Creating Open-Source Python Package for ECG Processing

POLINA TURISHCHEVA, Innopolis University
MANUEL MAZZARA, Innopolis University
KONSTANTIN USHENIN, UrFU

1 PROJECT PROPOSAL

Electrocardiography (ECG) is a process of recording electrical heart activity- voltage versus time. Classical ECG contains 5 peaks- P, Q, R, S, T- however, not all of them are distinguishable on all animals. In general, the Q-R-S complex can be seen on most animals, but P and T peaks can be hidden by signal noise/baseline wander, especially for small animals (such as mice). Information carried in ECG waves, such as T-T or QRS intervals or ST segments or any wave's amplitude can be extremely useful for clinical needs. For instance, T wave amplitude drift is often associated with genetically inherited illnesses.

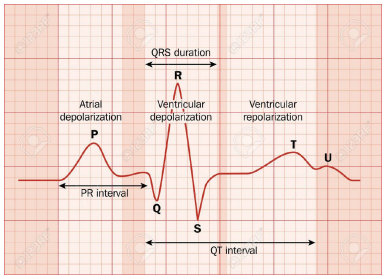


Fig. 1. ECG intervals/segments examples. U-wave is rarely distinguishable in real recordings.

Taking into account the great value of ECG, we would like to create an open-source python project for ECG processing. There is different software already available, e.g. ADInstruments and DSI ECG PRO. The problem is that this software is not free, it often uses proprietary extension (you should pay for month/annual license to have an ability to convert from that extension). Moreover, it often cannot be customized and fail in cases of unstable ECG, such as neonatal or pregnant ECG of lab animals. As for open-source packages, there exist BioSig package for C/C++ and Matlab, but not for python. Also PhisioZoo platform has a Matlab toolbox for ECG processing but the results of its processing are discussable and there are many things to improve there as well.

In python, there is a BioSPPy package, which can be used for any biological signal processing (ECG/EEG/EMG). Therefore, it does not take into account any data peculiarities and it is far from being comfortable for usage or widespread. For instance, BioSPPy also offers EEG processing but most of the users prefer mne package as it is created specially for EEG. We want to create mne analogy for ECG.

Additional Key Words and Phrases: ECG, biological signal processing

ACM Reference Format:
Polina Turishcheva, Manuel Mazzara, and Konstantin Ushenin. 2020. Creating Open-Source Python Package for ECG Processing . 1, 1 (March 2020), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: Polina Turishcheva, p.turischeva@innopolis.university, Innopolis University; Manuel Mazzara, m.mazzara@innopolis.ru, Innopolis University; Konstantin Ushenin, kostanew@gmail.com, UrFU.

XXXX-XXXX/2020/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 ITERATION ONE

Here is the plan created for the following semester. It may be corrected according to the challenges met during the research activity. As for now it is both initial and current plan.

Week	Activity
3-4	Literature review of the problems in ECG processing (e.g. baseline wander) and classical approaches to deal with them. Find standards for checking the correctness of ECG processing program.
5-6	Find preliminary datasets with human/lab animals ECG recordings. Most probably will be found on PhysioZoo and PhysioNet.
7-8	Create a draft program and validate its results - that Q-R-S -(P-Theta) peaks are defined correctly
9-10	Statistical analyses of the program output. Comparing the QRS definition with/without preprocessing. Enhancing preprocessing
11-12	Literature review of different king of arrhythmia. Trying to find addictions to any kind of diseases in the analyzed recording. Validating the results
13-14	Time to fix all the problems found out.
15-16	Refactoring the program to a python module. Creating a report.

2.1 Literature review

Most often problems with ECG preprocessing are baseline wander, signal to noise ratio, removing electrical wire noise, artefacts, e.g. rapid movements or detached electrode. During removing baseline wander/specific frequencies we have a chance to offset some particular peak intervals or segments. This will lead to biased data followed by incorrect diagnosis. Signal to noise ratio should be minimized during recording or on the hardware layer. Removing wire noise is classically done with notch filter. Techniques for removing artefacts depend on the artefacts - the most common approach is to compare the part of the recording with ECG pattern and remove it if it does not suit to the pattern.

