D1NAMO Data Wrangling Challenge

Author: Qiuhan Jin
Contact: jinqiuhan1996@gmail.com
Date: 21/08/2024

Introduction

The open-source D1NAMO dataset is a multi-modal dataset prepared for developing type-I diabetes management systems based on wearable devices in non-clinical conditions [1]. It consists of chest belt measured ECG signals, breathing signals, and accelerometers outputs, along with annotated glucose levels and food pictures, acquired on a study that included 20 healthy participants and 9 patients with type-I diabetes. In this report, a signal preprocessing pipeline is described for the manipulation, filtration, and segmentation of the raw measurements for analysis to predict blood glucose measures. Changes in blood glucose levels have been shown to correlate with changes in the shape of the ECG signals, heart rate variability, breathing patterns, food intake, and physical activity levels [2-4]. Food intake estimation based on single-view pictures requires the use of deep learning and is not discussed in this report. Since heart rate variability and breathing patterns can be estimated by ECG signals [5-8], and physical activity levels can be estimated by ECG signals and accelerometer signals [9-11], our method focuses on the preprocessing of ECG and accelerometer data. Codes and data are shared in the GitHub repository: https://github.com/qjin7796/BMSP. Preprocessing algorithms are collected in ../BMSP/bmsp_preprocessing_*.py, and the script and data for generating the figures are documented in ../BMSP/demo.

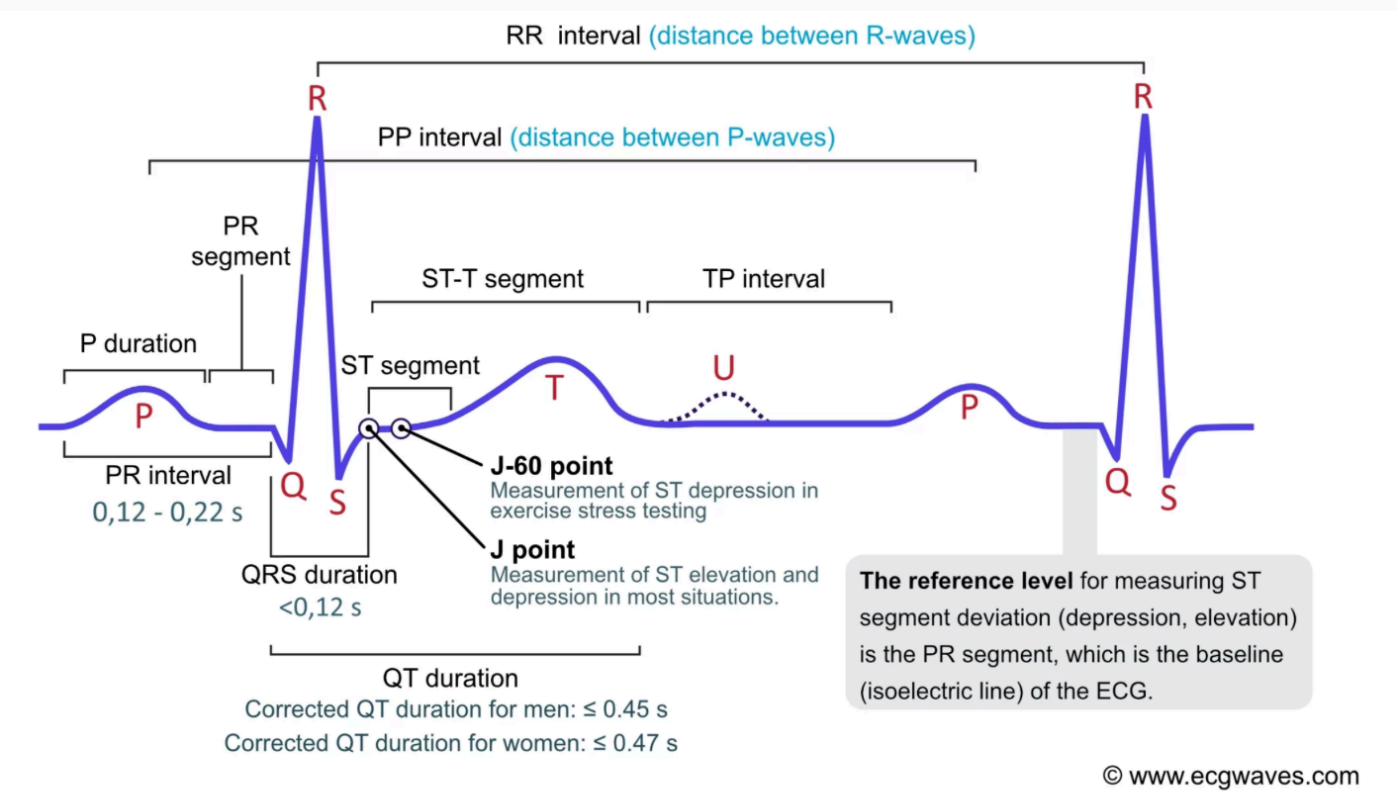

Figure 1. The classical ECG curve with its most common waveforms, important intervals and points of measurement.

