- Applied AI in Biomedicine Workshop Report - Development of the HTF Deep-Learning Heartbeat Classifier

Alberto Rota
Person Code: 10615751
Student Number: 964662
alberto2.rota@mail.polimi.it

Gabriele Santicchi
Person Code: 10579046
Student Number: 969088
gabriele.santicchi@mail.polimi.it

I. Context

The goal of this work is to build a classification model that properly annotates ECG heartbeats as Normal, PAC (supraventricular beats) or PVC (ventricular beats). The available dataset is composed of 2-lead ECG signal and R peaks position of 105 patient recordings.

II. Introduction

Electrocardiogram (ECG) signals records the electrical activity of the human hearts and consist of several waveforms (P, QRS, and T). The duration and shape of each waveform and the distances between different peaks are used to diagnose Cardiovascular Heart Diseases (CVDs). Premature Atrial Contractions (PACs) and Premature Ventricular Contractions (PVCs) are among the most common forms of arrhythmias; the first results from premature electrical activation originating in the atria of the heart, while the second one is caused by a premature electrical activation originating in the ventricles. As the presence of frequent PACs and PVCs is associated with a higher risk of unfavorable prognosis [3] [2], the accurate recognition of these abnormal beats is rigorously required. The aim of this work is to design a deep-learning ensemble model that properly classifies all the R-peaks in an ECG recording, without requiring the manual extraction of ECG features. The *History-Time-Frequency* classifier analyzes each heartbeat in both the time and frequency domains, and take as input also the labels assigned to the previous two beats.

A. Dataset

The dataset consists of 105 2-leads ECGs of 30 minutes recorded from 105 patients with two different sampling frequencies: more specifically, 65 patients were recorded with $f_s = 128Hz$ and the remaining 40 patients with $f_s = 250Hz$. The recordings are annotated with the position of the R-peaks and their ground truth classification. An example is provided in *Figure 1*.

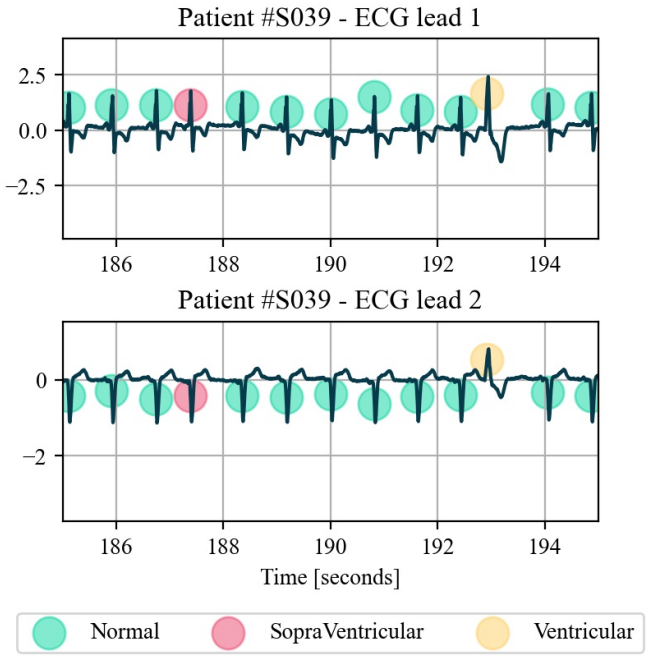

Fig. 1. Annotated ECGs from a sample patient

III. Data Preprocessing

The available raw dataset undergoes a pre-processing phase necessary for the optimal interfacing between the data and the model.

A. Exploratory Data Analysis

First and foremost, the acquisition frequencies are made uniform in the whole dataset by undersampling the recording acquired at $250Hz$: this is done in order to guarantee that the model inputs have the same size independently of the sampling frequency.

No missing data is found in the recordings, hence there is no need to handle this kind of issue.

