

Ten quick tips for electrocardiogram (ECG) signal processing

Davide Chicco^{1,2}, Angeliki-Ilektra Karaiskou³ and Maarten De Vos^{3,4}

¹ Dipartimento di Informatica Sistemistica e Comunicazione, Università di Milano-Bicocca, Milan, Italy

² Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

³ STADIUS Center for Dynamical Systems Signal Processing and Data Analytics, Katholieke Universiteit Leuven, Leuven, Belgium

⁴ Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium

ABSTRACT

The electrocardiogram (ECG) is a powerful tool to measure the electrical activity of the heart, and the analysis of its data can be useful to assess the patient's health. In particular, the computational analysis of electrocardiogram data, also called ECG signal processing, can reveal specific patterns or heart cycle trends which otherwise would be unnoticeable by medical experts. When performing ECG signal processing, however, it is easy to make mistakes and generate inflated, overoptimistic, or misleading results, which can lead to wrong diagnoses or prognoses and, in turn, could even contribute to bad medical decisions, damaging the health of the patient. Therefore, to avoid common mistakes and bad practices, we present here ten easy guidelines to follow when analyzing electrocardiogram data computationally. Our ten recommendations, written in a simple way, can be useful to anyone performing a computational study based on ECG data and eventually lead to better, more robust medical results.

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Data Science, Programming Languages, Software Engineering

Keywords ECG, Electrocardiography, Signal processing, Medical signal processing, Cardiology, Quick tips, Guidelines

Submitted 22 May 2024
Accepted 9 August 2024
Published 3 September 2024

Corresponding author
Davide Chicco,
davide.chicco@gmail.com

Academic editor
Valentina Emilia Balas

Additional Information and
Declarations can be found on
page 14

DOI 10.7717/peerj-cs.2295

© Copyright
2024 Chicco et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

An electrocardiogram (ECG) is a time-based recording of the electrical activity of the heart through several through repeated cardiac cycles (heartbeats). The ECG signal is usually recorded through electrodes put on the skin of a patient, and its representation shows the electrical action potential measured in volts on the y axis over time on the x axis. The most common ECG device today is the 12-lead ECG, which has twelve leads that record the signals of the heart (Sörnmo & Laguna, 2006).

The electrocardiogram (ECG) is a special case of electrogram, and should not be confused with the electroencephalogram (EEG), which is the recording of the electrical activity of the brain. The ECG recording can be interpreted manually by a physician or a medical doctor, who can notice some anomalies that could indicate potential pathological conditions. With the spread of computer sciences in the biomedical domain, researchers have started to analyze ECG data through computational algorithms and *ad hoc* software

programs able to reveal heart trends and information that otherwise would be unnoticeable by medical doctors.

This scientific area, mainly developed within the biomedical engineering community, gained the broad name of *ECG signal processing*. Since the 1980s, several computational projects for the analysis of electrocardiogram data have appeared, both from hi-tech companies and from academic scientific research ([Berkaya et al., 2018](#); [Clifford, Azuaje & McSharry, 2006](#); [Gupta et al., 2023a, 2023b, 2023c](#); [Gupta, Mittal & Mittal, 2022](#)). Even if performing computational analyzes of ECG has become easier for software developers, making mistakes during these analyses has become easier, too. We therefore decided to present these quick tips to avoid common errors and bad practices during a computational study on electrocardiogram signals.

To the best of our knowledge, no study reporting guidelines on ECG data analysis exists in the literature nowadays. The scientific literature includes articles of recommendations and guidelines on several topics but none on ECG signal processing. We fill this gap by presenting here our recommendations: even if we designed originally these guidelines for apprentices and beginners, we believe they should followed by experts, too.

TIP 1: BEFORE STARTING, STUDY HOW THE ELECTROCARDIOGRAM WORKS, WHAT IT SAYS AND WHAT IT DOES NOT SAY

This first guideline might seem trivial and straightforward to our readership, but it actually refers to an underrated aspect of electrocardiogram (ECG) analyses. Before turning on your computer to work on the computational analyses of electrocardiogram signals, we advise studying and understanding the electrocardiogram signal first.

In fact, the electrocardiogram signal has a unique data type and structure that requires specific training. Keep that in mind. Before starting an ECG signal processing analysis, learn the main concepts of an electrocardiogram, such as P wave, Q wave, R wave, S wave, J point, T wave, QRS complex, PR segment, ST segment, PR interval, QT interval, J point, corrected QT interval, and others [Fig. 1](#). Learn how to read an electrocardiogram by yourself with your own eyes, before turning the computer on. Then, only when you are able to understand what an ECG says and what it does not say, turn on your computer and work on your computational analysis.

Multiple resources to study the ECG are available online, including several free educational videoclips on YouTube. We recommend studying the content of ECGpedia ([De Jong, 2023](#)), an online encyclopedia similar to Wikipedia, started by Jonas de Jong and fully dedicated to the electrocardiogram theme.

And how do you know if you understood the meanings of the ECG well enough? Explain it to someone who has no medical knowledge, and see if they comprehend. The mathematician David Hilbert once said: “A mathematical theory is not to be considered complete until you have made it so clear that you can explain it to the first man whom you meet on the street”. That also applies to the ECG understanding.

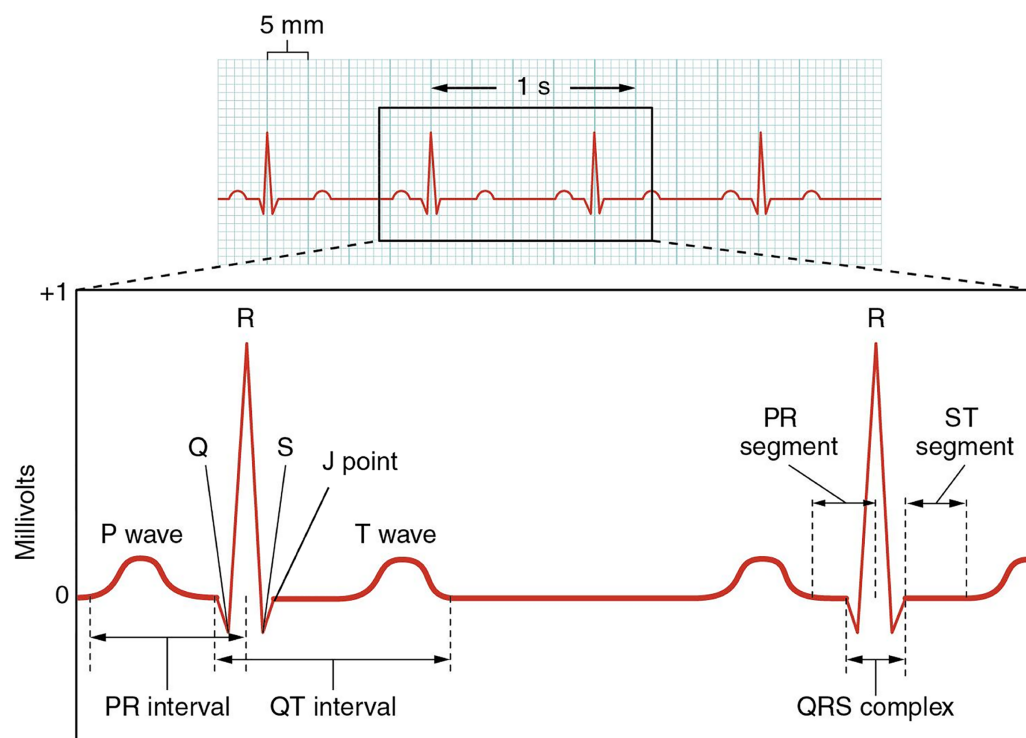


Figure 1 Schematic example of an electrocardiogram (ECG). Image publicly available under the Creative Commons Attribution 3.0 Unported (CC BY 3.0 DEED) license on Wikimedia Commons (CFCE, 2023). We advise learning and understanding the main concepts of an electrocardiogram (for example, P wave, Q wave, R wave, S wave, J point, T wave, QRS complex, PR segment, ST segment, PR interval, QT interval, J point and corrected QT interval) before performing the computational analysis. Would you be able to explain this image to first person around you?