Methods

The pipeline includes three main nodes: ECG signal preprocessing, accelerometer (ACC) signal preprocessing, and ECG-accelerometer signal alignment. As shown in Figure 2, the preprocessing nodes include signal cleaning (handling missing values, outliers, and noise), epoch creation (segmenting signal into epochs based on events), and basic feature extraction (such as heart rate, acceleration magnitude etc.). The pipeline is written in Python with a large part using functions adapted from the Neurokit2 pacakge [12].

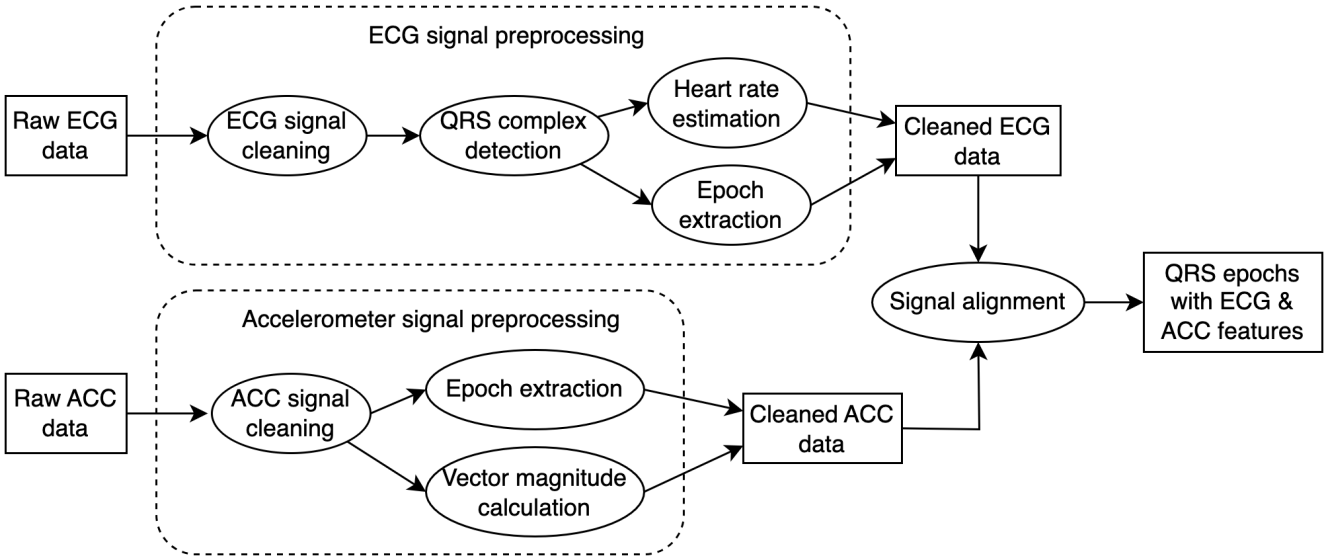

Figure 2. Overview of the signal preprocessing pipeline.

ECG signal preprocessing

The ECG signal preprocessing node first cleans the raw signal by (1) interpolating missing values and outliers (we use linear interpolation for simplicity), (2) filtering out potential noise, and (3) estimating a continuous index of signal quality.

There are three main types of noise present in raw ECG signals: the baseline wandering noise that generally occurs at a frequency below 0.5 Hz, powerline interference at 50 Hz (in Europe), and myoelectric interference the frequency of which can vary between 30 and 300 Hz [13-14]. In general, the majority energy of an ECG signal is in the range of 0.5-35 Hz [15]. So these two effects can be filtered out with little risk of distorting the true ECG signals. In our pipeline, we use Neurokit2's ecg_clean(method='neurokit') for signal denoising, which includes a 0.5 Hz high-pass butterworth filtering followed by a 50 Hz powerline filtering. The noise located in other frequency bands are not filtered in the preprocessing step, but it should be considered at a later stage of designing the glucose level prediction method.

The filtered ECG signals are then screened with signal quality assessment based on the QRS segments present in the samples. An electrical impulse from a heart beat captured by the ECG is called a QRS complex, which represents the ventricular depolarization, a complete circle of heart contraction and relaxation [16]. A typical QRS shape includes a sharp upward deflection (R wave) and two downward reflections occurring before (Q wave) and afterwards (S wave), as shown in Figure 1. We use the 'average QRS' method from Neurokit2 to assess the ECG signal quality. This

method first detects all potential QRS segments from the data, and then computes a continuous index of quality by interpolating the distance of each QRS segment from the average QRS segment. This index is relative to the average quality of the samples, and therefore should be used alongside an overall quality check (for example plotting the average QRS shape). If the majority of the samples are bad, this quality index should not be used [17]. Instead, one may consider dynamic or probabilistic artifact detection methods for signal quality assessment [18].

