Classification of 12-lead ECGs using Gradient Boosting on Features Acquired with Domain-Specific and Domain-Agnostic Methods

Durmus Umutcan Uguz[1], Felix Berief[1], Steffen Leonhardt[1], Christoph Hoog Antink[1]

[1]Medical Information Technology, Helmholtz-Institute for Biomedical Engineering,
RWTH Aachen University, Aachen, Germany

Abstract

This year, the objective of the CinC challenge was the classification of 12-lead electrocardiograms (ECG). The 12-lead ECG is an essential tool for detecting cardiac abnormalities and allows for a variety of analysis methods. The approach presented in this paper consists of two parts, feature extraction and classification. The extracted features can be separated into domain-specific and domain-agnostic features, where domain-specific features are based on known ECG processing methods such as QRS-detectors. Domain-agnostic features are generated by wavelet transforms that take the raw 12-lead ECG as input. Additionally, a novel beat-to-beat correlation analysis is proposed to identify arrhythmia occurring among other healthy beats. These features are then combined and classified by gradient-boosted trees implemented in Python. To account for the complexity of the multi-label and multi-class problem definition, a One-vs-Rest scheme is utilized, where distinct classifiers for each class determine whether a sample belongs to said class. The resulting imbalance in training sets for each classifier was compensated for by giving the positive samples a higher weight. The classifiers were trained using the XGBoost gradient boosting system. The proposed classification scheme of the team "Desafinado" received a score of 0.576 in the challenge.

1. Introduction

Cardiovascular diseases are the leading causes of mortality and account for $48\,\%$ of all deaths among the non-communicable diseases [1]. Advances in the electrocardiogram (ECG) monitoring in recent years has increased the demand in the automated and computer-aided ECG diagnosis. Machine learning tools are widely employed to meet this demand, have been tested in the problem of arrhythmia classification [2,3] and remain to be a focal point in the future of ECG diagnosis. The necessity of an automated arrhythmia classification is addressed in the PhysioNet/Computing in Cardiology Challenge 2020 [4] to encourage open-source approaches and obtain reproducible results.

This paper presents the approach developed by our team "Desafinado using gradient boosting on features generated from fiducial points, wavelet transform, higher order statistics and a novel beat-to-beat correlation analysis of the 12-lead ECG data. On the hidden set, our approach received a score of 0.576, and we were ranked 95$^{\text{th}}$ out of the 306 submission.

2. Methods

In the following, preprocessing, feature extraction, classification, and postprocessing are described.

2.1. Preprocessing

ECG recordings were low-pass filtered ($80\,\mathrm{Hz}$), after removing the baseline wander by using a two-step median filter (lengths of $200\,\mathrm{ms}$ and $600\,\mathrm{ms}$). To reduce the complexity of the feature vector from the beat-to-beat correlation analysis in Section 2.2.3, Kors transformation was used to transform 12-lead ECG recordings into orthogonal XYZ leads [5]. Both 12-lead ECG and obtained XYZ leads were segmented using BioSPPy [6].

2.2. Feature Extraction

For the classification, a wide range of features was used. These can be categorized into three groups.

2.2.1. ECG timing features

After segmenting the beats using Pan&Tompkins algorithm [7], following features ($\vec{v}_{\mathrm{timing}}$) were calculated from the resulting R-peak locations r_i:

- RR-interval $\delta_i = r_{i+1} - r_i$,
- ΔRR $\delta_i^2 = \delta_{i+1} - \delta_i$,
- pNN50,
- pNN20,
- number of beats.

In addition to these features, the following timing features were calculated from the segmented heartbeats:

- P-wave onset,
- QT interval,
- ST interval,
- T-wave onset,
- QRS width.

2.2.2. ECG Waveform Features

From the segmented heartbeats, the amplitudes of R, P, Q, T, and S-wave were calculated. In addition to these features calculated for 12 leads, the ratio of the amplitudes from the following waves were included:

- R/P,
- R/Q,
- R/T,
- R/S.

The set of waveform features ($\vec{v}_{\text{wav}}$) is completed by adding the average signal quality [8], average energy in the signal and the approximation coefficients of 3$^{\text{rd}}$-level wavelet decomposition using a $db1$ wavelet.

