Classifying Atrial Fibrillation in Electrocardiograms with Machine Learning

Caio Victor Gouveia Freitas
Elektro- und Informationstechnik
Technische Universität Darmstadt
Darmstadt, Germany
caio.freitas@stud.tu-darmstadt.de

Guilherme Fernandes G. Silva
Department of Information and Communication Engineering
Technische Universität Darmstadt
Darmstadt, Germany
guilherme.fernandes@stud.tu-darmstadt.de

Pablo Gómez Hidalgo
Elektro- und Informationstechnik (of Aff.)
Technische Universität Darmstadt
Darmstadt, Germany
pablo.gomez.hidalgo@stud.tu-darmstadt.de

***Index Terms*—Electrocardiogram, LGBM, Machine Learning, Atrial Fibrillation**

***Abstract*—With the rise of machine learning in medicine, diagnostics stands out one of the most interesting and most common applications. In this project, it is presented a platform that aims to predict atrial fibrillation in electrocardiogram signals. It is explored the pre-processing of the raw data with filtering, artifact removal, and normalization as well as extract metrics used in heart rate variability analysis (time and frequency domain). Besides that, the imbalanced data problem is also addressed, and the model selection is briefly discussed.**

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally [1]. From those, atrial fibrillation (AF) plays an important role, with a worldwide prevalence of atrial fibrillation is 37.574 million cases (0.51% of worldwide population in 2020) [10].

The goal of this project is to create an algorithm capable of detecting AF in electrocardiogram (ECG) signals. In this context, it is vital to achieve rapid, reliable, and safe diagnostics to prevent against consequential harmful diseases.

A. Foundations

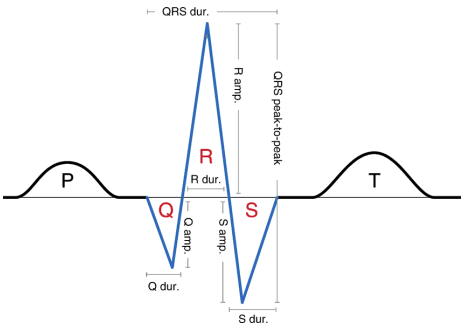

Fig. 1. QRS Complex [12]

The ECG is structured in a PQRST pattern (Figure 1), where the P wave corresponds to the filling of the atria with blood. The P-Q segment represents the travel time of the signal between the atria and the ventricles. The next step is the depolarization of the ventricles, reflected as the QRS complex. Then, in the S-T segment the ventricles contract and pump blood. Finally, the repolarization of the ventricles is shown by the T wave.

Types of heart rhythms depend on the origin of the initial electrical impulse. The sinus rhythm, initiated in the sinus node, is the normal, non-pathological. The aim of this project is to discuss the AF rhythm in which the initial electrical impulse is initiated in multiple foci in the cardiac atria, resulting in its abnormal contraction.

B. Atrial Fibrillation

This pathology generates a irregular rhythm and consequently a disordered depolarization of the atria. The irregularity of the beat is due to multiple attempts to depolarise the atrium but only some are able to pass to the AV node and depolarise the ventricles. This results in an irregularity of the rhythm. As the impulses from the atria become insufficient to pump blood to the ventricles clots may be formed, that can be ejected throughout the body and cause infarctions, or cause strokes when arriving the cerebral circulation.

To visualize AF in the electrocardiogram, two visible characteristics can be observed.

1) Absence of P wave, this is due to uncoordinated depolarisation of the atria.
2) Irregularity in heart frequency. Due to the randomness of the atrial impulses the ventricular contractions and the appearance of QRS complexes become irregular.

II. Methodology

To be able to create an algorithm to classify signals with AF, machine learning can be of great help, considering the great advancements and robustness of supervised learning methods for classification and clustering. Python has a great community and algorithms that can be used to design a solution for specific problems; thus it is the programming language considered for developing the procedures.