An example of the ECG signal cleaning effect is shown in Figure 3. Three segments (3-minute long) of raw ECG signal are plotted in gray lines and the cleaned signals in red lines. The signal quality index is illustrated by the size of the light green area with the 80% signal quality shown in green dashed lines.

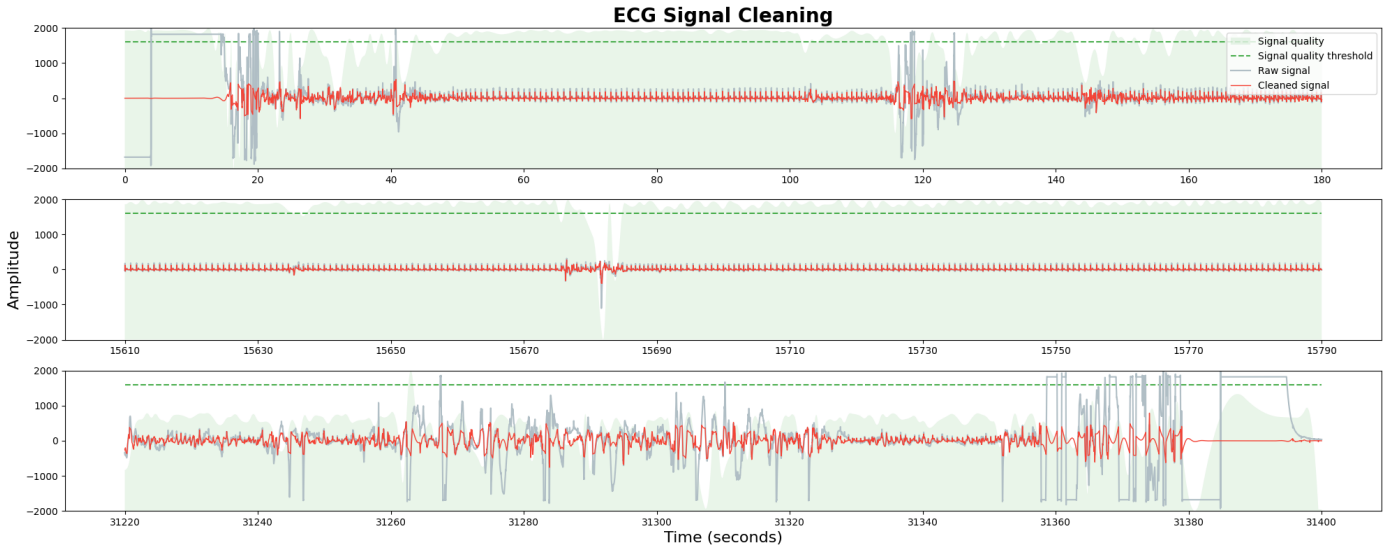

Figure 3. Example of ECG signal before and after cleaning. Three segments of ECG samples are selected from the D1NAMO healthy subject 001 measurement on 2014/10/01. The three segments are taken from the first, the middle, and the last 3 minutes of the measurement. Raw and cleaned signals are centered and plotted in gray and red lines, respectively. The signal quality index over time is indicated by the light green area. The green dashed line represents the 80% signal quality compared to the sample average.

Following ECG signal cleaning, heart rate is estimated based on the R-peaks extracted in the QRS segmentation step. The formula for computing the heart rate is 60/period (beats per minute), where period is the R-peak interval in seconds. All QRS segments are further delineated into a group of waveforms using the ecg_delineate() function from Neurokit2. Specifically, the following events are localized: P-waves (onsets, peaks and offsets), Q-waves (peaks), R-waves (onsets, peaks and offsets), S-waves (peaks), and T-waves (onsets, peaks and offsets). Based on the ECG signal quality and the QRS waveforms, we then extract QRS epochs that meet two conditions: (1) signal quality above 80% throughout the epoch time range; (2) R-peaks present within the epoch. Figure 4 illustrates the QRS epochs extracted from a one-minute ECG sample. The cleaned ECG signal, R-peaks, and heart rate are shown on the left panel. The overlay of all QRS epochs, averaged epoch shape, and the location of waveform peaks are displayed on the right panel. These features describe the fundamental properties of ECG signal patterns in the time domain. A great number of features can be derived from the waveforms, such as the kurtosis and skewness features of the epoch distribution, local energy of the peak frequency, autocorrelation of the waveforms over time, the time-frequency transform of the epochs and so on [19].

