

# Applied AI in Biomedicine

## Final Assignment

Simone Colombara - 962646  
Alessia Cotroneo - 963083  
Francesco De Caro - 976699  
Raffaella D'Anna - 963651  
Anna Iob - 963336



**POLITECNICO**  
**MILANO 1863**

# 1. Introduction

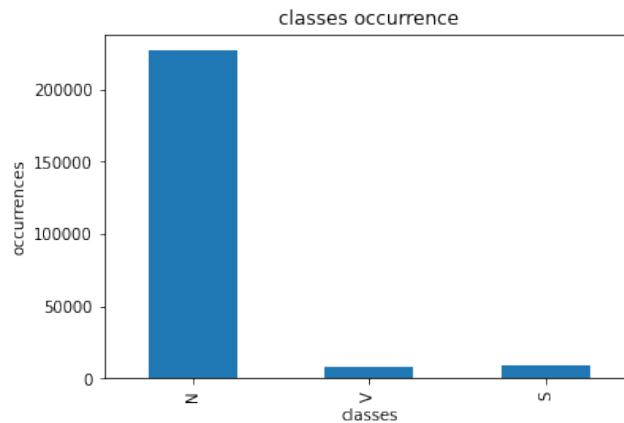
Given a training dataset concerning 2-lead ECG signals taken from 105 patients at 2 different frequencies (128 and 250 Hz), we had to train a model to classify R-peaks, i.e. labeling them as N (normal sinus rhythm beats), V (ventricular beats, typical of Premature Ventricular Complex) or S (supraventricular beats, typical of Premature Atrial Complex).

## 2. Materials and Methods

### 2.1 Dataset Analysis

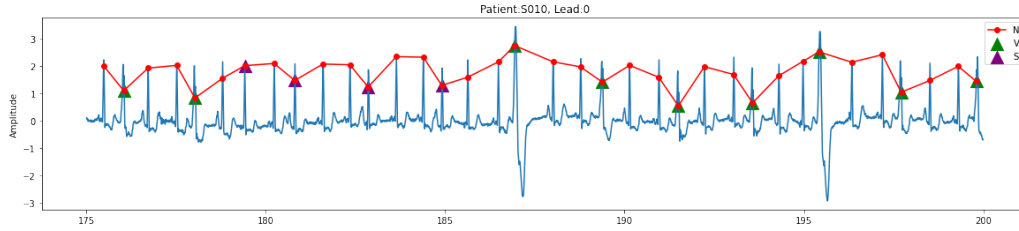
#### 2.1.1 Initial visualization

- By a preliminary dataset analysis focused on the labels, it is clear that, for each patient, N labels occur much more often than V and S.



labels distribution of the whole dataset

- One of the main aspects to manage was the poor R-peaks labeling, particularly evident when plotting the ECGs.



ECG plot of patient S010 from 175 to 200 seconds - The red line connects the R-peaks

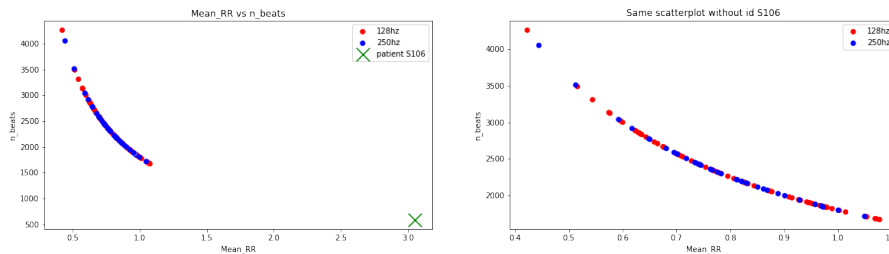
### 2.1.2 Feature extraction

To further explore and analyze the data, the following features were extracted for every patient from the initial ECGs, through an ad hoc function:

- mean\_RR: average distance between peaks
- median\_RR: median of the distances between peaks
- var\_RR: variance of the distance between peaks
- std\_RR: variance square root
- n\_beats: number of R peaks registered in the ECG, which is assumed to be the actual number of heartbeats measured