Full-size DOI: 10.7717/peerj-cs.2295/fig-1

TIP 2: UNDERSTAND THE MEDICAL CONTEXT AND THE MEDICAL PROBLEM YOU ARE TRYING TO SOLVE

Once you have clearly apprehended what the ECG says, it may be tempting to dive directly into your computational tools, but before starting this, you must clearly define your scientific question and be sure what your aim and potential pitfalls are. Furthermore, given the enormous number of medical data, it can be overwhelming to understand what is valuable information and what is a nuisance. Multiple factors can affect the data, and without the appropriate training, it is difficult to decipher if there is a problem with your methodology, if the medical context of your problem limits you, or if it is influenced by the data acquired from a population with specific characteristics. Naturally, this also applies in the case of ECG data. Thus, now it is time to focus on understanding the medical context and problem you are facing.

Before all else, you should learn how to interpret the electrocardiogram (ECG) signals you see based on the problem you are trying to solve. Medical doctors and physicians have the experience to distinguish with a bare eye not only the abnormalities of the ECG but also to give a more specific diagnosis of what these abnormalities represent. However, do not rely only on experts' labels; even experienced medical experts provide contradicting

diagnoses. Therefore, you should look into these abnormalities and learn how to extract patterns so that when you apply your algorithms, you are sure that what you captured is what you expected to catch. Take, for example, the medical problem of arrhythmia, where there are multiple categories with diverse characteristic manifestations, excessively slow or fast heartbeats such as sinus bradycardia and atrial tachycardia, irregular rhythm with missing or distorted wave segments and intervals, or both (Zheng et al., 2020). If your scientific problem involves a specific target group, such as athletes (Drezner, 2012), then focus on the ECG abnormalities relevant to this particular population. Additionally, on the assumption that you work with 12-lead ECG signal, pay attention to understanding what kind of information can be extracted and interpreted from each lead (Garcia, 2015).

Nevertheless, to master the electrocardiogram (ECG), collaboration with medical experts is key. Discuss your scientific problem with your clinicians, including cardiologists or medical doctors, to define and assess your goals from the beginning. Consistently share your results with them to evaluate and review them. Optimally, seek experts close to your workplace so that you can easily access their valuable knowledge; however, if this is impossible, consider connecting with healthcare professionals online. In that way, you are assured that your scientific goals stay relevant.

That being said, pay attention to the scientific breakthroughs relevant to the scientific problem you are working on. Dive into the literature and be informed of the most representative features and models used in previous works relevant to your medical problem. Then, try to reproduce those results. There might be a chance that your ECG entangles information for various medical problems, and therefore, you should be careful about what datasets you are going to use, as we see in the next tip.

TIP 3: UNDERSTAND IF YOU HAVE DATA SUITABLE FOR MAKING PROGRESS ON THE SCIENTIFIC PROBLEM YOU ARE INVESTIGATING

After gaining a solid grasp of the medical context and issue at hand, the next step is the selection of the appropriate dataset to address your scientific question. Understanding the suitability of data is a ground step in any scientific study, and it is no different when it comes to electrocardiogram (ECG) signal processing. The relevance and quality of the data you work with significantly impact the progress and success of your scientific work.

Above all else, your data must be directly relevant to the scientific question you are trying to solve. ECG signal studies typically focus on specific health issues, conditions, states, or populations. Being assured that the data you are going to analyze were collected from the appropriate medical context is of very high importance in order to achieve meaningful results. Irrelevant data can lead to inaccurate conclusions and wasted research efforts.

The quality of your data is of equal importance. You should be able to understand the quality of your signal and ensure that what you see is not an artifact (Fig. 2). Once again, this is directly connected with the medical context of the problem that you are facing. Was the ECG acquired during movement, sleep, or during a 24-hours span (Moeyersons et al., 2019; Pérez-Riera et al., 2018; Carr et al., 2018; Carr, de Vos & Saunders, 2018; Gregg et al., 2008; Clifford, 2006)? Comprehending this information and its effect on your

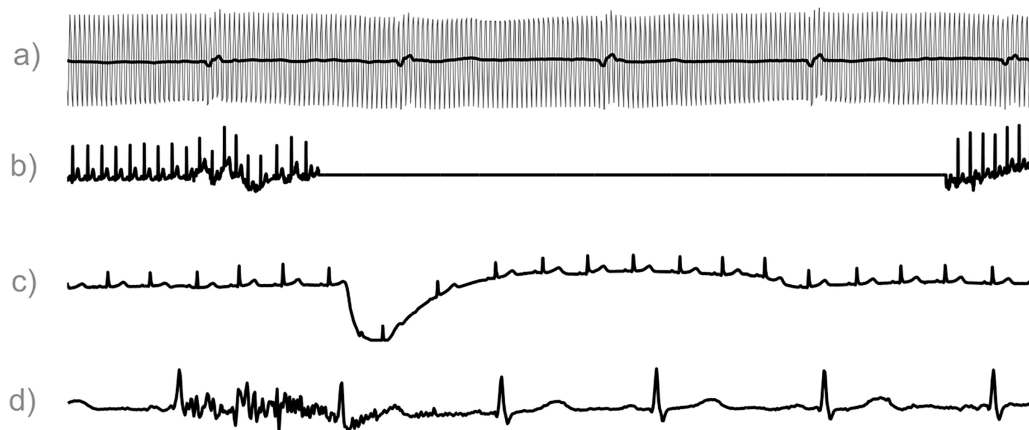


Figure 2 ECG Segments contaminated by artifacts. Image taken from *Varon, van Huffel & Suykens (2015)* with the authors' permission. (A) In grey the original electrocardiogram (ECG) contaminated by power line interference at 50 hertz is indicated, and in black its filtered version. (B) Contact noise. (C) Electrode motion. (D) Muscle artifact. [Full-size !\[\]\(b345a1c4255362eec3746050dd71ccac_img.jpg\) DOI: 10.7717/peerj-cs.2295/fig-2](https://doi.org/10.7717/peerj-cs.2295/fig-2)

electrocardiogram (ECG) data will save you a lot of time and energy when you are trying to find a solution. The assessment of ECG signal quality is an active topic of research, and various quality indices have been developed for this reason, as we will see later (Tip 5).

Data quantity is also critical. An insufficient amount of data can result in statistical insignificance, limiting the complexity of your analyses and the depth of your insights. A large dataset enhances the statistical power of your research, without falling into the *p*-hacking trap (*Chén et al., 2023; Makin & de Xivry, 2019*). The length of the data you are using must also be taken into account. Do you want to perform peak detection? Usually, the algorithms suggested in the literature require at least 10 seconds of continuous ECG recordings to have relatively robust results. Still, if you have longer windows, this would be much more beneficial. Is your aim to study the heart rate variability (*Clifford, 2002; Ernst, 2014*)? Then, you should possess recordings of at least 3 minutes of ECG length. Moreover, the time you record data can change the variability numbers a lot. So, remember to avoid comparing heart rate variability numbers from different time lengths (*Cygankiewicz & Zareba, 2013*).