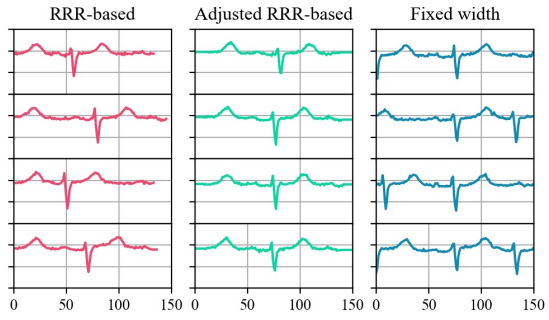

Fig. 2. Comparison of window splitting strategies. The *RRR-based* (left) clearly shows inconsistency in the width; the *fixed-width* (right) displays how adjacent peaks are included in the window; the *adjusted RRR-based* (center) shows, in cases like the uppermost window, the padding used to match the dimensions

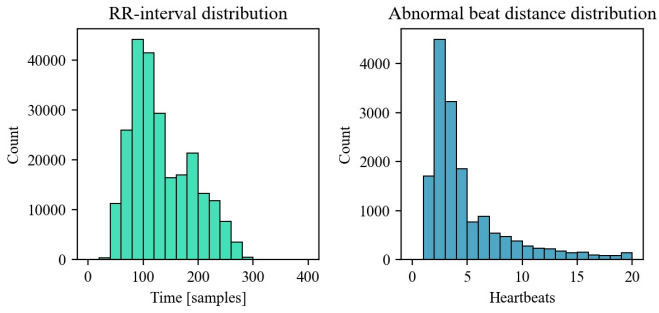

Fig. 3. *Left*: Distribution of the RR-interval in samples. *Right*: Distribution of the distances between abnormal heartbeats. For example, most abnormal heartbeats (over 4500) are located 2 beats after the previous abnormal heartbeat

An in-depth inspection of the dataset revealed, still, mismatches in the annotations of the R-peak positions, as for patient *S005* and *S009* the last heartbeat in the recording is labelled with a timestamp bigger than the signal length. In a conservative approach, given the very high number of heartbeats in the whole dataset (circa 244336), the two mislabelled R-peaks are removed.

B. Signal Windowing

The heartbeat-by-heartbeat classification that the model performs implies the necessity of splitting the 30 minutes recording of each patient in windows, each one to be processed separately by the network: choosing the proper width and the optimal windowing approach is key in obtaining the best model performance. Three different approaches have been taken into consideration when splitting the windows:

- **RRR-based:** Each heartbeat is considered as starting from the R-peak of the previous one and ending to the R-peak of the following one (with some margin, 5% of the window width by default, in order to not include the QR and RS slopes of the adjacent beats). This splitting strategy keeps the highest possible amount of information inside the window but, crucially, produces windows of inconsistent width, which depends on the patient heart-rate.
- **Fixed width:** A predefined width w (in samples) is chosen and each window is considered as starting from $x_{peak} - w/2$ to $x_{peak} + w/2$. This approach assures a consistent window size and, additionally, makes sure that the R-peak is centered in the window. For patients with a high heart-rate though, adjacent R-peaks may be unavoidably included in the window.
- **Adjusted RRR-based:** This approach combines the two previously described strategies by initially performing the *RRR-splitting* and later, after choosing a window width like in the *fixed-width* approach, padding or cropping to the desired width. This approach makes sure that a single R-peak is contained in the window, and it assures a uniform width among all the splittings: to this extent, the informative content of the window is reduced by the least possible amount.

Figure 2 shows the effect of the above strategies. The *adjusted RRR-based* is the employed windowing method, with the fixed width set at $w = 150$ samples, a value chosen based a qualitative analysis of the RR-interval distribution reported in *Figure 3*.

C. Class imbalance

As the following table highlights, the 244334 windows extracted (one for each available R-peak/heartbeat) are considerably unbalanced among the three labels of interest

	Normal	Sopraventricular	Ventricular
Count	226930	9609	7795
Fraction	92.87%	3.93%	3.19%

If such an unbalanced dataset would be used to train the model, learning would happen only for the overrepresented class. As a consequence, the underrepresented pathological heartbeats would not be confidently classified: such model would have a very high specificity and a very low sensitivity. In order to cope with this kind of issue, an undersampling of the most frequent class has been performed to balance the class distribution. To do so, $\frac{7795}{105}$ Normal heartbeats have been randomly extracted from the ECG of each patient, obtaining a total of 7795 Normal peaks, a number equal to the least represented among the classes.