During ECG processing the problem is to find wave peaks', starts and endings correctly. Solutions for this step depend on the preprocessing techniques.

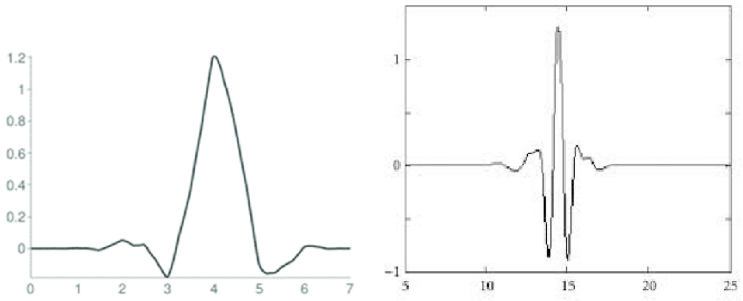


Fig. 2. Left - Wavelet sym4
Right - Wavelet Daubechies8

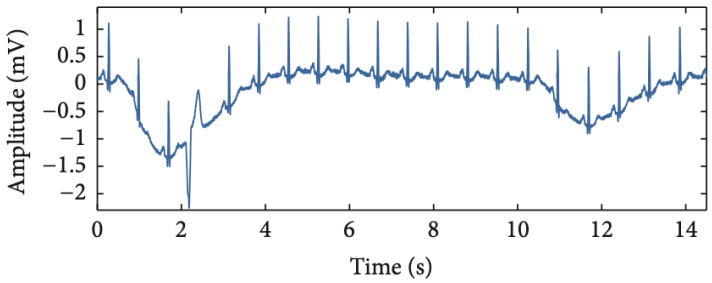


Fig. 3. Baseline wander is "a low frequency artifact in the ECG that arises from breathing, electrically charged electrodes, or subject movement and can hinder the detection of these ST changes because of the varying electrical isoline" - [4].
In ECG without baseline wander, the line should be near zero, except of the wave peaks.

In article [4] 5 most popular methods for ECG preprocessing -particularly baseline wander removal- were compared with each other. They are

- Butterworth High-Pass Filter - "transfer function of the filter had an order of 2 but the filtering process was performed in forward and reverse direction creating a zero-phase filtered signal and a resulting order of four." Filtering in straight and reverse direction is required not to offset peaks relative to time.
- Make Spline Approximation and Subtract it - detection of the "center of the PQ interval in every beat and to interpolating those points to create an estimate of the baseline wander."
- Find Moving Median and Subtract it - "concatenation of two moving median filters and subtracting that estimate from the corrupted signal."
- Wavelet-Based Baseline Cancellation (Daubechies 8 wavelet) - "the signal was decomposed using the discrete wavelet transform (DWT) and the approximation coefficients at the lowest frequency band were set to zero with the aim of fully cancelling baseline wander."
- Wavelet-Based High-Pass Filtering (Vaidyanathan-Hoang wavelet) - same a previous but "a high-pass filtering is used on the approximation coefficients instead of setting them to zero"

The result of the article says that "the best performing filter with respect to quality of the reconstruction turned out to be the wavelet-based baseline cancellation", hence, we would use this approach in our following studyings.

Solution in (4)	Solution in (3)	Our Planned Solution	Reasoning
Butterworth Filter 0.05 - - Hz	Butterworth Filter 0.5 - 100 Hz	Butterworth Filter 0.05 - 100 Hz	Filtering the signal with frequencies ≥ 0.05 Hz leads to QT segment and J point offset
No	No	Notch Filter for 45-55 or 55-65 Hz	To reduce the wire noise (50 Hz or 60 Hz, depending on the country).
Wavelet decomposition with Daubechies 8	No	Wavelet decomposition with Daubechies 8 for human and sym4 for animals	In lab animals' ECG P and T waves are less expressed than in human ones. The choise of wavelets is based on their form.