Last but not least, you should pay attention to the ECG's sampling rate since it can seriously affect various measures derived by the electrocardiogram (ECG) signal analysis. Reduced sampling rates can compromise the precision of detecting R-wave fiducial points and, therefore, greatly affect your Heart Rate Variability (HRV) analysis (*Garcia-Gonzalez, Fernandez-Chimeno & Ramos-Castro, 2004; Mahdiani et al., 2015; Ellis et al., 2015; Petelczyc et al., 2020; Lu et al., 2023*). In fact, for clinical applications, the ECG sampling frequency that is nowadays recommended is at least 512 hertz (Hz) (*Petelczyc et al., 2020*). Besides, the population on which your medical context is based plays an essential role in deciding the optimal sampling frequency. For instance, if your research involves the analysis of fetal ECG, it is recommended that a good dataset should be acquired with a sampling frequency of 2 kHz and resolution of 16 bits (*Sameni & Clifford, 2010*).

TIP 4: USE ONLY OPEN SOURCE PROGRAMMING LANGUAGES AND SOFTWARE PLATFORMS, IF SUFFICIENT FOR YOUR SCIENTIFIC PROJECT

When one starts working on a computational project, they have to face an important question: which programming language should I use? Our recommendation for electrocardiogram (ECG) analysis is the same we already provided in other studies: choose an open source programming language such as Python, R, or GNU Octave, if their functionalities are sufficient for your scientific goals.

Employing open source code, in fact, would bring multiple advantages. First, open source programming languages are free, and one does not need to spend their money or their institution's money to purchase a license, as it would happen for proprietary software. Second, an open source scripts can always be shared freely and openly among colleagues and collaborators. If one needs to send their software scripts to other people so that they could use them on their computers, only open source programs would give this possibility. On the contrary, if the ECG analysis was performed with proprietary software scripts, they could be shared and used only with other developers who have a license to that proprietary programming language.

Third, open source programs can be used and re-used by a developer even in case they changed laboratories, institutes, or companies. If one develops an ECG analysis script in Python, for example, and then moved to another scientific laboratory, they would still be able to re-use their own script in the new work environment. On the contrary, if they developed their script in an proprietary programming language, and the new work environment did not have the license for their usage, they would be unable to re-use it and would need to re-implement it.

It is also relevant to mention that currently, Python is the most popular programming language in the world, according to the PYPL index ([PYPL, 2023](#)), to the TIOBE index of November 2023 ([TIOBE, 2023](#)) and the Kaggle 2022 survey ([Kaggle, 2022](#)). R has excellent ranks in these standings as well: 7th most employed programming language worldwide in November 2023 for the PYPL index ([PYPL, 2023](#)), 19th at the same date for the TIOBE index ([TIOBE, 2023](#)), and 3rd in the Kaggle 2022 survey ([Kaggle, 2022](#)). Therefore, if a student or an apprentice coder decided to invest their time and energy to learn Python or R for ECG signal processing, it would be likely that they could take advantage of their gained skills in other projects as well.

Both Python and R provide several software packages for electrocardiogram signal processing. Python furnishes the popular HeartPy ([Van Gent et al., 2019](#); [Van Gent et al., 2018](#)) and other software libraries such as ecgtools ([pypi ecgtools, 2024](#)), ecg-quality ([pypi ecg-quality, 2023](#)), ndx-ecg ([pypi ndx-ecg, 2023](#)), PhysioZoo ([PhysioZoo, 2023](#); [Behar et al., 2023](#)), pyPPG ([Goda, Charlton & Behar, 2024](#)), and ecghelper ([pypi ecghelper, 2023](#)), while R supplies RHRV ([Martínez et al., 2017](#); [The Comprehensive R Archive Network, 2023](#)).

Of course, we recommend the usage of open source programming languages only if their libraries and functionalities are sufficient and suitable to solve the scientific problem analyzed.

If a user understood that a particular analysis on electrocardiogram data would be possible only with proprietary software, they should go for it. For all the other cases, we advise to choose open source.

TIP 5: TAKE CARE OF NOISE AND OF BASELINE WANDER

Broadly speaking, biomedical datasets for computational analyses can belong to two types: ready to be used and unready to be used. Usually, datasets containing data from electronic health records (EHRs) of patients with a particular disease, collected within a hospital, can be considered ready to be used for a computational analysis. These datasets often contain data from laboratory tests and need no preprocessing before usage (*Chicco & Jurman, 2020a*). However, for datasets of other types need an effective preprocessing step on the raw data before being ready to undergo a computational analysis. For example, medical images' data need noise removal (*Chicco & Shiradkar, 2023*), while microarray and RNA-seq gene expression data need batch correction (*Sprang, Andrade-Navarro & Fontaine, 2022; Chen et al., 2011*).

Electrocardiogram data belong to this second category: you need to remove noise before doing your scientific analysis on these data. Several filtering methods to remove electrocardiogram (ECG) noise exist in the scientific literature (*Sulthana, Rahman & Mirza, 2018; Salman, Rao & Rahman, 2018; Castroflorio et al., 2013; Zhang et al., 2014; Chui et al., 2016; Nejadgholi, Moradi & Abdolali, 2011; Afkhami, Azarnia & Tinati, 2016; Nuthalapati et al., 2019; Li, Rajagopalan & Clifford, 2013*). ECG data also suffer from *baseline wander*, which is an artifact noise signal generated by the movements of the patients, including breathing. Moreover, peripheral or implanted devices can also generate those artifacts (*Fig. 2*). ECG data therefore necessitate baseline wander correction (*Blanco-Velasco, Weng & Barner, 2008; Leski & Henzel, 2005; Kaur, Singh & Seema, 2011; Zhang, 2006; Sörnmo, 1993; Jane et al., 1992*) as one of the first preprocessing step.

Nevertheless, removing the baseline wander from the ECG does not guarantee that the signal is now ready for analysis. Thus, the use of multiple quality metrics for the electrocardiogram (ECG) is also essential. Various methods have been suggested to assess the quality of ECG recordings automatically. These methods, known as Signal Quality Indices (SQIs), measure the level of interference and evaluate the signal's suitability for its intended purpose.

Some of the benchmark ECG quality metrics that should always be used are the kurtosis of the signal, where larger values of this metric are associated with higher quality, and skewness, where ECG is expected to be highly skewed because of the QRS complex, and therefore, lower values would indicate lower quality of ECG (*He, Clifford & Tarassenko, 2006; Seeuws, De Vos & Bertrand, 2021; Behar et al., 2023*). Similarly, metrics extracted from the frequency domain are of equal significance. Those include the in-band (that means, 5–40 Hz) to out-of-band spectral power ratio (IOR) within the QRS complex, the relative power in the QRS complex (that means, 5–15 Hz), and the relative power in the

baseline (that means, frequency content below 1 Hz) where in all those metrics, the larger their value, the higher the ECG signal quality (Tobón, Falk & Maier, 2014; Seeuws, De Vos & Bertrand, 2021).