IV. The HTF Model

The proposed model is a *History-Time-Frequency* ensemble (HTF), as the classification is performed in parallel in the Time domain and in the Frequency domain, while also accounting for History in terms of the labels of the past heartbeats in the ECG recording.

A. H as History

During a first inspection phase, an analysis on the distribution of pathological beats has been performed. In *Figure 3*, the

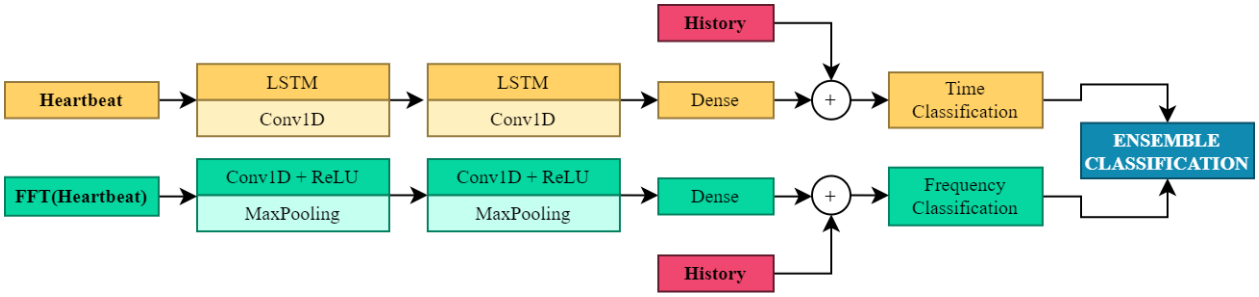

Fig. 4. Schematic of the History-Time-Frequency model ensemble, respectively color-coded in magenta, yellow and green

histogram reports the distribution of the abnormal interpeak distance computed over all the recordings, confirming that PACs and PVCs often occur in repeated patterns [4] [1]. For these reasons, the involvement of the labels assigned to the previous peaks could help in predicting more precisely the label of the current peak. Considering the interpeak distance distributions, in this classifier the two previous labels have been added as additional inputs to the model.

B. T as Time

One branch of the classifier consists of a recurrent architecture working on the ECG in time domain. Here, Long Short-Term Memories (LSTMs) in combination with 1D convolutions are employed to extract the most relevant features of the ECG morphology while accounting for their temporal dependencies. The heartbeat history is concatenated to the network prior to the last Dense+Softmax layer, so that the final labelling decision also accounts for the past two labels: the concatenation operation prior to the final classifier layer allows the model to learn how to combine the information about the current peak with the information related to the history.

C. F as Frequency

The last branch of this model consists in a frequency-domain classifier on a Fourier Transform of the inputs. In order to apply the Fourier Transform, the *quasi*-stationarity hypothesis has been assumed for each window. The classifier consist of multiple CNN-ReLU-MaxPooling stacks. A small convolutional kernel has been chosen to extract features at a reduced scale. After the application of the FFT on the input signal, the output of the CNN layer is concatenated with the two history labels like in the time model. Finally, a dense layer with a softmax activation function provides the probability assigned to the three classes.

The full ensemble model is schematically represented in *Figure 4*.

V. Training

The time and frequency models have been trained separately and then ensembled by summing the class probability vectors. The training is performed by minimizing a Categorical Crossentropy loss function with the ADAM optimizer. The training phase of the time model is set up for 100 epochs, while the frequency model fits over thrice the epochs as it has been observed converging at a slower rate. Batch size is of 256 heartbeats, learning rate starts from $1 \cdot 10^{-3}$ and decreases on eventual loss plateaus. Training is monitored on accuracy and on a recall metric calculated per class, as follows:

$$Recall_k = \frac{True_k}{True_k + \sum_{i=0}^{classes} False_i} \tag{1}$$

This metric allows to evaluate if the class imbalance issue has been correctly resolved, as an imbalanced dataset will reflect on a low recall for the underrepresented classes. Training histories are reported on *Figure 6*.

