Selected Features for Classification of 12-lead ECGs

Marek Żyliński*[1], Gerard Cybulski[1]

[1] Faculty of Mechatronics, Warsaw University of Technology, Warsaw, Poland

Abstract

In this paper we describe our algorithm develop by the Alba_W.O. team at The PhysioNet/Computing in Cardiology Challenge 2020. Our approach achieved a challenge validation score of 0.308 and a full test score of 0.102, placing us 31 out of 40 in the official ranking. Our final algorithm is based on bootstrap-aggregated (bagged) decision trees. For the classification task, we provide a set of features extracted from 12-lead ECG, in detail describe later. We use the method implemented in PhysioNet-Cardiovascular-Signal-Toolbox: Global Electrical Heterogeneity, arterial fibrillation features, and PVC detection. We also estimate ECG periods (PR, QS, QR, PT, TP) and morphology parameters (ST elevation, QRS area, ECG value at R points).

We also examine the importance of each predictor individually, for the classification task, using a t-test. All groups of used parameters, without sex shown utility in some class classification cases.

1. Introduction

In this paper we describe our algorithm develop at The PhysioNet/Computing in Cardiology Challenge 2020. The goal of the 2020 Challenge is to identify the clinical diagnosis from 12-lead ECG recordings, the dataset granted for the challenge contains beyond 40 000 recordings with clinician diagnoses from 111 classes (one recording can have multiple diagnose). In the challenge 27 different classes are evaluated, rest are omitted.

The goal of our Alba team was to design and implement, in our case in Matlab, a working algorithm for automatic detection and classification of cardiac abnormalities base on the provided dataset.

2. Algorithm

The algorithm has 3 steps: ECG pre-processing features extraction and classification. A sketch of the algorithm is shown in figure 1. Our solution is based on bootstrap-aggregated (bagged) decision trees.

For the classification task, we provide a set of features extracted from 12-lead ECG, in detail describe later. We use the method implemented in PhysioNet-Cardiovascular-Signal-Toolbox: Global Electrical Heterogeneity, arterial fibrillation features, and PVC detection. We also estimate ECG periods (PR, QS, QR, PT, TP) and morphology parameters (ST elevation, QRS area, ECG value at R points).

The classification algorithm was trained with all recordings from the data set, without sample with multiple or no one diagnose evaluated in the challenge.

In the official phase, submission is scored use not-public test data. The best score obtained by our team is 0,308.

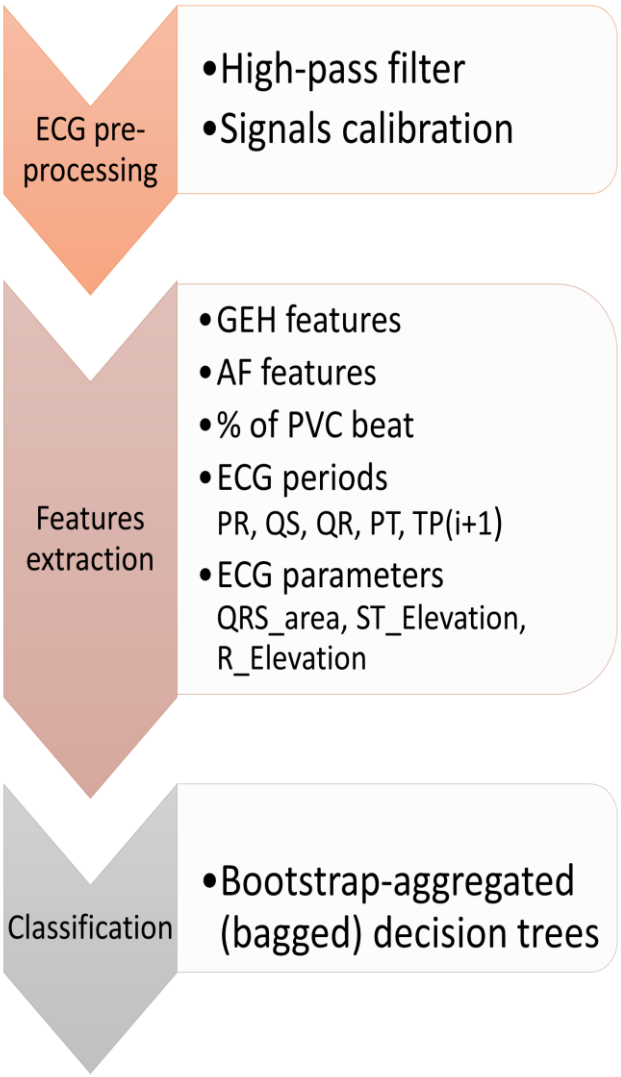

Figure 1: Sketch of our algorithm. It can be dividing into 3 steps: ECG pre-processing, features extraction, and classification

2.1. Signal preprocessing

First, we load data from a file, we used a similar approach as an example. From the header file, we utilize age and sex data as classification features and gain and sample frequency for signal calibration. After calibration, we perform signal filtration. We used a median filter to remove some noise and the Butterworth high pass filter at cut-off frequency = 1 Hz to remove isoline's floating.