Moreover, metrics depending on the beat detection should also be considered. Those include the rSQI metric, which is the ratio of the number of beats detected by the WQRS algorithm (Zong, Moody & Jiang, 2003) to the number of beats detected by the *Epilimited* algorithm (Hamilton, 2002), as well as the bSQI metric which is the ratio of beats detected by the WQRS that matched with beats detected by the *Epilimited*, where again in both metrics the higher the value indicates higher quality of the ECG signal (Behar et al., 2023; Seeuws, De Vos & Bertrand, 2021).

In a nutshell, we recommend you never analyze raw electrocardiogram data as they come from electrocardiogram machines. Always preprocess them, taking care of noise and of baseline wander. Similarly, remember to use SQIs to assess the quality and suitability of your data for the scientific problem you are facing. You can proceed with your computational analysis only when you know your data is adequately preprocessed and suited in terms of quality for your medical problem.

TIP 6: START WITH THE SIMPLEST METHODS, AND USE COMPLEX METHODS ONLY IF IT IS NECESSARY

Common analyses of electrocardiogram signals often include the feature extraction of temporal and morphological variables (Ardeti et al., 2023). To this end, researchers employed traditional electrocardiogram (ECG) signal processing algorithms such as several finite impulse response (FIR) (Lastre-Dominguez et al., 2019), derivative (Lastre-Dominguez et al., 2019), windowing (Oktivasari et al., 2019) and transformation-based algorithms (Fujita, Sato & Kawarasaki, 2015) in the past.

Especially with the spread of machine learning and deep learning, researchers analyzing ECG signals might be tempted to use more complex methods just because they are more trendy or *in fashion*. Instead, we recommend to start any analysis by utilizing simple algorithms, chosen among the traditional ones; they are after as good as complex ones (Andreotti et al., 2017). Do not start with complex stuff.

If the simple, traditional algorithms applied to the ECG data are sufficient to solve your scientific problem, just stick with them. Only if no simple computational technique can do it, move to more complex methods, such as machine learning and deep learning. Using simple methods will give you the chance to keep everything under control and to see results generated in an interpretable way, which is pivotal especially for the clinicians that will use your results (Rudin, 2019).

Along these lines, we suggest to start your preliminary analysis by using open source toolboxes or software libraries created specifically for dealing with ECG analysis (Van Gent et al., 2019; Martínez et al., 2017). In fact, other researchers have developed those libraries which were tested thoroughly before publication. Examples of such libraries are the already-mentioned HeartPy (Van Gent et al., 2019, 2018) package in Python and the RHRV (Martínez et al., 2017; The Comprehensive R Archive Network, 2023) package in R, both

open source and publicly available for free. Other software repositories on electrocardiogram signal processing are available on [GitHub \(2023\)](#).

TIP 7: CHOOSE THE MOST SUITABLE METRICS FOR YOUR SCIENTIFIC PROBLEM AND NEVER RELY ON A SINGLE ONE

When you have decided on the methodology you are going to follow, how will you know if it is actually good enough for your problem? Take, for instance, the task of heartbeat classification, a problem studied for decades but still needs to be solved. The problem of heartbeat classification is affected by the small number of training samples, data imbalance both in the number of samples for each class and the number of distinct features for each class, the large subject variability of the ECG signals, as well as the morphological similarity between different beats ([Wang, Zhou & Yang, 2022](#); [Jiang et al., 2019](#); [Villa et al., 2019](#)). Thus, employing a performance metric like accuracy would lead to misleading results. For a problem like this, it would be wise to use various metrics that consider the imbalance between the classes, the imbalance between the different features, and the variability between the subjects. That way, assessing and comparing different algorithms or methods would be more straightforward. Therefore, we recommend providing multiple performance metrics, like the ones mentioned later in this section, and good arguments on why each performance metric was chosen and what it represents ([Chicco & Shiradkar, 2023](#)).

Accuracy is the simplest and, thus, most typical way to evaluate a method performing ECG peak detection, beat or signal classification, clustering, and regression ([Jambukia, Dabhi & Prajapati, 2015](#); [Joloudari et al., 2022](#); [Nezamabadi et al., 2023](#)). However, accuracy is a measure that only shows the amount of correctly predicted classes and can be proven highly deceptive for datasets with unbalanced classes. For example, when you have thousands of hours of ECG recordings, and you create a model based on these that predicts an event that only occurs in a small portion of time relative to the whole duration of the recordings, you will have a high accuracy performance since the amount of the actual events is insignificant compared to the whole amount of your available data. Thus, it will not affect the overall accuracy even if it is not predicted.

For that reason, it is highly advised always to avoid accuracy and to include more metrics that account for imbalanced datasets while also considering the importance of each metric for each scientific problem. When dealing with a classification problem, always include sensitivity, specificity, precision, and negative predictive value in your reported results. Moreover, remember to include a metric that considers the whole confusion matrix for evaluating your model, the Matthews correlation coefficient (MCC) ([Jurman, Riccadonna & Furlanello, 2012](#); [Chicco, 2017](#); [Chicco & Jurman, 2020b](#); [Chicco & Jurman, 2023a, 2023b, 2022](#); [Chicco, Starovoitov & Jurman, 2021](#); [Chicco, Tötsch & Jurman, 2021](#)) metric, which will clearly indicate whether your model performs well in all the corresponding classes. An alternative to MCC is the Kappa Cohen coefficient, which is quite famous among applications with imbalanced datasets like sleep stage detection;

however, it should be used more cautiously (Delgado & Tibau, 2019; Chicco, Warrens & Jurman, 2021).

If you still want to report accuracy for interpretation reasons, consider a weighted accuracy measure (Li, Rajagopalan & Clifford, 2014), which has also been proposed as a way to describe the performance of models dealing with imbalanced datasets, though always combined with those measures as mentioned above. The j index (Soria & Martinez, 2007), which adds the sensitivity and specificity of the most important classes, has also been suggested for beat classification problems that deal with imbalanced, like arrhythmia datasets (Villa et al., 2019). For applications like seizure detection, a pivotal evaluation measure is the false detection/alarm rate per day (Jeppesen et al., 2023; Vandecasteele et al., 2017; De Cooman et al., 2017; Bhagubai et al., 2023). Further, if you aim to create a robust electrocardiogram (ECG) peak detection algorithm, you should also consider checking the error rate. Concerning regression analysis, as already mentioned in other tips (Chicco & Shiradkar, 2023), the way to proceed is by always including the coefficient of determination (R-squared or R^2), symmetric mean absolute percentage error (SMAPE), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and build the rankings of the methods on R-squared (Chicco, Warrens & Jurman, 2021; Chicco & Shiradkar, 2023).

Concerning clustering scientific problems, you should always report measures like the Davies-Bouldin index (Davies & Bouldin, 1979), the Calinski-Harabasz index (Łukasik et al., 2016), the Dunn index (Dunn, 1974), Gap statistic (Tibshirani, Walther & Hastie, 2001), and the Silhouette score (Kaufman & Rousseeuw, 2009; Chicco & Shiradkar, 2023) in combination with other metrics relevant to your research objective. Other similarity metrics which can be considered are the Jaccard coefficient and normalized mutual information (Nezamabadi et al., 2023). Nonetheless, if you already have ground truth labels of what you will expect, you can also use this information. For statistical analyses, always include the adjusted p -value (Jafari & Ansari-Pour, 2019; Chén et al., 2023), by using a 0.005 significance threshold (Benjamin et al., 2018). The take-home message here is that you should always check the recent literature to find out which metrics are used with respect to your scientific problem, use multiple metrics and not only one, and be careful that using the wrong metrics can fool you and fool the readership.