2.2 Standards for checking ECG processing

For human ECG there are a lot of standards defined. One of the most popular (5, 6) consists of 10 short recordings are specified by the current American National Standard for testing various devices that measure heart rate.

PhysioNet provides a Matlab toolkit for validating ECG processing programs. We will use these recordings a standard, however, the following soft validation with some of the marked databases is also planned to be made.

2.3 What was done

(1) Literature review - 8 hours - to be continued
(2) Searching for standards - 2 hours - done
(3) Writing a report - 3 hour - to be continued

3 ITERATION TWO

3.1 Criteria for database

(1) It should be an open-source dataset from a reliable source
(2) The dataset should be marked in case if it is recorded from an ill animal (e.g. arrhythmia)
(3) The recordings should contain different real-world artifacts, for instance, noisy sequences because of electrode estrangement - this is important for real-world applications to mark such artifacts and deal with them
(4) The sampling frequency is no less than 500 Hz - this may occur to be important for the following digital signal processing.
(5) There should be no pregnant or fetal ECGs as they are extremely different, therefore, their processing is different as well
(6) The recording should be rather short. We are mostly interested in processing basic ECG, the statistical analyses to validate the processing and maybe the following classifying of arrhythmia or other widespread decease. For this we do not need one-day long halter-recording. We are interested in recordings no less than 3 minutes but also no more than 5 hours.

3.2 Database resources and final choice

I have used three acknowledged resources with biological data: Zenodo, PhysioZoo, and PhysioNet.
Three datasets were chosen from the mentioned resources (the names are clickable):

(1) Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020
- Human dataset from PhysioNet. All of the recordings are 500Hz. It was chosen as it contains marked data and represents many different widespread cardial deceases:
  - Normal sinus rhythm (Normal)
  - Atrial fibrillation (AF)
  - First-degree atrioventricular block (I-AVB)
  - Left bundle branch block (LBBB)
  - Right bundle branch block (RBBB)
  - Premature atrial complex (PAC)
  - Premature ventricular complex (PVC)
  - ST-segment depression (STD)
  - ST-segment elevation (STE)

(2) PhysioZoo - mammalian NSR databases
- The dataset contains ECGs from most popular laboratory animals: dogs, rabbits, and mice. Dogs recording were made at 500Hz, mice and rabbits at 1 kHz. All rabbits were female and none of the animals were under sedative. All of this is relevant as far as drugs or gender may influence the average heart rate. Nevertheless, this database is not big - it is one of best animal ECG databases I have managed to found. It is also relevant as it contains real-world noise artifacts and some preliminary statistics.

	Dog	Rabbit	Mouse
Number of records	17	20	8
Number of mammals	17	4	8
Average length (hr:min:sec)	00:05:31	00:10:34	00:29:44
Min length (hr:min:sec)	00:04:09	00:04:53	00:13:53
Max length (hr:min:sec)	00:06:48	00:26:00	00:39:38
Total length (hr:min:sec)	01:33:55	03:31:13	03:28:07
Total R-peak annotations	10,871	50,452	109,865
Bad quality, Gross (%)	2.7%	1.0%	0.4%

Fig. 4. The authors'statistics about PhysioZoo - mammalian NSR databases

(3) ANSI/AAMI EC13 Test Waveforms
- The database to check that our program meets American National Standards (were chosen during Iteration one). All of the recordings are 720 Hz.

3.3 Preparations for the following iteration

As looking for a database occurred to be a bit faster than we have expected, preliminary steps for program writing were made:

- Uploading files (edf/txt/csv -formats)
- Found python functions for filters planned on iteration one - *scipy.signal.butter, scipy.signal.filtfilt, scipy.signal.iirnotch*

Also thought about the algorithm to detect different parts of ECG signal.