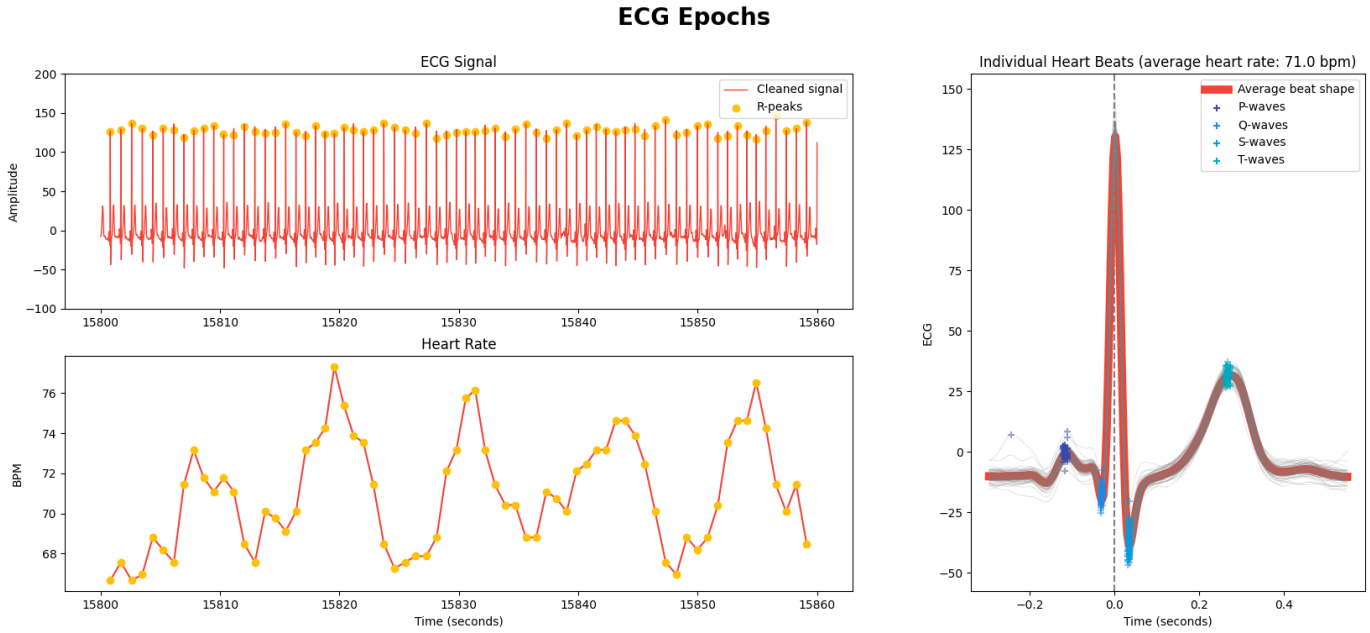

Figure 4. Example of QRS epochs extracted from cleaned ECG signal. A one-minute ECG signal sample is selected from the D1NAMO healthy subject 001 measurement on 2014/10/01. The upper left plot shows the cleaned signal in red lines and the R-peaks in yellow dots. The estimated heart rate over the period is plotted in red lines in the lower left graph. The right panel plots all QRS epochs in light gray lines with R-peaks aligned, shows an averaged epoch shape in red curve, and marks the peaks of the P-, Q-, S- and T-waves in blue pluses (exact color legend shown at the top right).

Accelerometer signal preprocessing

The accelerometer (denoted as ACC) signal preprocessing node first cleans the raw signal by interpolating missing values and filtering out potential noise. Ideally the accelerometer output should be calibrated to local gravity. But such information is lacking in the D1NAMO dataset, so we design the preprocessing steps based on the signal's relative values. In the cleaning step here, we do not apply the outlier remover implemented in the ECG preprocessing node due to two reasons: (1) the variance of an ACC signal sample can be very small if physical activity is sparse, so thresholding the signal by variance risks smoothing out meaning peaks; (2) unlike ECG signals, the outliers in ACC signals occur at a relatively lower frequency, making a high-pass filter sufficient for removing the outliers. With regard to signal denoising, we apply a 0.5-10 Hz band-pass filter using Neurokit2's signal_filter(). We set the low frequency cut-off at 0.5 Hz due to the fact that the accelerometer sensor used for obtaining the D1NAMO dataset was embedded together with the ECG sensors in the chest belt [1], and the low frequency cut-off is therefore set at the same level as that used for denoising the ECG signals. For the high frequency cut-off, we choose 10 Hz because the acceleration related to human physical activity is found at frequencies up to 10 Hz, with the majority of normal human step frequency below 4 Hz [20, 21].

An example of the signal cleaning effect is shown in Figure 5, a 3*3 plot that displays three two-minute segments of an ACC signal sample in rows and the three channels of the signal in columns. The three channels shown in red, green, and blue lines correspond to the acceleration in the vertical, lateral, and sagittal directions, respectively. The raw signals are centered and shown in light gray lines beneath the cleaned signals in color.