TIP 8: PREPARE YOUR RESULTS IN A FORMAT THAT CAN BE CLEARLY UNDERSTOOD, INTERPRETED, AND USED BY THE CLINICIANS

Collaboration with medical experts in the field is one of the cornerstones of computational medical research. In electrocardiogram (ECG) research, this is not different. You have already asked their opinion concerning the medical context and defined your scientific goal with them. It is now time to present your results to either take their feedback or prepare a manuscript together. Yet, medical experts may not have the computational background to fully follow the steps your methods are based on. They also might not be aware of certain metrics you used without further explanation from your side. Therefore, you need to be sure that your results are in a format that is easily comprehended.

A picture means a thousand words. Or, for the scientific case, a figure means a thousand words. Try to create figures showing all the appropriate information and the message you want to convey. Is the message that one method works better than the other? Smartly use your predefined evaluation metrics and create figures that clearly show your novelty. If, for example, you have worked on an electrocardiogram (ECG) peak detection algorithm that can clearly distinguish the PQRST wave of ECG even in very noisy data, then visualize a noisy ECG and what your algorithm captures. If you study features that are difficult to understand by scientists without a technical background, find a way to create a self-explanatory visualization or an analogy that can easily explain what you are taking into account and what it represents. Create figures that could be directly used in a publication, both for your convenience and mostly to be sure that what you present is readable and comprehensible (Rougier, Droettboom & Bourne, 2014; Chicco & Jurman, 2023c).

Does the clinician want to carry out some further analysis? Ask them in what format they would want to see those features you extracted so they can do their own analysis afterward. For example, provide them a data frame in comma-separated values (CSV) format that includes all the information you have extracted, making abstraction of the method itself. In those tables, we suggest you include the information on all the variables relevant to the problem, like age, sex, and different health conditions related to the scientific problem as different categorical values, and finally, all the other measures you calculated during your analysis while ensuring the patients' confidentiality. Thereon, they can use tools or software they prefer to perform their own analysis.

Finally, make sure to write a report explaining your methods and all the figures you created. Preferably, write this report as you would for a scientific paper, namely, explain thoroughly the methodology you used and the results you achieved, and add clear captions to your figures and tables (Ehrhart & Evelo, 2021). Subsequently, this report could be directly used as part of a potential manuscript draft, helping you save time and reduce the need for revisions when that moment arrives (Chicco & Jurman, 2023c).

TIP 9: LOOK FOR A VALIDATION COHORT DATASET TO CONFIRM THE RESULTS YOU OBTAINED ON THE PRIMARY COHORT DATASET

After performing computational analysis on your main electrocardiogram dataset, which we can call *primary cohort*, you might have found interesting results that can tell you something relevant about the cardiac cycles of the patients. Even if these results might seem promising, if they were found on a single dataset, they can be considered specific to that cohort of patients, and they might lack generalizability.

How can you assess if your scientific results stand not only for the patients of the dataset analyzed but rather for all the patients with the same conditions? An effective way to tackle this issue is to look for an alternative dataset of data recorded from other patients but having the same features of the primary cohort dataset. This dataset is usually called *validation cohort*.

Repeating your computational analysis on a validation cohort dataset would allow you to confirm or reject the discoveries made on the primary dataset. See for example this study on electronic health records of patients diagnosed with sepsis ([Chicco & Jurman, 2020a](#)).

Of course, often it is difficult to find a validation dataset compatible with the primary one, but it is always worth trying. Therefore, this is our recommendation: look for an ECG dataset of the same type of your primary dataset from the clinicians you know or online on open data repositories such as [Kaggle \(2023\)](#), [Figshare \(2023\)](#), [Zenodo \(2023\)](#), [PhysioNet \(2023\)](#), the University of California Irvine Machine Learning Repository ([University of California Irvine, 2023](#)), re3data ([Registry of Research Data Repositories, 2023](#)), or Google Dataset Search ([Google, 2023](#)). Several electrocardiogram signal datasets (such as PTB-XL ([Wagner et al., 2020](#); [Strodthoff et al., 2023](#)), MedalCare-XL ([Gillette et al., 2023](#)), and others ([Khan, Hussain & Malik, 2021](#); [Silva et al., 2014](#); [Zhang et al., 2015](#); [Liu et al., 2022](#); [Zheng et al., 2020](#))) have been released and published with peer-reviewed articles within journals like *Scientific Data* in the last years.

If you find one, repeat your computational analysis on it and include its results in your study.

TIP 10: RELEASE YOUR DATA AND SOFTWARE CODE PUBLICLY ONLINE AND SUBMIT YOUR ARTICLE TO AN OPEN ACCESS JOURNAL, IF POSSIBLE

In the previous tip (Tip 9), we recommended to look for an electrocardiogram (ECG) validation cohort dataset online, where to re-run your computational analysis and perhaps verify your results obtained on the primary cohort dataset. These online data repositories are useful for this scope, since they can have multiple datasets to offer for secondary analyses. However, these online data repositories should not only be exploited, but also helped to grow: since your ECG dataset could serve as a validation cohort for someone else's study, we recommend you to openly share it online.

Of course, this sharing is possible only when you have the necessary authorization to openly share the ECG data of the patients, obtained from the ethical committee of your hospital or of your institution. If you have this authorization, we advise you to proceed and share both your raw ECG data and your ECG preprocessed data on public repositories (such as the already-mentioned [Kaggle \(2023\)](#), [Figshare \(2023\)](#), [Zenodo \(2023\)](#), [PhysioNet \(2023\)](#), or University of California Irvine Machine Learning Repository ([University of California Irvine, 2023](#))).

Sharing your data online would make it possible to anyone around the world to re-analyze your dataset and to discover something new regarding cardiology, allowing the progress of scientific research and facilitating new scientific discoveries.

Regarding reproducibility, we recommend to make your scripts public online on GitHub or GitLab. This way, anyone will be able to re-use your computational methods, understand how you implemented them and use them, make changes to discover new scientific outcomes and verifying if you made any mistake in your analyses ([Barnes, 2010](#)). If your software code is publicly available online, your scientific discoveries can be

considered more robust and transparent, since you allowed the reproducibility of your study ([Markowetz, 2015](#)).

Moreover, if you have a say on which scientific journal you and your collaborators will choose for the paper submission, we recommend you to select an open access one. If published in an open access venue, your study's article will be available to anyone in the world, and not only to members of academic institutions or to members of high-tech companies which pay restricted-access journals to read their articles. Open access articles, in fact, can be read for free even by people in the least developed countries, currently facing wars (in Somalia and South Sudan, for example), and by high schools students who would like to delve into a specific theme. A list of health informatics open access journals where to submit an article on electrocardiogram data can be found on ScimagoJR ([Scimago Journal Ranking, 2023](#)).

CONCLUSIONS

Electrocardiogram (ECG) is a powerful and useful tool to understand the health of a heart: anomalies caught by an ECG can mean the difference between life and death for a patient. The precious information provided by an electrocardiogram can lead to important decisions on the patient's health, such as the need to have a surgery or to start a new therapy.

Computational analyses of ECG signal, usually called *ECG signal processing*, can provide additional insights and pieces of information about the patient's health to the medical doctors. Results of ECG signal processing, in fact, can highlight trends and outcomes that would otherwise be impossible to noticed by the physicians. We prepared our tips by considering only ECG data acquired digitally, but there is also an older subfield of ECG data analysis which involves the digitization of ECG recorded on paper ([Lence et al., 2023](#)), through devices such as Biopac modules ([Maheshkumar et al., 2016](#)). We have no experience with these methods and therefore we provided no recommendations on this specific ECG theme.