VI. Results

Results have been evaluated in terms of the model precision, recall and f1-score on the 3 classes separately. The test set used for such evaluation consists of 10% of the available dataset.

Label	Precision	Recall	f1-score	Support
Normal	0.98	0.96	0.97	699
Sopraventricular	0.96	0.98	0.97	865
Ventricular	0.98	0.98	0.98	702
Per-Class Accuracy	0.97	0.97	0.97	2266
Weighted Accuracy	0.97	0.97	0.97	2266
Overall Accuracy				**0.97**

The same results are graphically represented in the form of a confusion matrix in *Figure 5*, this time differentiating the results of the time model, the frequency model and the ensemble.

VII. Discussion

VIII. Conclusions

References

[1] Boon-Hor Chong, Vincent Pong, Kwok-Fai Lam, Shasha Liu, Ming-Liang Zuo, Yuk-Fai Lau, Chu-Pak Lau, Hung-Fat Tse, and Chung-Wah Siu. Frequent premature atrial complexes predict new occurrence of atrial fibrillation and adverse cardiovascular events. *EP Europace*, 14(7):942–947, 12 2011.

[2] Hideo Hirose, Shizukiyo Ishikawa, Tadao Gotoh, Tomoyuki Kabutoya, Kazunori Kayaba, and Eiji Kajii. Cardiac mortality of premature ventricular complexes in healthy people in japan. *Journal of Cardiology*, 56(1):23–26, 2010.

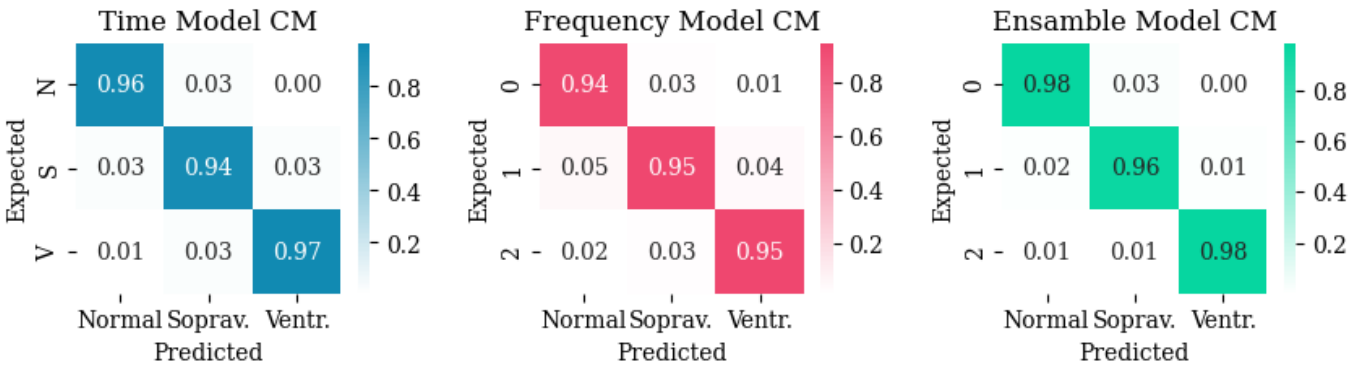

Fig. 5.

[3] Bao-tao Huang, Fang-yang Huang, Yong Peng, Yan-biao Liao, Fei Chen, Tian-li Xia, Xiao-bo Pu, and Mao Chen. Relation of premature atrial complexes with stroke and death: Systematic review and meta-analysis. *Clinical Cardiology*, 40(11):962–969, 2017.

[4] Khashayar Farzam; John R. Richards. Premature ventricular contraction. *StatPearls*, 2021.