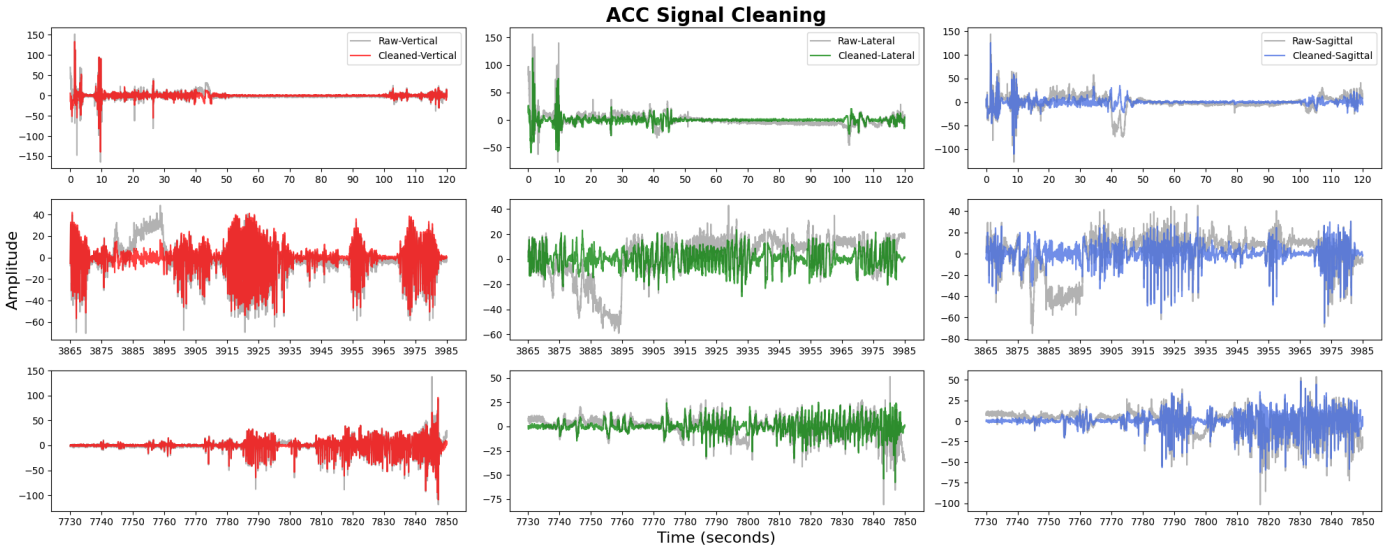

Figure 5. Example of ACC signal before and after cleaning. Three segments of ACC samples are selected from the D1NAMO healthy subject 001 measurement on 2014/10/01. The three segments are taken from the first, the middle, and the last 2 minutes of the measurement, and are plotted in three rows. The three columns correspond to the acceleration signal in the vertical (red lines), lateral (green lines), and sagittal (blue lines) direction, respectively. Raw and cleaned signals are centered and plotted in gray and colored lines, respectively.

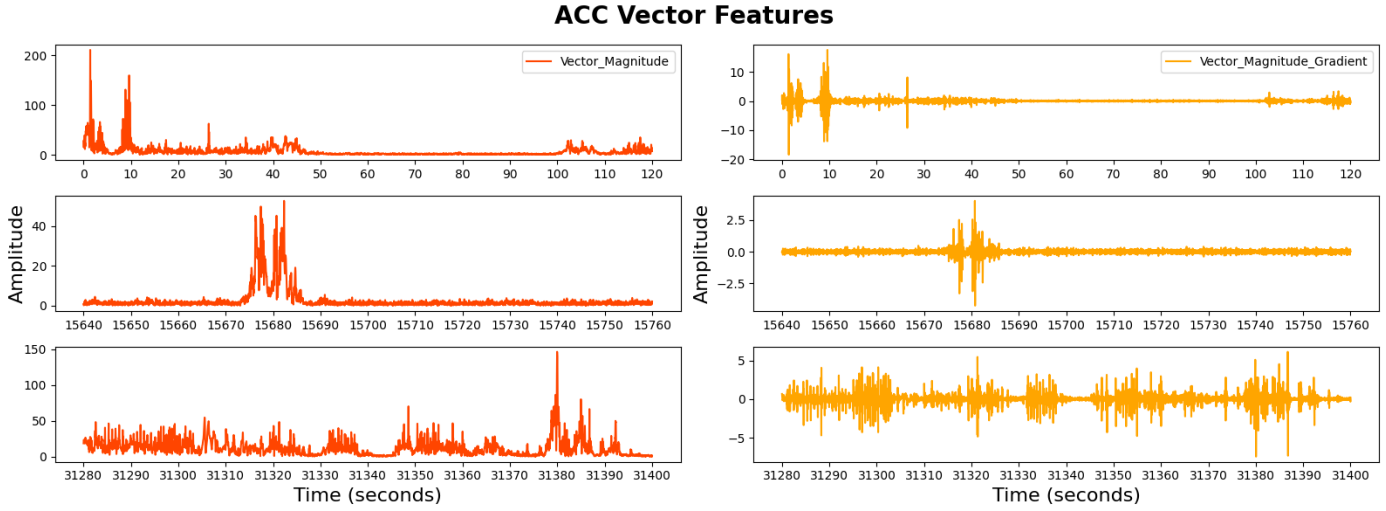

Figure 6. Example of acceleration vector magnitude and gradient of cleaned ACC signal. The three rows correspond to the same samples used in Figure 5. Acceleration vector magnitude is shown in red lines and its gradient is shown in orange lines.