A. Overview

The data set provided for this project contains 6000 signals, all with the same sampling frequency of $300Hz$ but with

different lengths. The majority have around 30 seconds of data recorded in all classes; the classes are:

- Normal (N) - normal ECG signal, without AF.
- Atrial Fibrillation (A) - signal with AF
- Other signal (O) - not ECG signal
- Noise (˜) - Noise signal

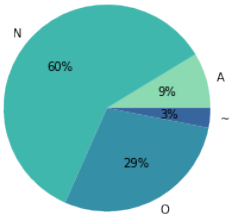

Fig. 2. Distribution of classes in training data

One of the problems regarding the data is the fact that the classes are unequally distributed [5]. This needs to be considered to avoid biasing the model during training and when using accuracy as an evaluation metric, not to misinterpret the results. Therefore, the multi-label and binary F1-Score metrics will be used, which considers both the precision and recall perspectives.

Considering that our data is a biological time series signal, one could say that there are two possible approaches. The first would consider the usage of Neural Networks, because is a strong method that can process the data in raw form, consequently without losing information. Although in practice this creates a black-box, because the decision process of the model cannot be easily interpretable, which is an important factor in the medicine environment. Yet another disadvantage is that this approach requires lots of data, in practical terms, proportional to the square value of the number of parameters [15] which was a determinant factor considering the available data set for this project.

The second approach consists of extensive processing of the ECG to be able to extract essential information to the recognition of AF. That option can be easily interpretable and repliable for a given ECG signal.

Given the whole scenario, the second approach was chosen. The Figure 3 contains the proposed methodology and transformation steps in the data.

B. Preprocessing

This step starts with a Noise Removal process, where the raw data is cleaned from known sources of noise, such as electrical interference from other devices or noise from the electronics measurements. Then, high-amplitude noise and artifacts are removed (such as baseline wander [9]). Finally, an amplitude normalization is performed for better results in the training process.

1) Filtering: For filtering, a bandpass filter was chosen, which according to the literature [7], [9], [11] is a standard

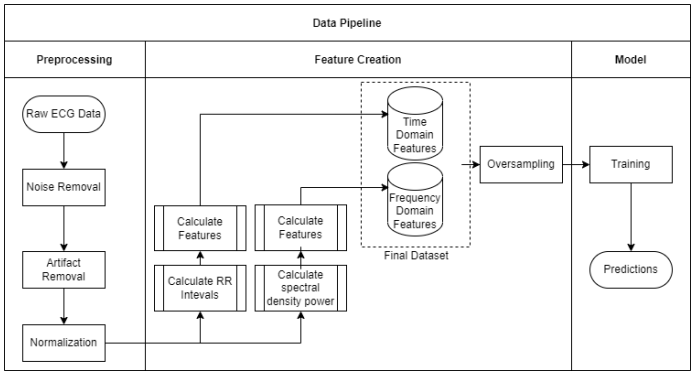

Fig. 3. Data Pipeline Block Diagram

method, with good performance and computationally cheap [9], with cutoff frequencies of 3 and 45 Hz. After trying different values for the band limits, the ones proposed in [7] showed better performance in our data.

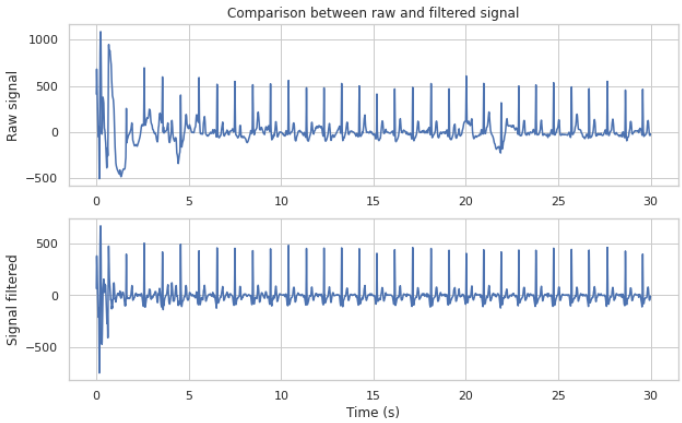

Fig. 4. Comparison between raw and filtered signal

2) Artifact Removal: Artifacts in the ECG context are defined as alterations in the signal that are not related to cardiac electrical activity. Some of the most common alterations are motion artifacts, misplacement of electrodes, ventricular tachycardia *et cetera* [14]. They can alter important characteristics of the signal such as the baseline. The baseline wander is a low-frequency artifact usually caused by subject movement [14], that can resemble atrial fibrillation [9] - hence our interest in removing this specific kind of distortion.