Although ECG computational analysis has become fundamental in cardiac research, it is sometimes carried out in the wrong way, with researchers making mistakes that lead to misleading results. Here, we presented a list of guidelines for the electrocardiogram signal processing to avoid common mistakes and wrong practices we noticed in several studies in the past.

Our list of quick tips, although partial, can serve as a starting point for good practices for researchers performing this kind of computational analysis. Although we designed these recommendations for apprentices and beginners, we believe they should be followed by experts, too. Complying with these guidelines can lead to better, more robust, and more reliable scientific results, that in turn can mean a better understanding of how patients' hearts work, and therefore can lead to better therapies and better clinical outcomes.

LIST OF ABBREVIATIONS

CC	Creative Commons
CSV	comma-separated values

ECG	electrocardiogram
FIR	finite impulse response
HRV	Heart Rate Variability
Hz	hertz
IOR	in-band to out-of-band spectral power ratio
MAE	mean absolute error
MCC	Matthews correlation coefficient
MAPE	mean absolute percentage error
MSE	mean square error
PYPL	PopularitY of Programming Language
R-squared	R^2 coefficient of determination
RMSE	root mean square error
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
SMAPE	symmetric mean absolute percentage error
SQIs	Signal Quality Indices.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The work of Davide Chicco was funded by the European Union—Next Generation EU programme, in the context of The National Recovery and Resilience Plan, Investment Partenariato Esteso PE8 “Conseguenze e sfide dell’invecchiamento”, Project Age-It (Ageing Well in an Ageing Society), and was supported by Ministero dell’Università e della Ricerca of Italy under the “Dipartimenti di Eccellenza 2023–2027” ReGAIInS grant assigned to Dipartimento di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Ministero dell’Università e della Ricerca of Italy.

Dipartimento di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca.

Competing Interests

Davide Chicco is an Academic Editor for PeerJ Computer Science.

Author Contributions

- Davide Chicco conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables,

authored or reviewed drafts of the article, designed and conceived the quick tips, and approved the final draft.

- Angeliki-Ilektra Karaiskou performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, designed and conceived the quick tips, and approved the final draft.
- Maarten De Vos conceived and designed the experiments, performed the experiments, performed the computation work, authored or reviewed drafts of the article, designed and conceived the quick tips, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This is a literature review.

REFERENCES

- Afkhami RG, Azarnia G, Tinati MA. 2016.** Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals. *Pattern Recognition Letters* **70**(3):45–51
DOI [10.1016/j.patrec.2015.11.018](https://doi.org/10.1016/j.patrec.2015.11.018).
- Andreotti F, Carr O, Pimentel MAF, Mahdi A, De Vos M. 2017.** Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG. In: *Proceedings of CinC 2017—the 2017 Computing in Cardiology Conference*. Piscataway: IEEE, 1–4.
- Ardeti VA, Kolluru VR, Varghese GT, Patjoshi RK. 2023.** An overview on state-of-the-art electrocardiogram signal processing methods: traditional to AI-based approaches. *Expert Systems with Applications* **217**(7):119561 DOI [10.1016/j.eswa.2023.119561](https://doi.org/10.1016/j.eswa.2023.119561).
- Barnes N. 2010.** Publish your computer code: it is good enough. *Nature* **467**(7317):753
DOI [10.1038/467753a](https://doi.org/10.1038/467753a).
- Behar JA, Levy J, Zvuloni E, Gendelman S, Rosenberg A, Biton S, Derman R, Sobel JA, Alexandrovich A, Charlton P, Goda MA. 2023.** PhysioZoo: the open digital physiological biomarkers resource. In: *Proceedings of CinC 2023—Computing in Cardiology*. Vol. 50. Piscataway: IEEE, 1–4.
- Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field AP, Forster M, George EI, Gonzalez R, Goodman S, Green E, Green DP, Greenwald AG, Hadfield JD, Hedges LV, Held L, Ho TH, Hoijtink H, Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Jones JH, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore DA, Morgan SL, Munafó M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J, Johnson VE. 2018.** Redefine statistical significance. *Nature Human Behaviour* **2**(1):6–10
DOI [10.1038/s41562-017-0189-z](https://doi.org/10.1038/s41562-017-0189-z).
- Berkaya SK, Uysal AK, Gunal ES, Ergin S, Gunal S, Gulmezoglu MB. 2018.** A survey on ECG analysis. *Biomedical Signal Processing and Control* **43**:216–235 DOI [10.1016/j.bspc.2018.03.003](https://doi.org/10.1016/j.bspc.2018.03.003).
- Bhagubai M, Vandecasteele K, Swinnen L, Macea J, Chatzichristos C, De Vos M, van Paesschen W. 2023.** The power of ECG in semi-automated seizure detection in addition to two-channel behind-the-ear EEG. *Bioengineering* **10**(4):491 DOI [10.3390/bioengineering10040491](https://doi.org/10.3390/bioengineering10040491).

- Blanco-Velasco M, Weng B, Barner KE. 2008.** ECG signal denoising and baseline wander correction based on the empirical mode decomposition. *Computers in Biology and Medicine* **38**(1):1–13 DOI [10.1016/j.compbiomed.2007.06.003](https://doi.org/10.1016/j.compbiomed.2007.06.003).
- Carr O, de Vos M, Saunders KEA. 2018.** Heart rate variability in bipolar disorder and borderline personality disorder: a clinical review. *BMJ Mental Health* **21**(1):23–30 DOI [10.1136/eb-2017-102760](https://doi.org/10.1136/eb-2017-102760).
- Carr O, Saunders KEA, Tsanas A, Bilderbeck AC, Palmius N, Geddes JR, Foster R, Goodwin GM, De Vos M. 2018.** Variability in phase and amplitude of diurnal rhythms is related to variation of mood in bipolar and borderline personality disorder. *Scientific Reports* **8**(1):1649 DOI [10.1038/s41598-018-19888-9](https://doi.org/10.1038/s41598-018-19888-9).
- Castroflorio T, Mesin L, Tartaglia GM, Sforza C, Farina D. 2013.** Use of electromyographic and electrocardiographic signals to detect sleep bruxism episodes in a natural environment. *IEEE Journal of Biomedical and Health Informatics* **17**(6):994–1001 DOI [10.1109/JBHI.2013.2274532](https://doi.org/10.1109/JBHI.2013.2274532).
- CFCF. 2023.** Wikimedia Commons—file: 2022 Electrocardiogram.jpg. Available at https://commons.wikimedia.org/wiki/File:2022_Electrocardiogram.jpg.
- Chén OY, Bodelet JS, Saraiva RG, Phan H, Di J, Nagels G, Schwantje T, Cao H, Gou J, Reinen JM, Xiong B, Zhi B, Wang X, de Vos M. 2023.** The roles, challenges, and merits of the p value. *Patterns* **4**(12):100878 DOI [10.1016/j.patter.2023.100878](https://doi.org/10.1016/j.patter.2023.100878).
- Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C. 2011.** Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLOS ONE* **6**(2):e17238 DOI [10.1371/journal.pone.0017238](https://doi.org/10.1371/journal.pone.0017238).
- Chicco D. 2017.** Ten quick tips for machine learning in computational biology. *BioData Mining* **10**(1):1–17 DOI [10.1186/s13040-017-0155-3](https://doi.org/10.1186/s13040-017-0155-3).
- Chicco D, Jurman G. 2020a.** Survival prediction of patients with sepsis from age, sex, and septic episode number alone. *Scientific Reports* **10**:17156 DOI [10.1038/s41598-020-73558-3](https://doi.org/10.1038/s41598-020-73558-3).
- Chicco D, Jurman G. 2020b.** The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**:6 DOI [10.1186/s12864-019-6413-7](https://doi.org/10.1186/s12864-019-6413-7).
- Chicco D, Jurman G. 2022.** An invitation to greater use of Matthews correlation coefficient in robotics and artificial intelligence. *Frontiers in Robotics and AI* **9**:876814 DOI [10.3389/frobt.2022.876814](https://doi.org/10.3389/frobt.2022.876814).
- Chicco D, Jurman G. 2023a.** The Matthews correlation coefficient (MCC) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining* **16**(1):1–23 DOI [10.1186/s13040-023-00322-4](https://doi.org/10.1186/s13040-023-00322-4).
- Chicco D, Jurman G. 2023b.** A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and fowlkes–mallows index. *Journal of Biomedical Informatics* **144**(1):104426 DOI [10.1016/j.jbi.2023.104426](https://doi.org/10.1016/j.jbi.2023.104426).
- Chicco D, Jurman G. 2023c.** Ten simple rules for providing bioinformatics support within a hospital. *BioData Mining* **16**(1):1–12 DOI [10.1186/s13040-023-00326-0](https://doi.org/10.1186/s13040-023-00326-0).
- Chicco D, Shiradkar R. 2023.** Ten quick tips for computational analysis of medical images. *PLOS Computational Biology* **19**(1):e1010778 DOI [10.1371/journal.pcbi.1010778](https://doi.org/10.1371/journal.pcbi.1010778).
- Chicco D, Starovoitov V, Jurman G. 2021.** The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment. *IEEE Access* **9**:47112–47124 DOI [10.1109/ACCESS.2021.3068614](https://doi.org/10.1109/ACCESS.2021.3068614).
- Chicco D, Tötsch N, Jurman G. 2021.** The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining* **14**(1):1–22 DOI [10.1186/s13040-021-00244-z](https://doi.org/10.1186/s13040-021-00244-z).