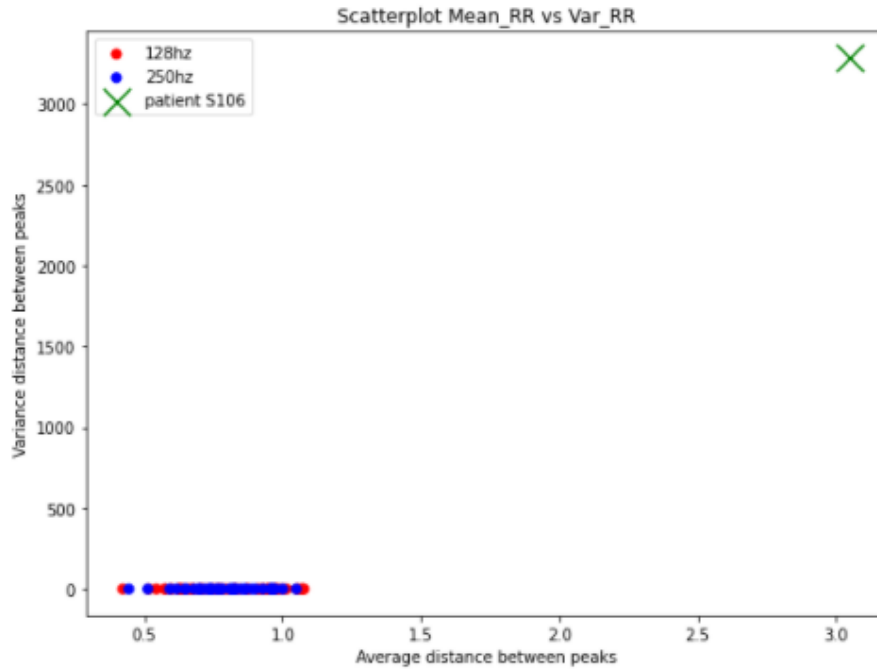
All these features but n\_beats were standardized with respect to the sampling frequency of the ECG.

Possible correlations were studied analyzing appropriate plots. From the following scatterplots (studying the correlation between mean\_RR and n\_beats), it is clear that patient S106 is an outlier.

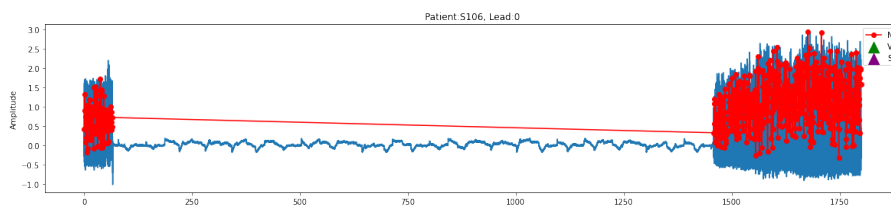


A clear dependence between the number of beats and mean\_RR is also present, as expected.

Further plots regarding other features, such as the following (studying a dependence between mean\_RR and var\_RR), highlighted once again that patient S106 is an outlier, also in terms of variance.



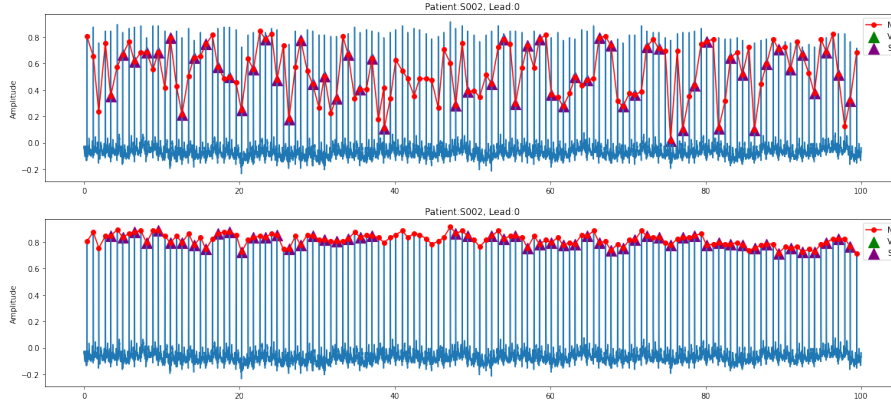
Investigating more on the strange behaviour of this outlier, it was found that the majority of its ECG is actually not recorded properly as shown in the following plot.