(1) Filter the signal from baseline drift
(2) Find R -peak - it is reasonable to start from it as R-peak is the most different part of ECG signal. Also it has the greatest amplitude.
(3) Find minimum point to the right and to the left from each R peak - these will be Q and S.
(4) Take all point between current S and following Q and try to find P and Theta waves in this area.

Actually, this algorithm does not solve the problem of detached electrode. The current solution I suggest is to track the average amplitude of R peaks and if the amplitude of the signal in some point is twice bigger than the average R-amplitude, we mark it as noise. The area with noise it discarded from the recording, we create two separate recordings instead, but save the start and finish time of the discarded area - it may be crucial for further statistical analyses. However, this solution is very rough and should be improved during the coding stage. Also in some recordings in the chosen databases are reversed - the R-peaks are oriented down, not up. We should somehow detect the orientation of the recording and in case it is oriented down - multiply it on (-1) to reverse the orientation again and apply the same algorithms.

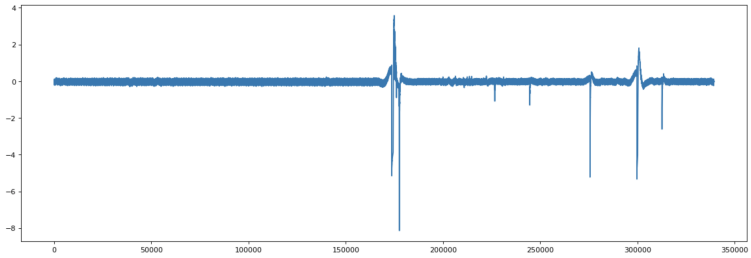


Fig. 5. Example of mice ECG to show artefacts. The extremely high peak is a detached electrode. We cut out such areas if the amplitude of the peak is more than average value of R-peaks+ variance. We remember the start and stop indexed of cut out sequence because it should be useful for the following statistical analyses.

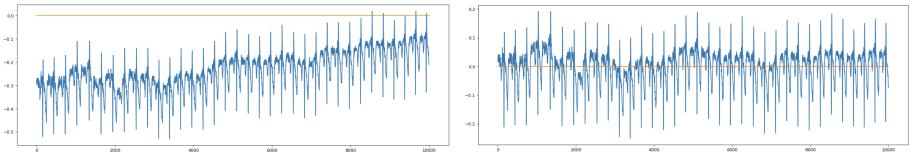


Fig. 6. The left picture is example of ECG with baseline drift. The right one if filtered. The orange line is zero and ECG without baseline drift should be close to it. However, filtering may be improved more. This is an example of rabbit ECG from the 2nd dataset.

3.4 What was done

(1) Designed criteria to select databases
(2) Found 3 databases meeting the selected criteria - 4 hours - finished
(3) Preparations for coding - 3 hours - to be continued
(4) Writing a report - 3 hours - to be continued

4 ITERATION THREE

4.1 Results:

Link to the first Implementation
During this 2 weeks I have implemented the draft version of ECG processing tools. However, they have been validated only on the 2nd dataset with animals. For the results of the work see Figures 5-14.

4.2 What was done

- The program for Q-R-S-T peaks definition was created - 15 hours - to be continued
- Its results were partially validated - only on animals dataset - 4 hours - to be continued
- - writing a report - 2 hours - to be continued

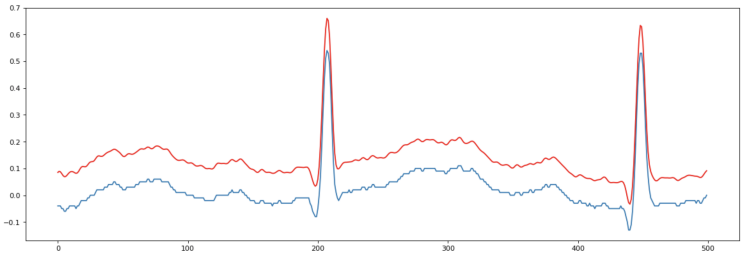


Fig. 7. The picture shows that Butterworth filtering does not affect peaks position vs time, which is extremely important for correct analyses. The red line is before filtering, blue one is after. This is an example of a rabbit ECG.