2.2.3. Beat-to-beat correlation analysis

Using an analysis window of $600\,\text{ms}$ centered at R-peak locations r_i, the median beat was calculated for each XYZ lead. Using this segmentation, beat waveforms were extracted for an analysis window of $600\,\text{ms}$ centered at r_i. Using these segmented heartbeats, the median beat for each lead was calculated. Using Pearson's correlation, a correlation vector consisting of ρ_i elements was calculated as

$$\rho_i = \frac{1}{L-1}\sum_{n=1}^{L}\Big(\frac{x_i[n]-\mu_i}{\sigma_i}\Big)\Big(\frac{x_{\text{med}}[n]-\mu_{\text{med}}}{\sigma_{\text{med}}}\Big), \quad (1)$$

where ρ_i, μ_i and σ_i represent the correlation coefficient, the mean and the standard deviation of the i^{th} beat, respectively. $x_i[n]$ is the i^{th} beat among N beats from the segmentation and $x_{\text{med}}[n]$ is the median beat of these N beats.

Segmented heartbeats were then sorted according to their ρ_i with the median heartbeat of the section. The most correlated (*accordant*) heartbeat and the least correlated (*dissonant*) heartbeat were retrieved, as illustrated in Figure 1. This attempt aims to isolate arrhythmias that do not occur at each heart cycle, and are thus prominent with their dissimilarity to the neighboring beats.

The following features were calculated from the correlation vector $\vec{\rho}$ for XYZ lead:

- std, mean, median, max, min of $\vec{\rho}$,
- std($\vec{\rho}$)/mean($\vec{\rho}$),
- std($\Delta\rho$), where $\Delta\rho = \rho_{i+1} - \rho_i$,
- std($\Delta\rho$)/mean($\vec{\rho}$).

In addition to these features from the beat-to-beat correlation vector $\vec{\rho}$, the following waveform features were calculated for both the dissonant and the accordant beats at each XYZ lead:

- skewness and kurtosis of each $50\,\text{ms}$ long section,
- argmax, max and min for the region [$0\,\text{ms}$ to $110\,\text{ms}$],
- max and min for the region [$240\,\text{ms}$ to $360\,\text{ms}$],
- argmax, max and min for the region [$410\,\text{ms}$ to $570\,\text{ms}$].

2.3. Classifier

The problem of assigning one or more of K classes to the 12-channel ECG recording and its corresponding features is a so-called multi-class multi-label task. The presented classifier first transforms this problem into K binary classification problems by using a One-vs-Rest approach, where the j^{th} classifier decides whether the ECG recording belongs to class j or not. This problem transformation was needed since the utilized gradient boosting classifier can not handle multiple labels per input. Due to the application of the One-vs-Rest approach, binary classifiers receive highly imbalanced datasets even with slightly more positive samples [9]. To account for this imbalance, positive samples for class k were re-weighted during the training of classifier k by

$$w_k = \frac{N_{\text{tot}} - N_k}{N_k}, \quad (2)$$

where N_k and N_{tot} are the number of samples for class k and the total number of samples, respectively. The negative samples received a weight of 1. Classes considered to be equivalent by the organizers were treated as the same class and only the scored classes were used in the classifier. Thus the total number of classes was $K = 25$.

The classification was handled by the XGBoost algorithm utilizing the idea of gradient boosted trees [10]. Boosted trees are decision trees used as an ensemble to build a single stronger classifier from many weak classifiers. Gradient boosting refers to the gradient descent method employed to find the best decision tree, whereas boosting itself indicates that samples previously misclassified receive larger weights for the next training steps. XGBoost utilize a variety of highly optimized techniques to make the training both robust and fast while still maintaining a high degree of customizability.