In the next step, we compute the magnitude of the acceleration vector and its gradient to quantify the intensity level of acceleration and its variation over time [22]. The magnitude of the acceleration vector is simply taken as the vector's l2-norm. And the gradient is computed by division of the differences in the vector magnitude and the differences in time.

Similar to the idea of QRS segmentation in processing ECG signals, splitting ACC signals into short segments quantifies the signal local patterns and provides the basis for identifying physical active time windows [23]. For simplicity, we only extract ACC epochs of fixed length and overlapping size. For each epoch, we compute the mean and variance of every signal channel as

well as the acceleration vector magnitude and gradient. Figure 6 visualizes the acceleration vector magnitude and its gradient of the same three two-minute ACC signal samples used in figure 5. Comparing the second row of the two figures, it can be seen that vector magnitude and gradient provide different information about the ACC signals - the acceleration in different directions may show high variance at multiple time points while the vector magnitude and gradient reveal a time point that is significantly different from others in activity intensity. These simple features are useful for estimating physical activity level and localizing activity in detailed temporal positions. Based on the epochs of varying acceleration magnitude variances, one can compute cumulative relative time active or activity volume, temporal correlation and regularity, activity trajectories, and subdivision of activity categories [24-26].

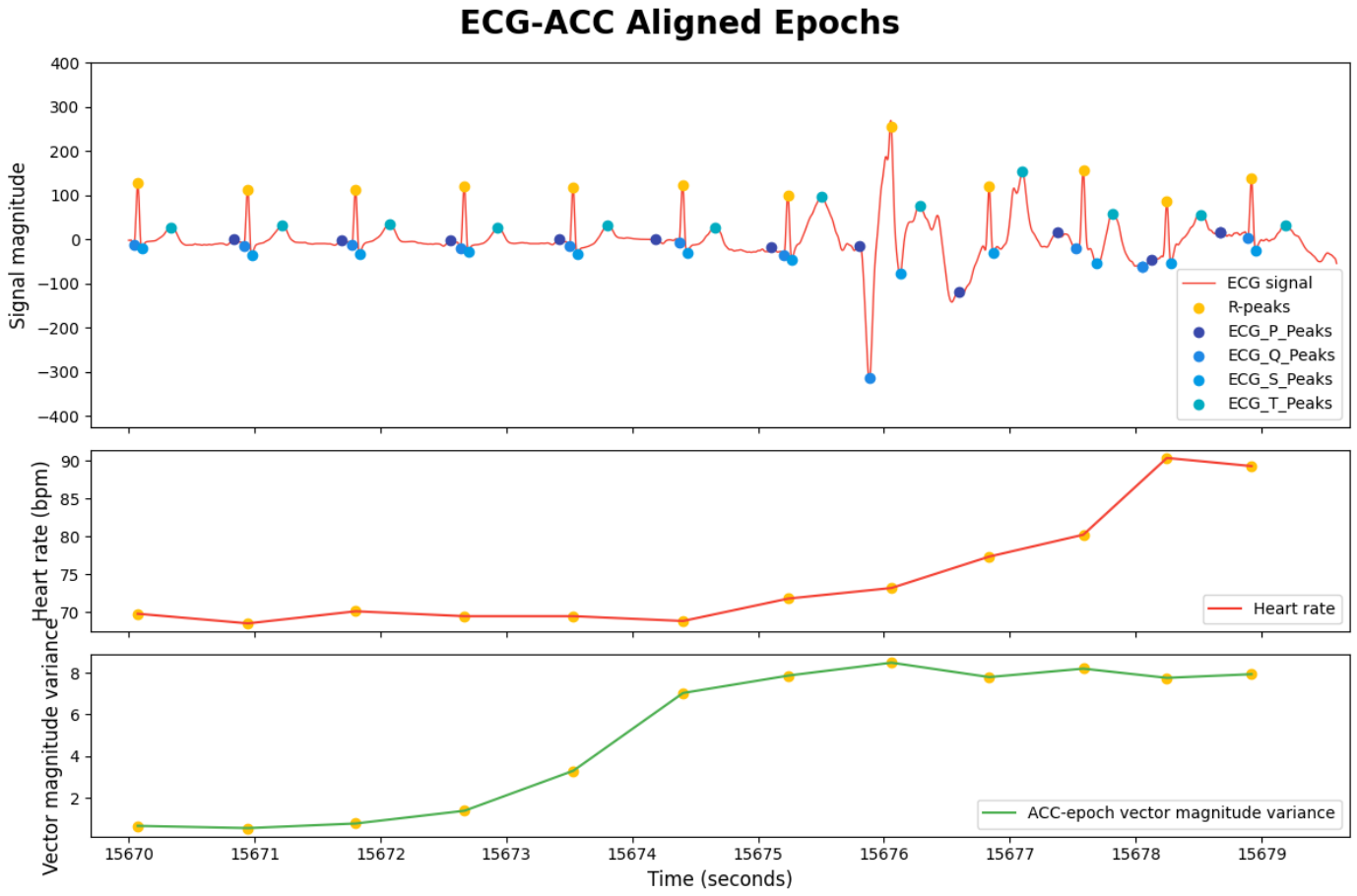

Figure 7. Example of QRS epochs extracted from cleaned ECG signal with aligned acceleration features. A ten-second sampling period is selected from the D1NAMO healthy subject 001 measurement on 2014/10/01. The cleaned ECG and accelerometer signals for the specific time period are extracted from their preprocessing readout. The top plot shows the cleaned ECG signal in red lines, the R-peaks in yellow dots, and the peaks of the P-, Q-, S- and T-waves in blue pluses (exact color legend shown at the bottom right). The middle plot shows heart rate changes in red lines and R-peaks in yellow dots. The bottom plot shows acceleration vector magnitude in green lines, which each point corresponding to a four-second epoch centered at the R-peaks marked in yellow dots.

Multi-modal signal alignment

In practice, the collected multi-modal signal samples may not be synchronized, especially when measured by different devices. Therefore, we add a signal alignment module for matching ECG and ACC signals in time and updating QRS complexes readout with additional information of the