2.2. Features

We base on the built-in function from the PhysioNet-Cardiovascular-Signal-Toolbox. We used the following features:

Global Electrical Heterogeneity (GEH features) – such as in example code, this group contains 22 parameters based on spatial ventricular gradient vector (SVG) such as SVG magnitude, SVG elevation, SVG azimuth, etc. [4]

AF features – group of features obtained from AF_features.m function at ECG_Analysis_Tools – this subset of features provide analysis of variability in RR interval and some sophisticated features such as coefficient if fuzzy measure entropy etc. [5]

The ratio of PVC beat – we use a modified PVC_detect.m function to detect premature ventricular contraction beats. The function uses a convolutional neural network and features obtained from the wavelet transform. A feature in our algorithm we use the ratio of beat detected as PVC to all ECG beats in the recording.

ECG periods –wavedet_3D_ECGKit function return time of characteristic points in ECG – based on it we write a function that calculates some ECG periods, selected base on literature [7] , [8] ,[9] : PR, QS, QR, PT, and TP (Figure 2). The mentioned periods are shown in image 2. We also calculate RAPR which is a ratio of PR and RR [10] :

$$RAPR_i = \frac{PR_i}{PP_{i-1}}$$

If it was possible, ECG period times were calculated for each beat. if some necessary point in cardiac beat was not detected, that period was omitted. As the classification features, we used the 7 metrics:

- Minimum time of the period
- Maximum time of the period
- The median time of the period
- The standard deviation of the time of the period
- The minimum value of intra beat difference
- The maximum value of intra beat difference
- Medium value of intra beat difference

Intra-beat difference (IBD) was calculated as [10] :

$$IBD_i = \frac{t_i - t_{i-1}}{mean(t)}$$

Where: t is a period's time.

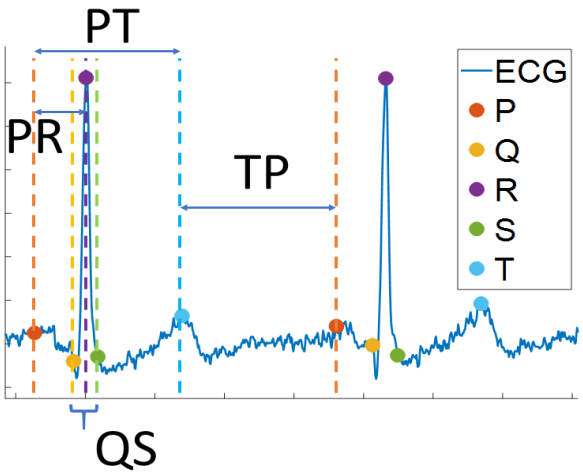

Figure 2: ECG periods use for classification. Times were calculated as the difference of time is given characteristic points detected be wavedet_3D_ECGKit function.

ECG morphology parameters – three time-domain ECG signal describe its morphology was also use:

– we estimate QRS area [10] (as integrated of ECG signal between Q and S points) for each lead (12 features)

- ST-elevation as the difference between the value of ECG in J-point and isoline. J-point was defined as local extremum after QRSend points. Elevation was estimated for each beat in each lead. For classification, we use the same set of metrics as for ECG periods. We obtained 84 features (12 leads x 7 metrics).

- R elevation was defined as the difference in the value of ECG in R points and isoline. The median value for each lead was used as features.

The illustrative plot of the estimation method of R and ST elevations are shown in figure 3:

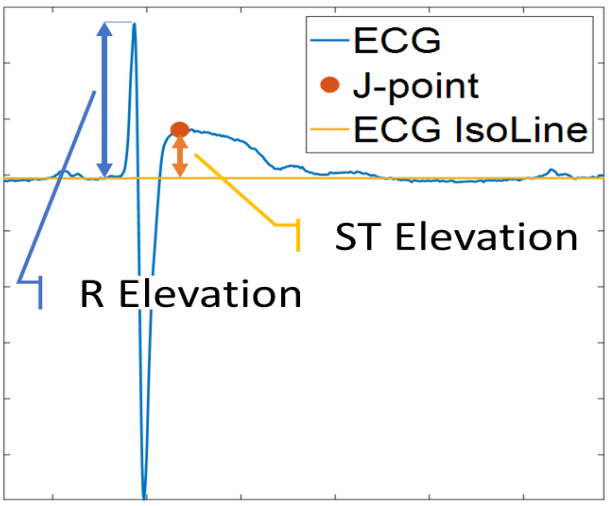

Figure 3: ST elevation was estimated as a difference between ECG isoline and ECG value at J-point – defined as local extremum after the end of QRS. R elevation is ECG value at R.