Among 20 tunable parameters offered by XGBoost, 3 of the most impactful ones were tested extensively for this challenge. All of these parameters, namely *gamma*, *min_child_weight* and *max_depth*, play an important role for avoiding overfitting without losing the ability to generalize to unseen data. As the hidden test data partly includes samples from a new data source, the parameter combinations were optimized not only through 5-fold

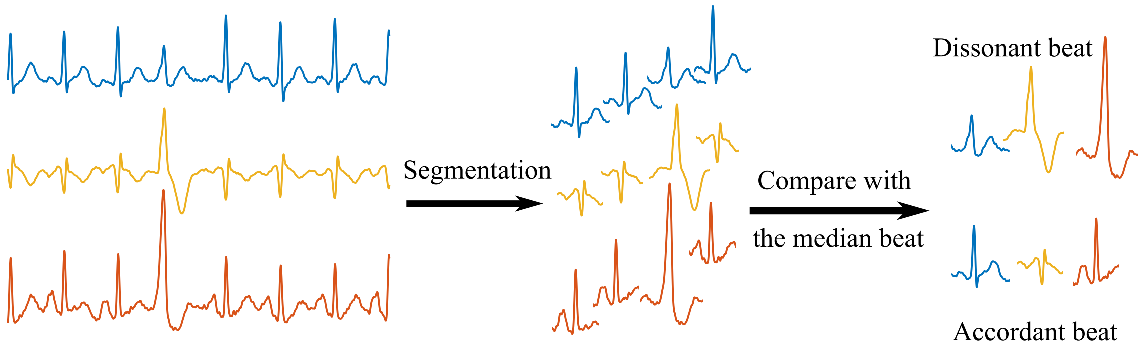

Figure 1. The correlation analysis conducted on the orthogonal XYZ leads. After segmenting the heartbeats, the found median beat is Pearson-correlated with each individual beat to find the least and the most similar beats, called as dissonant and accordant beats, respectively.

cross-validation but also by holding out entire data sources for evaluation.

2.4. Postprocessing

Since the classifier makes use of the One-vs-Rest technique, it consists of independent binary classifiers for each of the K classes. One of the downsides of this approach is the independent decisions as some diagnoses have a higher probability to occur jointly whereas other diagnoses might not be able to appear together in the same recording. This could also lead to the case where no output class is assigned.

In order to reduce the impact of these limitations, each class was assigned a specific threshold of t_k instead of relying on the default binary classification threshold of 0.5. The thresholds t_k were found by optimizing the challenge score as a function of t_k during cross-validation.

3. Results

During the official phase of the challenge, four versions of the algorithm were tested by our team “Desafinado”, see Table 1.

Submission	Runtime	Score
1	39:55:00	0.552
2	**38:17:00**	0.541
3	38:26:00	0.546
4	101:28:00	**0.576**

Table 1. Runtime and Score of the entries submitted during the official phase of the challenge.

While the final entry had a significantly higher runtime of more than 100 hours, it achieved the highest score and was thus selected for the final evaluation yet to be performed.

4. Outlook

As the results of the final evaluation of the challenge will be announced at the conference, an in-depth analysis of our submitted approach will be presented in the final version of the manuscript after the conference.

References

[1] Mendis S, Puska P, Norrving B, Organization WH. Global atlas on cardiovascular disease prevention and control. World Health Organization, 2011.

[2] Sahoo S, Dash M, Behera S, Sabut S. Machine learning approach to detect cardiac arrhythmias in ecg signals: A survey. IRBM 2020;.

[3] Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A. A review on deep learning methods for ecg arrhythmia classification. Expert Systems with Applications X 2020;100033.

[4] Alday EAP, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad AB, Elola A, Seyedi S, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. medRxiv 2020;.

[5] Kors J, Van Herpen G, Sittig A, Van Bemmel J. Reconstruction of the frank vectorcardiogram from standard electrocardiographic leads: diagnostic comparison of different methods. European heart journal 1990;11(12):1083–1092.

[6] Carreiras C, Alves AP, Lourenço A, Canento F, Silva H, Fred A, et al. Biosppy: Biosignal processing in python. Accessed on 2015;3(28):2018.

[7] Pan J, Tompkins WJ. A real-time qrs detection algorithm. IEEE transactions on biomedical engineering 1985;(3):230–236.

[8] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, S H Chen A. Neurokit2: A python toolbox for neurophysiological signal processing, 2020. URL `https://github.com/neuropsychology/NeuroKit`.

[9] Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance met-

rics based on the binary confusion matrix. Pattern Recognition 2019;91:216–231.

[10] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016; 785–794.

Address for correspondence:

Durmus Umutcan Uguz
Medical Information Technology
RWTH Aachen University
Pauwelsstr. 20, 52074 Aachen, Germany
uguz@hia.rwth-aachen.de