The allegedly best performing techniques for baseline wander cancellation include the wavelet-based, moving median and subtraction, and Butterworth high-pass filters, respectively. From these, the Butterworth filter was chosen for being computationally cheap (with little loss of performance) and widely spread across medical applications [9].

3) Normalization: In order to facilitate the model training, a normalization was applied to get a zero mean and unity standard deviation, removing the DC offset and disconsidering the amplitude variance of the signals, as suggested in the work of [4], [17].

C. Feature Engineering

RR peaks interval is the time between two heartbeats (consecutive R-peaks) in the QRS-Complex; this signal is non-stationary and may contain indicators of heart diseases, because can indicate irregularity in the heart frequency.

There were strong efforts on research of how to extract the R-peaks from the electrocardiogram signal. The most used method is the Pan and Tompkins algorithm [13] used for real-time ECG, the method consists of differentiate the signal, highlighting the larger variations between two data points squared, accentuating the R peaks. Although the signals face a filtering process, it is not possible to extract information about the presence of the P wave with reliability.

According to [6], it is possible to extract time domain, frequency domain and non linear features to the HRV analysis. In this project were considered time and frequency domain features.

The time domain features are the most easy to be extracted, because its calculation is straight forward on the data and are the metrics are easily replicable.

- **SDNN (ms)**: standart deviation (SD) of RR intervals series.
- **SDSD (ms)**: SD of differences between adjacent RR intervals.
- **pNN50 (ms)**: proportion of adjacent RR intervals differing by more than 50 ms.
- **RMSSD (ms)**: the square root of the mean of the squares of differences between adjacent RR intervals.

Besides those, given the relation between atrial fibrillation and heart insufficiency (which can be seen by the increase in the patient's heart rate due to the lower cardiac efficiency) [16], features involving the Heart Rate were also considered. In essence the average, median, maximum and minimal heart rate were also added to the features.

According to [2], the spectral power analysis of heart rate series provide quantitative information about the function of the cardiovascular system. It breaks the frequency spectrum in three components: Very low frequency (VLF, ¡ 0.03Hz), low frequency (LF, 0.03-0.15 Hz) and high frequency component (HF, 0.15-0.4 Hz). Each one can provide certain information regarding the nervous system. Thus, also the spectral power in those bands and ratios between them were added to the feature space.

D. Imbalanced Data

The data imbalancement has a major negative impact on the model performance by performing particularly bad on minority classes [8]. To deal with it, a number of strategies are available; SMOTE is a strategy that oversamples the object of minority class by generating synthetic samples in the feature space (taking each minority class sample and introducing synthetic examples along the line segments joining the k minority class nearest neighbors) [5], and has shown to improve the accuracy scores (more precisely the area under the ROC curve) for the minority classes.

E. Model

The approach for model selection involves [3], an open-source library in Python that automates machine learning workflows and model management tool that speeds up experiments in the data. This technology makes model selection faster because the library sorts the best model considering the chosen metrics. In all of the experiments the best model evaluated by PyCaret was tree based models, specifically Extra-tree Classifier and Light Gradient Boosting Machine (LGBM). The latter was the model selected given the previous experience with this technology in Python.

III. Results and Discussion

Working with time series data brings several challenges, specially regarding data analysis, exploration, and understanding the singularities of the signals to create a model that can generalize it. In addition to that, to create a procedure that would handle all the cases and particularities was also challenging.

Upon getting familiar with the data, it was found an imbalance between classes representation in the data set and between the lengths of the records. A simple process that helped avoiding mistakes while evaluating the model results and understanding the pattern in the signals.

In practical terms, extracting information of each signal is the most important thing. This task was done by extracting the R peaks interval for each signal, that again results in a series for each signal, which can be complex to analyse for the amount of data. The results of all major iterations are presented in the table below, and will be explained in detail up next.

TABLE I
Evolution of the classification score

Iteration	Multi-label Accuracy	Binary Accuracy
1	0.7142	0.9515
2	0.7084	0.9532
3	0.8034	0.9743
4	0.9028	0.982

A. First iteration

The first attempt of predictor was based purely in Heart Rate Variability HRV metrics [6] - based in the RR peaks values. At this stage, there was still no filtering process or artifact removal.