subject's acceleration patterns. This signal alignment module includes a signal resampling step that aligns the signal sampling rate, along with a QRS epoch refining step that selects QRS epochs with available acceleration signal and appends acceleration signal features into the QRS epochs' output. Figure 7 shows a one-minute time window with the refined QRS epochs, heart rate, and acceleration vector magnitude.

Conclusion and discussion

This report describes a biomedical signal preprocessing pipeline featuring ECG and accelerometer measurements provided by the D1NAMO dataset. The pipeline efficiently cleans, segments, and extracts basic features from the raw signal, and aligns the multi-modal features in time. This includes filling missing values, handling outliers, filtering noises, locating important events, and summarizing basic features of the signal. The extracted features encompass heart rate and QRS waveforms from the ECG signal, and acceleration magnitude and its variation over time from the accelerometer signal. Our pipeline prepares the ECG and accelerometer data for further analysis of feature engineering and machine-learning based glucose level prediction.

Limitations and future work

Limitations of the pipeline and potential solutions are discussed below. Firstly, the pipeline's data processing speed exhibits a slightly exponential decrease trend with increasing amount of data (over 8 hours). This is mainly due to the ECG signal delineation algorithm we use from Neurokit2, where the signal indexing can be improved. Leveraging a multiprocessing system such as Joblib will further enhance the performance of the pipeline.

In addition, a limited number of denoising methods are explored in this report for efficiency concerns, but in practice it's worth the test depending on the problems we face and the machine learning methods we use in the downstream analysis. We will explore methods such as signal detrending to handle signal that shows linear or nonlinear trends, as well as machine learning based artifact classification methods for adaptive denoising [17, 27]. The flexible structure of our pipeline makes it simple to embed different algorithms.

Last but not least, the pipeline's epoch extraction searches for single epochs at a fixed rate, which is not optimal when certain events that are located slightly distant from each other should be grouped into one epoch. Data segregation methods will be explored to dynamically adjust epoch size. We will also identify autocorrelation patterns in the time and frequency domains to link epochs across time.

References

1. Dubosson F, Ranvier JE, Bromuri S, Calbimonte JP, Ruiz J, Schumacher M. The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management. Informatics in Medicine Unlocked. 2018;13:92-100. doi:10.1016/j.imu.2018.09.003