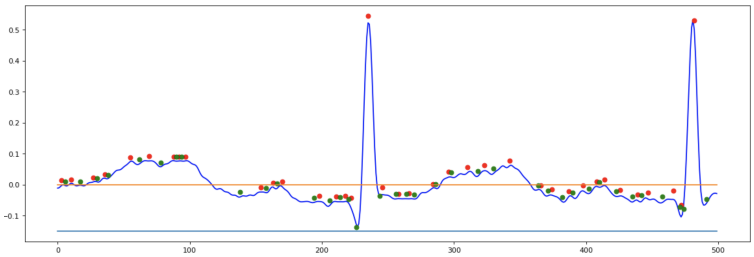


Fig. 8. We also use wavelet decomposition for filtering. Here is an example of a rabbit's ECG after butterworth filter. The green dots are minimums of the wavelet approximation, the red dots are the maximums of the wavelet approximation. Wavelet decomposition used 6 levels, 3 of them were used for signal reconstruction. If using less levels, peaks displacement becomes significant. For mice 4 levels are required.

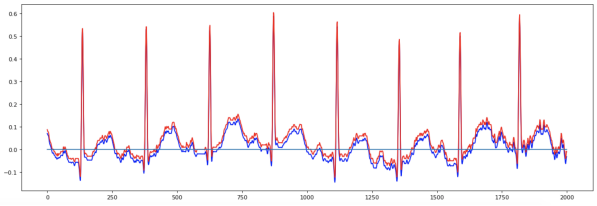


Fig. 9. Wavelet reconstruction without 1 layer

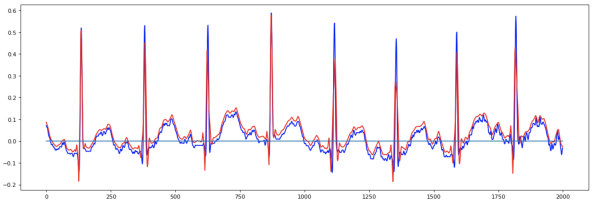


Fig. 10. Wavelet reconstruction without 3 layers

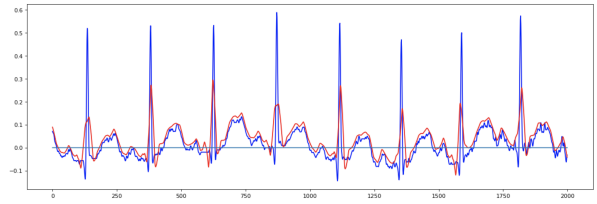


Fig. 11. Wavelet reconstruction without 4 layers

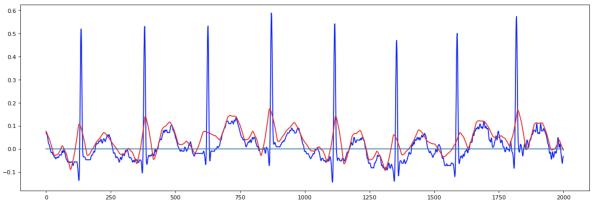
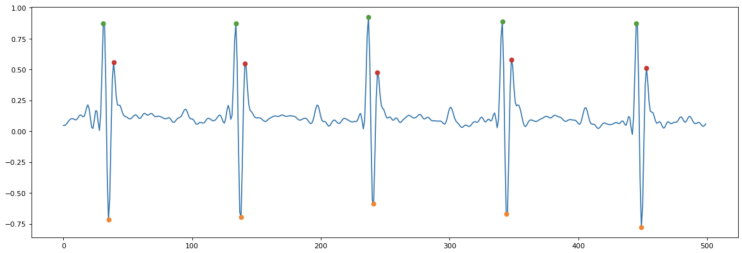
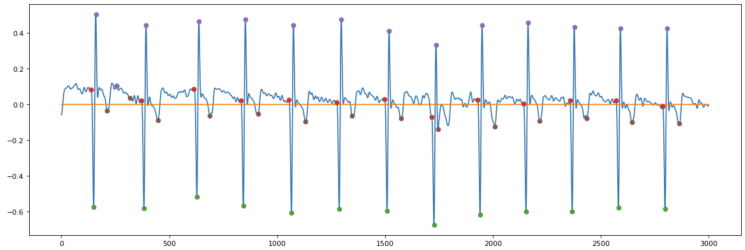