In this trial, a data pipeline (Figure 3) was created to feed the raw data into the model in tabular format. After generating the tabular data set with some time domain features, PyCaret was used for comparing multiple models. Between the models ran in this analysis, the one that presented the best multi-label F1 score was the Light Gradient Boosting Machine (LGBM).

There was no testing for binary classification with PyCaret in this process, thus LGBM was the model chosen based on the multi-label experiment. The model was tested applying cross-validation with a splitting strategy of 5 folds, the results were initially satisfactory, and through the confusion matrix it

was verified that most of the errors were between the classes N and O, A and O and the model was not being able to classify correct the noisy data in general. As we can see in the confusion matrix below.

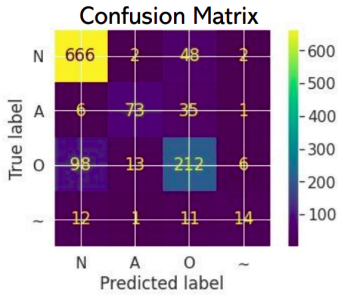

Fig. 5. Confusion matrix of the first iteration

B. Second iteration

In order to improve the results from the first tests, the focus was on improve the quality of the data before the feature extraction, for that purpose both a filter and a baseline wander removal process were added to our data pipeline before the RR peaks extraction. After those upgrades the model performance scores changed very little, evidenciating a different approach was necessary.

C. Third iteration

After the first presentation, the used features were expanded to get more information about the data. Taking as a reference [6] frequency domain and heart rate features were added to our analysis, which increased the prediction performance in both models.

Nevertheless, in this stage, problems arose when running the code in the hidden data set of the competition. And all efforts were put into understanding and minimizing these errors. After some research, it turned out to be associated with some particularly high-amplitude noisy records. Since it was not possible to have any information about the hidden data, logging functionalities were added to the code to understand the problem. Which turned out to be problems computing the features.

When running the RR peak detectors on such data, the relevant parts of the signal were sometimes ignored as having an amplitude significantly smaller than the noise. That caused the detector to return very few (or none at all) RR peaks, which is the base for calculating all the HRV features and lead the code to break during the tests. To deal with this problem, different RR peak detectors were used to look for a better performing one, ending up with the $ecg.ecg()$ from biosspy. Which was based on the Hamilton detector, it also performs a peak correction after the original detection, and was the one that performed best in the tests.

Still, there were few examples in this problem persisted, so in order to allow the model to deal with this information the number of detected R peaks was added to our tabular data as a feature. This was an attempt to help the model make classifications by adding the problems of the signal as features.

D. Fourth and final iteration

Until this moment the battle was against the fact that the code was not being able to predict the hidden data sets of the challenge. The best solution for the high amplitude noise signals involved extracting the R peaks intervals in sections of the signal that did not contain the noise. This is a costly method that might work for the project but was not implemented for this iteration.

The imbalancement of the data set was to blame for the heretofore poor performance on the minority classes (specially noise). Therefore, to approach this issue with oversampling (since the given data set is not particularly big, undersampling methods were not taken into account). Here is where the SMOTE algorithm came into the solution, to generate more sample of the less represented classes in our generated tabular data set right before training (as shown in 3).

The new features and the oversampling improved the classification of the noise signals, and helped the distinction between the classes N and O, which was a problem that was persisting since the first iteration.

Below, it is possible to observe the difference in the results with oversampling for each class. Those results will be detailed in the next section.

TABLE II
PERFORMANCE SCORES BEFORE AND AFTER OVERSAMPLING

Classes	Initial Samples	Initial F1	Samples after [1]	Final F1
N	521	0.89	3581	0.87
A	3581	0.72	3581	0.95
O	1713	0.67	3581	0.81
Noise	185	0.46	3581	0.97

[1] After oversampling with SMOTE.

IV. CONCLUSION

At the end of the project, an automated, modular system was developed with a data pipeline that runs all the way from preprocessing until training, model selection, and prediction. From the raw, imbalanced set of records, it is possible to get a considerably well-performing and generalizing model that can be used even to predict between other HRV detectable diseases like hypoglycemic episodes, ischemic episodes, ventricular tachyarrhythmia *et cetera* as gathered by [6].