2.3. Classification algorithm

We use build-in MATLAB, The Classification Learner app to choose the best classifier. Forests classification supported vector machine; k-nearest neighbor classifier was tested. As training data, we use all described features obtained for all samples in the challenge data set. Only valid recording, with one scored diagnose, is evaluated.

Bootstrap-aggregated (bagged) decision trees shown the best accuracy and was chosen to perform the classification task. Our forest contains 50 random trees.

Our algorithm was scored several times in the official phase, the best challenge score obtained be our team is 0,308.

3. Feature selection

Last, we tried to reduce the dimension of our set of features. Reducing features can also save storage and computation time and increase comprehensibility.

We use a simple filter approach to select the feature subsets. For this analysis, we used a set of features extracted from recordings of the Challenge's training set. We apply the t-test on each feature in each class. For each class, we divided the data into two groups: samples from in-class recording and the rest group. We exanimated the p-value to evaluate how effective that features separate groups.

In table 1 obtained p-value for each group of features is summarized. Features are grouped based on the source for example, all features obtained from AF_features function are grouped as one. We found that every group of features that we selected can be utilized to identify the clinical diagnosis from 12-lead ECG recordings. Every feature can help in the classification of the same classes. But also, parameters are not useful in all cases.

GEH features and QRS area have shown an advantage for all diagnosis's classification. Sex and percentage of PVC beat differentiate the smallest number of classes.

Despite all, as a result, none of the features that we use were removed from our algorithm.

4. Conclusion

Alba team gets involved in the Physionet/CinC 2020 challenge. We design and implement, a working algorithm for automatic detection and classification of cardiac abnormalities base on the provided dataset.

Our algorithm code will be available in the git-hub repository, after the end of the CinC 2020 conference (at the beginning of October 2020):

https://github.com/AlbaWOinPhysio/Physionet2020

The code is released under the GPL license.

Acknowledgments

We want to thank Alba for being a mascot of our team. She is only one known albino orangutan. Please visit: https://savetheorangutan.org/

References

[1] Erick A. Perez Alday, Annie Gu, Amit Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, Matthew A. Reyna. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol. Meas.

[2] John G. R. Kohavi, "Wrappers for feature subset selection", Artificial Intelligence, No.1-2, pp.272-324. 1997

[3] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, … & H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. "Circulation [Online]. 101 (23), pp. e215–e220, 2000

[4] L. G. Tereshchenko, "Global electrical heterogeneity: mechanisms and clinical significance." In 2018 Computing in Cardiology Conference (CinC) IEEE. September 2018

[5] Q. Li, C. Y. Liu, J. Oster, and G. D. Clifford, Chapter: "Signal processing and feature selection preprocessing for classification in noisy healthcare data." In book: Machine Learning for Healthcare Technologies, 2016

[6] Vest A, Da Poian G, Li Q, Liu C, Nemati S, Shah A, Clifford GD, "An open source benchmarked toolbox for cardiovascular waveform and interval analysis", Physiological measurement: 105004. DOI:10.5281/zenodo.1243111, 2018

[7] L. Mao, H. Chen, J. Bai, J. Wei, Q. Li, & R. Zhang, "Automated detection of first-degree atrioventricular block using ECGs." In International Conference on Health Information Science (pp. 156-167). Springer, Cham. October, 2019

[8] M. Elgendi, M. Jonkman, & F. De Boer, "Premature atrial complexes detection using the Fisher Linear Discriminant." In 2008 7th IEEE International Conference on Cognitive Informatics (pp. 83-88). IEEE. August, 2008

[9] Rafał Baranowski, "25 EKG na XXV-lecie SENIT" ECG atlas in Polish, from Kasprowisko 2019 conference. 2019

[10] V. T. Krasteva, I. I. Jekova, & I. I. Christov, "Automatic detection of premature atrial contractions in the electrocardiogram." Electrotechniques Electronics E & E, 9, 10. 2006

Address for correspondence:
Marek Żyliński
Wydział Mechatroniki
Ul. Św. Andrzeja Boboli 8
02-525 Warszawa
Poland
zylinski@mchtr.pw.edu.pl

Table 1. The result of the analysis of the features' importance is presented. The minimal value of the p-value for each feature group is listed. Features are grouped based on the source, for example, all features obtained from the AF_features function are grouped as one also all metrics for the given ECG period are grouped as one, etc. If any feature from a group shown statistical significance in a t-test ($p<0,05$) whole group is marked as important (green). GEH features and QRS area shown an advantage for all diagnosis's classification, every group of features can be utilized in some cases.

