 446 |
| LBBB | 982 | 20 | 0.80 | 0.85 | 123 |
| LPR | 338 | 6 (3 AVB, 3 challenge) | 0.17 | 0.11 | 13 |
| LQRSV | 526 | 5 | 0.03 | 0.15 | 3 |
| LQT | 1090 | 5 (challenge) | 0.75 | 0.33 | 57 |
| NSIVCB | 897 | 0 | 0.60 | 0.04 | 4 |
| PR | 299 | 0 | 0.03 | 0.51 |  |
| PAC | 1337 | 12 | 0.0 | 0.76 | 277 |
| PVC | 552 | 26 | 0.42 | 0.18 | 631 |
| QAb | 824 | 0 | 0.01 | 0.01 | 233 |
| RAD | 403 | 0 | 0.13 | 0.22 | 224 |
| RBBB | 2018 | 20 (the ones from LBBB) | 0.80 | 0.70 | 475 |
| SA | 1087 | 0 | 0.44 | 0.50 |  |
| SB | 1606 | 0 | 0.80 | 0.64 | 3861 |
| SNR | 12019 | 0 | 0.80 | 0.80 | 1989 |
| STach | 1555 | 0 | 0.90 | 0.82 | 1536 |
| TAb | 1865 | 0 | 0.46 | 0.14 | 1869 |
| Tinv | 832 | 0 | 0.04 | 0.03 | 157 |

**Recapitulative table**

Pathologies for which the score is satisfying: AF, LBBB, RBBB, SBR, SB, STach

Pathologies that have consequently benefitted from the Multi-Label Classification: IAVB, LAnFB, LAD, LQT, NSIVCB, PVC, SB, TAb, Stach, RBBB

Pathologies that have consequently suffered from the Multi-Label Classification: LQRSV, PR, PAC

Pathologies fo which Data-Augmentation is a valuable option: AFL, Brady, LPR, LQRSV, PR, PVC, RAD, TInv (pathologies for which I will first extract new features before Augmenting Data)

Pathologies for which I will extract new features: AFL (in order to separate with AF), Brady, IRBBB, LAnFB, LAD, NSIVCB, PR, QAb, RAD, TAb, SA, TInv (pathologies for which I have already found interesting features).