Fig. 12. Wavelet reconstruction without 5 layers

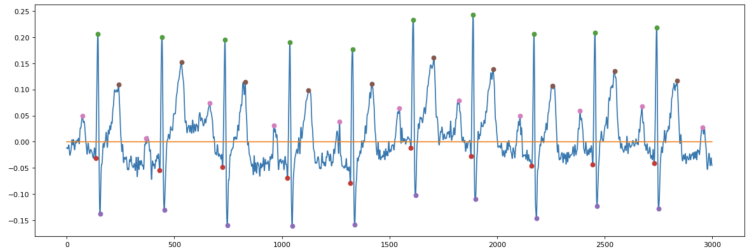
Fig. 13. From this decomposition of a sample rabbit ECG we conclude that no more than 3 levels should be discarded in wavelet deconstruction, otherwise the minimums and maximums start shifting. Same situation for dogs. For mice, whose amplitude of signal is much less and more frequent, no more than 2 levels can be discarded.



Mouse: Orange dots - R peak, Green dots -Q peaks, Red dots - S peaks



Dog : Green dots - R peak, Pink dots - P peak, Red dots -Q peaks, Violet dots - S peaks



Rabbit : Green dots - R peak, Pink dots - P peak, Red dots -Q peaks, Violet dots - S peaks, Brown dots - T peak

Fig. 14. These are examples of mouse's, dog's and rabbit's ECGs. For mice there is no sence in finding P or T waves due to their very low amplitude for this animals. For dog this relates only to T-peak. For rabbits, whose ECG is most close to a human one, we can find all the parameters.*Note that for mouse and dos ECG the R peaks are directed down but the program is still able to recognize them*

5 REFERENCES

(1) Baumert, M., Porta, A., Vos, M. A., Malik, M., Couderc, J. P., Laguna, P., ... Volders, P. G. (2016). QT interval variability in body surface ECG: measurement, physiological basis, and clinical value: position statement and consensus guidance endorsed by the European Heart Rhythm Association jointly with the ESC Working Group on Cardiac Cellular Electrophysiology. Europace, 18(6), 925-944.
(2) Schlögl, A., Brunner, C. (2008). BioSig: a free and open source software library for BCI research. Computer, 41(10), 44-50.
(3) Silva, I., Moody, G. B. (2014). An open-source toolbox for analysing and processing physionet databases in matlab and octave. Journal of open research software, 2(1)
(4) Lenis, G., Pilia, N., Loewe, A., Schulze, W. H., Dössel, O. (2017). 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.
(5) Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.
(6) Cardiac monitors, heart rate meters, and alarms [American National Standard (ANSI/AAMI EC13:2002)]. Arlington, VA: Association for the Advancement of Medical Instrumentation, 2002.
(7) Graps, A. (1995). An introduction to wavelets. IEEE computational science and engineering, 2(2), 50-61.
(8) Zenodo:
https://zenodo.org/search?page=1size=20q=EEGtype=dataset
(9) Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020:
https://physionetchallenges.github.io/2020/rules-and-deadlines
(10) PhysioZoo - mammalian NSR databases :
https://physionet.org/content/physiozoo/1.0.0/