APPENDIX A
TRAINING PROCESS

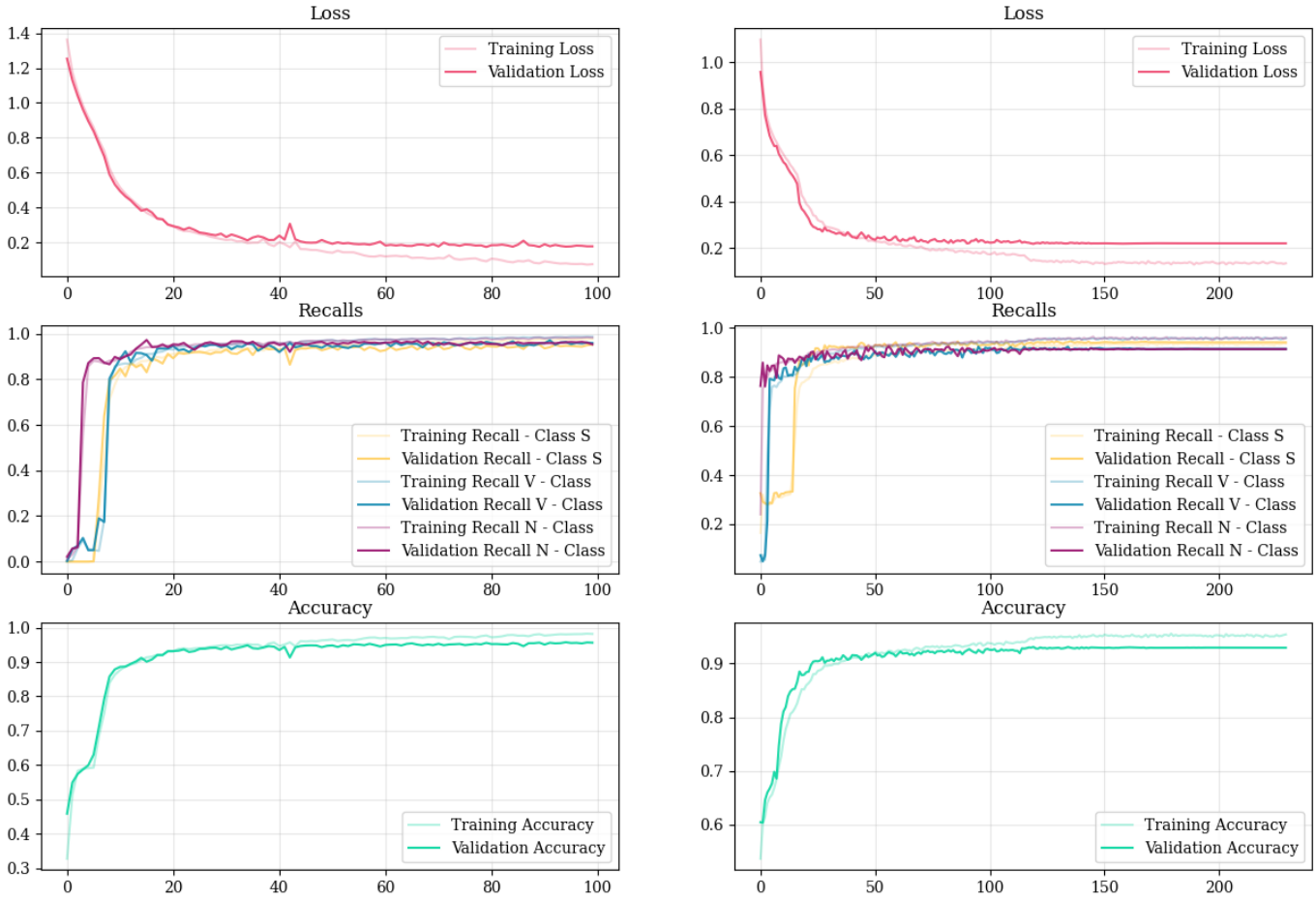
Loss
Training Loss
Validation Loss
Recalls
Training Recall - Class S
Validation Recall - Class S
Training Recall V - Class
Validation Recall V - Class
Training Recall N - Class
Validation Recall N - Class
Accuracy
Training Accuracy
Validation Accuracy
Loss
Training Loss
Validation Loss
Recalls
Training Recall - Class S
Validation Recall - Class S
Training Recall V - Class
Validation Recall V - Class
Training Recall N - Class
Validation Recall N - Class
Accuracy
Training Accuracy
Validation Accuracy

Fig. 6.