Conducting this analysis on different patients, lead the same results.

### 2.1.3 R-peaks repositioning

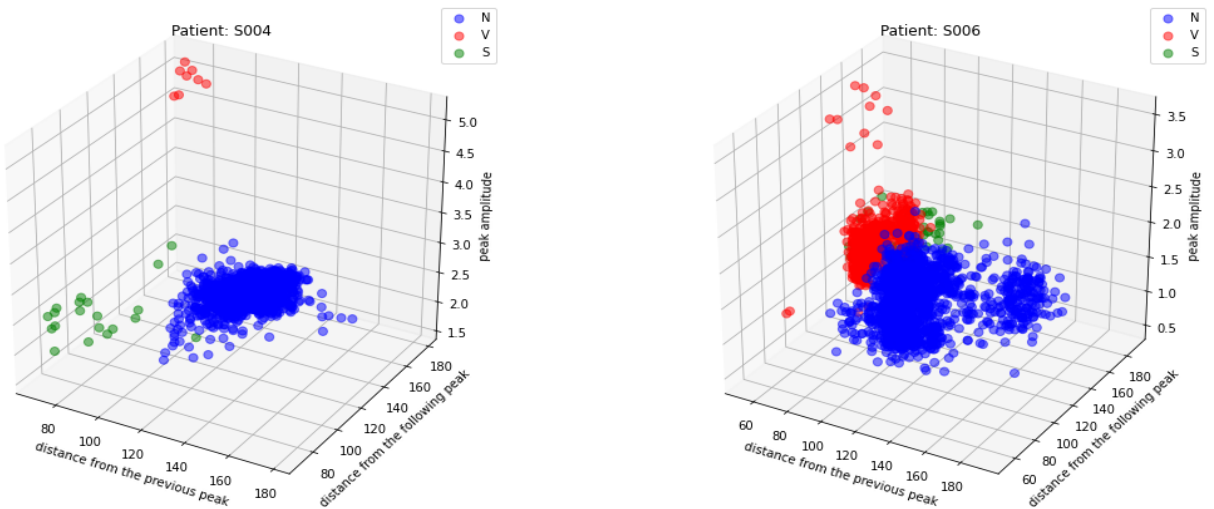
To deal with the R-peaks repositioning problem, some functions were introduced to adjust the samples. The results were satisfying, as shown in the following plot, which highlights the improvement between the original peaks and the repositioned ones.



After peaks repositioning, all the features mentioned above were extracted again, in order to check for changes, and it seemed appropriate to consider two additional local features, namely:

- the distance of each heartbeat from the previous and next ones.
- the amplitude of a single heartbeat

Considering the distance mentioned above, seems relevant in order to understand whether the single heartbeat considered is normal (N) or not (not N). Indeed, both ventricular beats and supraventricular beats are anomalies in the heart rhythm. On the other hand, considering the amplitude of the single heartbeat can be relevant in distinguishing the two anomalies S and V as an "excessive amplitude" is typical of ventricular (V) beats.

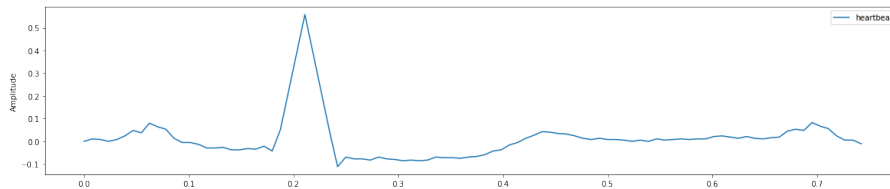


These 2 patients' plots show that, in fact, there is already some sort of clustering in the data.