- Chicco D, Warrens MJ, Jurman G. 2021.** The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access* 9:78368–78381 DOI 10.1109/ACCESS.2021.3084050.
- Chicco D, Warrens MJ, Jurman G. 2021.** The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science* 7(3):e623 DOI 10.7717/peerj-cs.623.
- Chui KT, Tsang KF, Chi HR, Ling BWK, Wu CK. 2016.** An accurate ECG-based transportation safety drowsiness detection scheme. *IEEE Transactions on Industrial Informatics* 12(4):1438–1452 DOI 10.1109/TII.2016.2573259.
- Clifford G. 2002.** Signal processing methods for heart rate variability. Doctoral thesis, Oxford University, Oxford, England, United Kingdom.
- Clifford GD. 2006.** ECG statistics, noise, artifacts, and missing data. *Advanced Methods and Tools for ECG Data Analysis* 6(1):18.
- Clifford GD, Azuaje F, McSharry P. 2006.** *Advanced methods and tools for ECG data analysis*. Vol. 10. MA: Artech House Boston.
- Cygankiewicz I, Zareba W. 2013.** Heart rate variability. *Handbook of Clinical Neurology* 117:379–393 DOI 10.1016/B978-0-444-53491-0.00031-6.
- Davies DL, Bouldin DW. 1979.** A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1*(2):224–227 DOI 10.1109/TPAMI.1979.4766909.
- De Cooman T, Varon C, Hunyadi B, van Paesschen W, Lagae L, van Huffel S. 2017.** Online automated seizure detection in temporal lobe epilepsy patients using single-lead ECG. *International Journal of Neural Systems* 27(07):1750022 DOI 10.1142/S0129065717500228.
- De Jong J. 2023.** Welcome to ECGpedia, part of CardioNetwork.org. Available at https://en.ecgpedia.org/wiki/Main_Page.
- Delgado R, Tibau XA. 2019.** Why Cohen's Kappa should be avoided as performance measure in classification. *PLOS ONE* 14(9):1–26 DOI 10.1371/journal.pone.0222916.
- Drezner JA. 2012.** Standardised criteria for ECG interpretation in athletes: a practical tool. *British Journal of Sports Medicine* 46(Suppl 1):i6–i8 DOI 10.1136/bjsports-2012-091703.
- Dunn JC. 1974.** Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics* 4(1):95–104 DOI 10.1080/01969727408546059.
- Ehrhart F, Evelo CT. 2021.** Ten simple rules to make your publication look better. *PLOS Computational Biology* 17(5):e1008938 DOI 10.1371/journal.pcbi.1008938.
- Ellis RJ, Zhu B, Koenig J, Thayer JF, Wang Y. 2015.** A careful look at ECG sampling frequency and r-peak interpolation on short-term measures of heart rate variability. *Physiological Measurement* 36(9):1827 DOI 10.1088/0967-3334/36/9/1827.
- Ernst G. 2014.** Heart rate variability. Technical report. London: Springer.
- Figshare. 2023.** Store, share, discover research. Available at <https://www.figshare.com>.
- Fujita N, Sato A, Kawarasaki M. 2015.** Performance study of wavelet-based ECG analysis for ST-segment detection. In: *Proceedings of TSP, 2015—the 38th International Conference on Telecommunications and Signal Processing*. Piscataway: IEEE, 430–434.
- Garcia TB. 2015.** *12-lead ECG: the art of interpretation*. Burlington, Massachusetts, USA: Jones & Bartlett Learning.
- Garcia-Gonzalez MA, Fernandez-Chimeno M, Ramos-Castro J. 2004.** Bias and uncertainty in heart rate variability spectral indices due to the finite ECG sampling frequency. *Physiological Measurement* 25(2):489 DOI 10.1088/0967-3334/25/2/008.