SNOMED CT Code	Name	age	sex	GEH	AF feature	% PVC beat	PR	QS	QR	PT	TP	RAPR	ST elevation	QRS area	R elevation
10370003	pacing rhythm	0,01	0,15	0,00	0,00	0,20	0,00	0,00	0,00	0,00	0,05	0,02	0,00	0,00	0,00
17338001	ventricular premature beats	0,87	0,75	0,01	0,00	0,03	0,03	0,18	0,18	0,12	0,01	0,17	0,09	0,01	0,17
39732003	left axis deviation	0,12	0,64	0,00	0,00	0,00	0,00	0,04	0,04	0,00	0,02	0,13	0,01	0,00	0,00
47665007	right axis deviation	0,24	0,41	0,00	0,02	0,56	0,00	0,03	0,03	0,11	0,43	0,19	0,00	0,00	0,00
59118001	right bundle branch block	0,00	0,16	0,00	0,02	0,10	0,03	0,00	0,00	0,02	0,04	0,05	0,00	0,00	0,00
59931005	t wave inversion	0,79	0,72	0,00	0,06	0,30	0,14	0,01	0,01	0,07	0,13	0,23	0,01	0,00	0,02
111975006	prolonged qt interval	0,00	0,66	0,00	0,00	0,25	0,08	0,21	0,21	0,00	0,03	0,18	0,00	0,00	0,00
164889003	atrial fibrillation	0,00	0,98	0,00	0,00	0,10	0,01	0,01	0,01	0,01	0,00	0,01	0,03	0,00	0,03
164890007	atrial flutter	0,22	0,05	0,01	0,00	0,63	0,21	0,00	0,00	0,23	0,02	0,03	0,06	0,00	0,05
164909002	left bundle branch block	0,00	0,15	0,00	0,07	0,00	0,00	0,01	0,01	0,06	0,00	0,09	0,00	0,00	0,00
164917005	Q wave abnormal	0,01	0,67	0,00	0,00	0,00	0,00	0,02	0,02	0,25	0,05	0,09	0,00	0,00	0,00
164934002	t wave abnormal	0,73	0,09	0,00	0,05	0,00	0,00	0,00	0,00	0,06	0,52	0,05	0,00	0,00	0,01
164947007	prolonged pr interval	0,63	0,64	0,00	0,01	0,30	0,00	0,02	0,02	0,01	0,31	0,01	0,00	0,00	0,00
251146004	low qrs voltages	0,56	0,47	0,00	0,21	0,17	0,02	0,00	0,00	0,02	0,09	0,07	0,01	0,00	0,00
270492004	1st degree av block	0,28	0,65	0,00	0,02	0,11	0,00	0,06	0,09	0,00	0,04	0,00	0,18	0,00	0,15
284470004	premature atrial contraction	0,00	0,23	0,01	0,00	0,70	0,45	0,06	0,06	0,23	0,00	0,13	0,18	0,00	0,27
426177001	sinus bradycardia	0,77	0,87	0,00	0,00	0,99	0,21	0,18	0,33	0,00	0,00	0,00	0,03	0,00	0,02
426783006	sinus rhythm	0,00	0,03	0,00	0,00	0,00	0,00	0,14	0,14	0,11	0,00	0,00	0,00	0,00	0,00
427084000	sinus tachycardia	0,00	0,16	0,00	0,00	0,09	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
427393009	sinus arrhythmia	0,99	0,74	0,01	0,01	0,00	0,00	0,35	0,35	0,15	0,11	0,21	0,26	0,00	0,18
445118002	left anterior fascicular block	0,06	0,42	0,00	0,37	0,14	0,14	0,01	0,01	0,05	0,23	0,22	0,01	0,00	0,00
698252002	nonspecific intraventricular conduction disorder	0,88	0,93	0,00	0,53	0,67	0,06	0,04	0,11	0,15	0,24	0,03	0,00	0,00	0,00
713426002	incomplete right bundle branch block	0,99	0,48	0,00	0,30	0,00	0,00	0,04	0,04	0,00	0,34	0,00	0,00	0,00	0,00
713427006	complete right bundle branch block	0,00	0,69	0,02	0,09	0,27	0,00	0,03	0,03	0,01	0,23	0,09	0,15	0,00	0,31