Unfortunately, besides the time spent attempting to protect the code from errors and perform logging for debugging, the run of the code was not successful for 2 of the 3 final hidden data sets of the competition, as errors arose and our predictions could not be correctly evaluated, penalizing performance in the competition.

Finally, it is important to state that regardless of the obtained performance scores, it is necessary to correlate the ECG analysis with the clinical history in order to provide a proper diagnosis [14]. Therefore if the developed tools in this project were to be used in real situations, they would be an auxiliary medium for quicker detection of the disease.

REFERENCES

[1] Wold Health Organization (WHO). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds),.

[2] Solange Akselrod, David Gordon, F Andrew Ubel, Daniel C Shannon, A Clifford Berger, and Richard J Cohen. Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. *science*, 213(4504):220–222, 1981.

[3] Moez Ali. *PyCaret: An open source, low-code machine learning library in Python*, April 2020. PyCaret version 1.0.0.

[4] Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M. Bilginer Gulmezoglu. A survey on ECG analysis. *Biomedical Signal Processing and Control*, 43:216–235, 2018.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[6] Constantino Antonio Garcia Martinez, Abraham Otero Quintana, Xose A. Vila, Maria Jose Lado Tourino, Leandro Rodriguez-Linares, Jesus Maria Rodriguez Presedo, and Arturo Jose Mendez Penin. *Heart Rate Variability Analysis with the R package RHRV*. Use R! Springer-Verlag, s.l, 2017.

[7] Sebastian D Goodfellow, Andrew Goodwin, Robert Greer, Peter C Laussen, Mjaye Mazwi, and Danny Eytan. Atrial fibrillation classification using step-by-step machine learning. *Biomedical Physics & Engineering Express*, 4(4):045005, May 2018.

[8] Pradeep Kumar, Roheet Bhatnagar, Kuntal Gaur, and Anurag Bhatnagar. Classification of Imbalanced Data:Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering*, 1099(1):012077, March 2021.

[9] Gustavo Lenis, Nicolas Pilia, Axel Loewe, Walther H. W. Schulze, and Olaf Dössel. Comparison of Baseline Wander Removal Techniques considering the Preservation of ST Changes in the Ischemic ECG: A Simulation Study. *Computational and Mathematical Methods in Medicine*, 2017:9295029, 2017.

[10] Giuseppe Lippi, Fabian Sanchis-Gomar, and Gianfranco Cervellin. Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge. *International Journal of Stroke*, 16(2):217–221, February 2021.

[11] Shen Luo and Paul Johnston. A review of electrocardiogram filtering. *Journal of Electrocardiology*, 43(6):486–496, 2010.

[12] Kristjan Norland, Gardar Sveinbjornsson, Rosa Thorolfsdottir, Olafur Davidsson, Vinicius Tragante, Sridharan Rajamani, Anna Helgadottir, Solveig Gretarsdottir, Jessica Van Setten, Folkert Asselbergs, Jon Sverrisson, Sigurdur Stephensen, Gylfi Oskarsson, Emil Sigurdsson, Karl Andersen, Ragnar Danielsen, Gudmundur Thorgeirsson, Unnur Thorsteinsdottir, David Arnar, and Kari Stefansson. Sequence variants with large effects on cardiac electrophysiology and disease. *Nature Communications*, 10:4803, 10 2019.

[13] Jiapu Pan and Willis J Tompkins. A real-time QRS detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236, 1985.

[14] Andrés Ricardo Pérez-Riera, Raimundo Barbosa-Barros, Rodrigo Daminello-Raimundo, and Luiz Carlos de Abreu. Main artifacts in electrocardiography. *Annals of Noninvasive Electrocardiology: The Official Journal of the International Society for Holter and Noninvasive Electrocardiology, Inc*, 23(2):e12494, March 2018.

[15] Eduardo D Sontag et al. VC Dimension of Neural Networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998.

[16] S Stewart. Population prevalence, incidence, and predictors of atrial fibrillation in the Renfrew/Paisley study. *Heart*, 86(5):516–521, November 2001.

[17] Qing Wu, Yangfan Sun, Hui Yan, and Xundong Wu. ECG signal classification with binarized convolutional neural network. *Computers in Biology and Medicine*, 121:103800, June 2020.