## 2.2 Classification - preliminary considerations

### 2.2.1 Beat visualization

Before proceeding with classification, the signals were preprocessed with a band-pass filter in order to remove low and high-frequency noises. An algorithm to extract only ECGs' portions related to single heartbeats was built in order to address the labels classification problem.



Feature extraction was performed again, in order to focus on local ECG's characteristics of a single beat after the band-pass filtering. Along with the previous features, new local ones, necessary to build the model, were extracted, i.e.:

- min: minimum amplitude value in an ECG beat
- diff: difference between maximum (R-peak) and minimum amplitude
- area\_under\_the\_curve: approximated area underneath the beat

It was also necessary to perform a features preprocessing by introducing a robust scaler, given the presence of some outliers.

## 2.3 Classification models

### 2.3.1 Grey box models

First of all, a "grey box" model was implemented to take advantage of the domain knowledge, represented by the global and local features previously extracted from the dataset, and to implement a deep learning model that is particularly suited when large quantities of data are available, as in this case.

The result is that the model takes in input not only the ECG, but also a table of "hand made" and meaningful features that help the algorithm in its purpose.

### 2.3.2 Multi-class classification

The starting idea was to train a multi-class classification model, but the unbalanced dataset led to unsatisfying results, making it difficult to distinguish V and S peaks. The problem was then split in two steps:

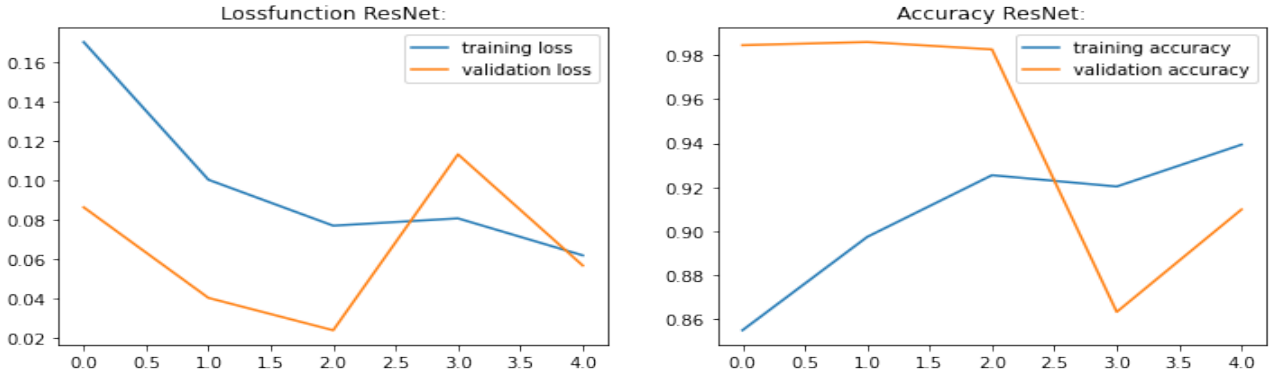
- First, building a model to classify peaks as N or not-N.
- Secondly, trying to deal with the problem of classifying not-N peaks into V or S ones.

Doing so, the task was easier for the network and good performances were reached on the first model.