2. Stern S, Sclarowsky S. The ECG in diabetes mellitus. Circulation. 2009;120(16):1633-1636. doi:10.1161/CIRCULATIONAHA.109.897496
3. Harms PP, van der Heijden AA, Rutters F, et al. Prevalence of ECG abnormalities in people with type 2 diabetes: The Hoorn Diabetes Care System cohort. Journal of Diabetes and its Complications. 2021;35(2). doi:10.1016/j.jdiacomp.2020.107810
4. Im S il, Kim SJ, Bae SH, et al. Real-time heart rate variability according to ambulatory glucose profile in patients with diabetes mellitus. Frontiers in Cardiovascular Medicine. 2023;10. doi:10.3389/fcvm.2023.1249709
5. C.Dela Cruz J, Ibera J, M. Alcoy J, Erick R. Tulio C. Deriving Heart Rate and Respiratory Rate from ECG Using Wavelet Transform. In: Proceedings of the 2021 11th International Conference on Biomedical Engineering and Technology. ICBET '21. Association for Computing Machinery; 2021:100-105. doi:10.1145/3460238.3460254
6. Roberts JD, Walton RD, Loyer V, Bernus O, Kulkarni K. Open-source software for respiratory rate estimation using single-lead electrocardiograms. Scientific Reports. 2024;14(1). doi:10.1038/s41598-023-50470-0
7. Charlton PH, Birrenkott DA, Bonnici T, et al. Breathing Rate Estimation from the Electrocardiogram and Photoplethysmogram: A Review. IEEE Reviews in Biomedical Engineering. 2018;11:2-20. doi:10.1109/RBME.2017.2763681
8. Bao X, Abdala AK, Kamavuako EN. Estimation of the respiratory rate from localised ecg at different auscultation sites. Sensors (Switzerland). 2021;21(1):1-11. doi:10.3390/s21010078
9. Wientzek A, Vigl M, Steindorf K, et al. The improved physical activity index for measuring physical activity in EPIC Germany. PLoS ONE. 2014;9(3). doi:10.1371/journal.pone.0092005
10. Cosoli G, Antognoli L, Scalise L. Wearable Electrocardiography for Physical Activity Monitoring: Definition of Validation Protocol and Automatic Classification. Biosensors. 2023;13(2). doi:10.3390/bios13020154
11. Karas M, Bai J, Strączkiewicz M, et al. Accelerometry Data in Health Research: Challenges and Opportunities: Review and Examples. Statistics in Biosciences. 2019;11(2):210-237. doi:10.1007/s12561-018-9227-2
12. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods. 2021;53(4):1689-1696. doi:10.3758/s13428-020-01516-y
13. Leif Sörnmo, Pablo Laguna. Bioelectrical Signal Processing in Cardiac and Neurological Applications.; 2005.
14. Al Hinai N. Chapter 1 - Introduction to biomedical signal processing and artificial intelligence. In: Zgallai W, ed. Biomedical Signal Processing and Artificial Intelligence in Healthcare. Developments in Biomedical Engineering and Bioelectronics. Academic Press; 2020:1-28. doi:https://doi.org/10.1016/B978-0-12-818946-7.00001-9
15. Mvuh FL, Ebode Ko'a COV, Bodo B. Multichannel high noise level ECG denoising based on adversarial deep learning. Scientific Reports. 2024;14(1). doi:10.1038/s41598-023-50334-7
16. Bagliani G, de Ponti R, Gianni C, Padeletti L. The QRS Complex: Normal Activation of the Ventricles. Cardiac Electrophysiology Clinics. 2017;9(3):453-460. doi:10.1016/j.ccep.2017.05.005
17. Rahman S, Karmakar C, Natgunanathan I, Yearwood J, Palaniswami M. Robustness of electrocardiogram signal quality indices. Journal of the Royal Society Interface. 2022;19(189). doi:10.1098/rsif.2022.0012
18. Van der Bijl K, Elgendi M, Menon C. Automatic ECG Quality Assessment Techniques: A Systematic Review. Diagnostics. 2022;12(11). doi:10.3390/diagnostics12112578

19. Singh AK, Krishnan S. ECG signal feature extraction trends in methods and applications. BioMedical Engineering Online. 2023;22(1). doi:10.1186/s12938-023-01075-1
20. Fridolfsson J, Börjesson M, Arvidsson D. A biomechanical re-examination of physical activity measurement with accelerometers. Sensors (Switzerland). 2018;18(10). doi:10.3390/s18103399
21. Fridolfsson J, Börjesson M, Buck C, et al. Effects of frequency filtering on intensity and noise in accelerometer-based physical activity measurements. Sensors (Switzerland). 2019;19(9). doi:10.3390/s19092186
22. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Scientific Reports. 2018;8(1). doi:10.1038/s41598-018-26174-1
23. Doherty A, Jackson D, Hammerla N, et al. Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. PLoS ONE. 2017;12(2). doi:10.1371/journal.pone.0169649
24. Bai J, He B, Shou H, Zipunnikov V, Glass TA, Crainiceanu CM. Normalization and extraction of interpretable metrics from raw accelerometry data. Biostatistics. 2014;15(1):102-116. doi:10.1093/biostatistics/kxt029
25. Bai J, Di C, Xiao L, et al. An activity index for raw accelerometry data and its comparison with other activity metrics. PLoS ONE. 2016;11(8). doi:10.1371/journal.pone.0160644
26. Backes A, Gupta T, Schmitz S, Fagherazzi G, van Hees V, Malisoux L. Advanced analytical methods to assess physical activity behavior using accelerometer time series: A scoping review. Scandinavian Journal of Medicine and Science in Sports. 2022;32(1):18-44. doi:10.1111/sms.14085
27. Moeyersons J, Smets E, Morales J, et al. Artefact detection and quality assessment of ambulatory ECG signals. Comput Methods Programs Biomed. 2019;182:105050. doi:10.1016/j.cmpb.2019.105050