### 2.3.3 Model description

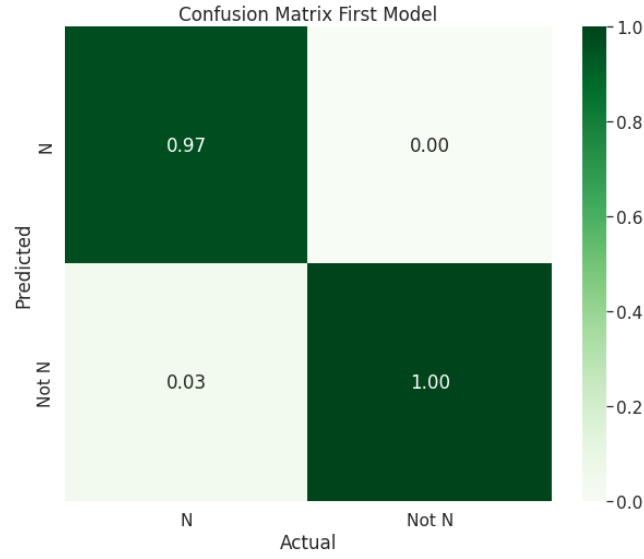
The choice of splitting the problem in two steps was justified by the fact that, in this way, the second model worked with a balanced dataset, obtained from the first one. The first model is a Res Net neural network. This model was chosen because it allows to build a deep network relying on a small number of parameters w.r.t. other networks (i.e. VGG16). It was trained using a weighted cross-entropy loss function, since it was a classification problem, and 90% of the training set was made of normal beats. The second model is instead an ensemble of a Res Net neural network, a KNN and a random forest (the last two models classification is only based on the features). This choice was driven by the fact that the performance of the models above of their own was quite poor, so the natural solution was to try all of them together. Moreover, the contribution of each model is weighted according to their accuracy. It was decided to avoid training two different models for each sampling frequency, since it would have been too time consuming and computationally expensive.

## 3. Results

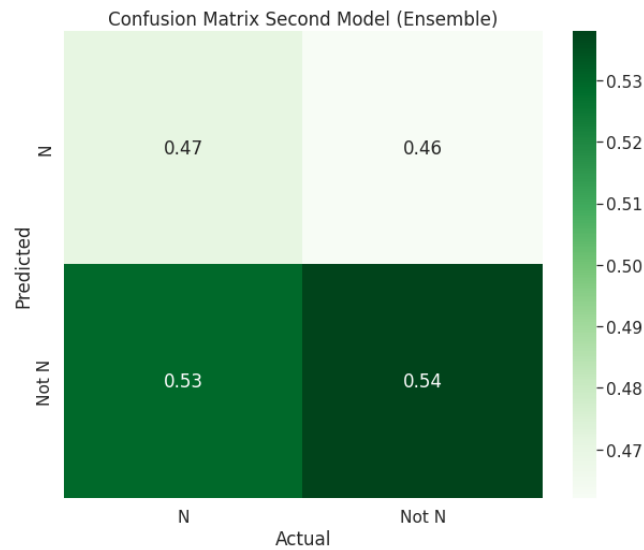


first model

The third checkpoint was saved as it was the one with lowest validation loss value. Only 5 epochs were trained in order to save computational power.



The first model, as clearly shown in the confusion matrix, performs very well in the test set.



The second model, as shown in the above confusion matrix, does not perform well. In particular, it seems that none of the models of the ensemble learned to distinguish between S and V labels. One reason for this outcome could be the various issues of the dataset. Indeed, the difference between S and V wasn't that clear even in the original labeling. Moreover, the overall performance of our model may have been influenced by the collinearity of the dataset, since the global features of the beats of each patient don't change.



## 4. Discussion and Conclusion

The model seems good in identifying anomalous heartbeats, so we can conclude that using a grey box model seems a good choice when dealing with these kind of medical problems. Indeed, it is of paramount importance to include some medical knowledge to our model together with the data in order to have more valuable and intelligible results. However, it's also reasonable to think that some of the N beats will be misclassified as not N, since we introduced a loss function in the model that doesn't penalize much these kind of errors. This is acceptable, given our initial goal, which is to prioritize anomalous beats over normal ones. Indeed, we decided to give more importance to detecting potentially sick patients, rather than erroneously classifying sick patients as healthy ones, which was our main goal.

# Bibliography

- [1] <https://drive.google.com/drive/folders/1e1g0htpfBXe-CP1woGaLWOnp2wiX8ze5>
- [2] *Interpretazione dell'elettrocardiogramma*, Derek J Rowlands, Pharma