- Gillette K, Gsell MAF, Nagel C, Bender J, Winkler B, Williams SE, Bär M, Schäffter T, Dössel O, Plank G, Loewe A. 2023. MedalCare-XL: 16,900 healthy and pathological synthetic 12 lead ECGs from electrophysiological simulations. *Scientific Data* 10:531 DOI 10.1038/s41597-023-02416-4.
- GitHub. 2023. ECG repositories. Available at <https://github.com/topics/ecg>.
- Goda MÁ, Charlton PH, Behar JA. 2024. pyPPG: a Python toolbox for comprehensive photoplethysmography signal analysis. *Physiological Measurement* 45(4):045001 DOI 10.1088/1361-6579/ad33a2.
- Google. 2023. Dataset search. Available at <https://datasetsearch.research.google.com/>.
- Gregg RE, Zhou SH, Lindauer JM, Helfenbein ED, Giuliano KK. 2008. What is inside the electrocardiograph? *Journal of Electrocardiology* 41(1):8–14 DOI 10.1016/j.jelectrocard.2007.08.059.
- Gupta V, Mittal M, Mittal V. 2022. A novel FrWT based arrhythmia detection in ECG signal using YWARA and PCA. *Wireless Personal Communications* 124:1–18 DOI 10.1007/s11277-021-09403-1.
- Gupta V, Mittal M, Mittal V, Diwania S, Saxena NK. 2023a. ECG signal analysis based on the spectrogram and spider monkey optimisation technique. *Journal of the Institution of Engineers (India): Series B* 104(1):153–164 DOI 10.1007/s40031-022-00831-6.
- Gupta V, Mittal M, Mittal V, Gupta A. 2023b. Adaptive autoregressive modeling based ECG signal analysis for health monitoring. In: *Optimization Methods for Engineering Problems*. USA: Apple Academic Press, 1–15.
- Gupta V, Sharma AK, Pandey PK, Jaiswal RK, Gupta A. 2023c. Pre-processing based ECG signal analysis using emerging tools. *IETE Journal of Research* 2023:1–12 DOI 10.1080/03772063.2023.2202162.
- Hamilton P. 2002. Open source ECG analysis. In: *Computers in Cardiology*. Piscataway: IEEE, 101–104.
- He T, Clifford G, Tarassenko L. 2006. Application of independent component analysis in removing artefacts from the electrocardiogram. *Neural Computing & Applications* 15(2):105–116 DOI 10.1007/s00521-005-0013-y.
- Jafari M, Ansari-Pour N. 2019. Why, when and how to adjust your p values? *Cell Journal (Yakhteh)* 20(4):604 DOI 10.22074/cellj.2019.5992.
- Jambukia SH, Dabhi VK, Prajapati HB. 2015. Classification of ECG signals using machine learning techniques: a survey. In: *Proceedings of ICACEA, 2015—the 2nd International Conference on Advances in Computer Engineering and Applications*. Piscataway: IEEE, 714–721.
- Jane R, Laguna P, Thakor NV, Caminal P. 1992. Adaptive baseline wander removal in the ECG: comparative analysis with cubic spline technique. *Computers in Cardiology* 52:143 DOI 10.1109/CIC.1992.269426.
- Jeppesen J, Christensen J, Johansen P, Beniczky S. 2023. Personalized seizure detection using logistic regression machine learning based on wearable ECG-monitoring device. *Seizure* 107:155–161 DOI 10.1016/j.seizure.2023.04.012.
- Jiang J, Zhang H, Pi D, Dai C. 2019. A novel multi-module neural network system for imbalanced heartbeats classification. *Expert Systems with Applications: X* 1:100003 DOI 10.1016/j.eswax.2019.100003.
- Joloudari JH, Mojriani S, Nodehi I, Mashmool A, Zadegan ZK, Shirkharkolaie SK, Alizadehsani R, Tamadon T, Khosravi S, Kohnehshari MA, Hassannatajjeloudari E, Sharifrazi D, Mosavi A, Loh HW, Tan RS, Acharya UR. 2022. Application of artificial intelligence techniques for

- automated detection of myocardial infarction: a review. *Physiological Measurement* **43(8)**:08TR01 DOI 10.1088/1361-6579/ac7fd9.
- Jurman G, Riccadonna S, Furlanello C. 2012.** A comparison of MCC and CEN error measures in multi-class prediction. *PLOS ONE* **7(8)**:e41882 DOI 10.1371/journal.pone.0041882.
- Kaggle. 2022.** State of data science and machine learning 2022. Available at <https://www.kaggle.com/kaggle-survey-2022>.
- Kaggle. 2023.** Kaggle datasets—Explore, analyze, and share quality data. Available at <https://www.kaggle.com/datasets>.
- Kaufman L, Rousseeuw PJ. 2009.** *Finding groups in data: an introduction to cluster analysis*. Hoboken: John Wiley & Sons.
- Kaur M, Singh B, Seema. 2011.** Comparison of different approaches for removal of baseline wander from ECG signal. In: *Proceedings of the ICWET '11—the 2011 International Conference & Workshop on Emerging Trends in Technology*. 1290–1294.
- Khan AH, Hussain M, Malik MK. 2021.** ECG images dataset of cardiac and COVID-19 patients. *Data in Brief* **34**:106762 DOI 10.1016/j.dib.2021.106762.
- Lastre-Dominguez C, Shmaliy YS, Ibarra-Manzano O, Vazquez-Olguin M. 2019.** Denoising and features extraction of ecg signals in state space using unbiased fir smoothing. *IEEE Access* **7**:152166–152178 DOI 10.1109/ACCESS.2019.2948067.
- Lence A, Extramiana F, Fall A, Salem JE, Zucker JD, Prifti E. 2023.** Automatic digitization of paper electrocardiograms—a systematic review. *Journal of Electrocardiology* **80**:125–132 DOI 10.1016/j.jelectrocard.2023.05.009.
- Leski JM, Henzel N. 2005.** ECG baseline wander and powerline interference reduction using nonlinear filter bank. *Signal Processing* **85(4)**:781–793 DOI 10.1016/j.sigpro.2004.12.001.
- Li Q, Rajagopalan C, Clifford GD. 2013.** Ventricular fibrillation and tachycardia classification using a machine learning approach. *IEEE Transactions on Biomedical Engineering* **61(6)**:1607–1613 DOI 10.1109/TBME.2013.2275000.
- Li Q, Rajagopalan C, Clifford CD. 2014.** Ventricular fibrillation and tachycardia classification using a machine learning approach. *IEEE Transactions on Biomedical Engineering* **61(6)**:1607–1613 DOI 10.1109/TBME.2013.2275000.
- Liu H, Chen D, Chen D, Zhang X, Li H, Bian L, Shu M, Wang Y. 2022.** A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Scientific Data* **9(1)**:272 DOI 10.1038/s41597-022-01403-5.
- Łukasik S, Kowalski PA, Charytanowicz M, Kulczycki P. 2016.** Clustering using flower pollination algorithm and Calinski-Harabasz index. In: *Proceedings of IEEE CEC, 2016—the 2016 IEEE Congress on Evolutionary Computation*. IEEE, 2724–2728.
- Lu L, Zhu T, Morelli D, Creagh A, Liu Z, Yang J, Liu F, Zhang YT, Clifton DA. 2023.** Uncertainties in the analysis of heart rate variability: a systematic review. *IEEE Reviews in Biomedical Engineering* **17**:180–196 DOI 10.1109/RBME.2023.3271595.
- Mahdiani S, Jeyhani V, Peltokangas M, Vehkaoja A. 2015.** Is 50 hz high enough ECG sampling frequency for accurate HRV analysis? In: *Proceedings of IEEE EMBC, 2015—the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Piscataway: IEEE, 5948–5951.
- Maheshkumar K, Dilara K, Maruthy KN, Sundareswaren L. 2016.** Validation of PC-based sound card with Biopac for digitalization of ECG recording in short-term HRV analysis. *North American Journal of Medical Sciences* **8(7)**:307 DOI 10.4103/1947-2714.187150.

- Makin TR, de Xivry JJO. 2019.** Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife* 8:e48175 DOI 10.7554/eLife.48175.
- Markowitz F. 2015.** Five selfish reasons to work reproducibly. *Genome Biology* 16(1):1–4 DOI 10.1186/s13059-015-0850-7.
- Martínez CAG, Quintana AO, Vila XA, Touriño MJL, Rodríguez-Liñares L, Presedo JMR, Penín AJM. 2017.** *Heart rate variability analysis with the R package RHRV*. Cham: Springer.
- Moeyersons J, Smets E, Morales J, Villa A, De Raedt W, Testelmans D, Buyse B, van Hoof C, Willems R, van Huffel S, Varon C. 2019.** Artefact detection and quality assessment of ambulatory ECG signals. *Computer Methods and Programs in Biomedicine* 182(3):105050 DOI 10.1016/j.cmpb.2019.105050.
- Nejadgholi I, Moradi MH, Abdolali F. 2011.** Using phase space reconstruction for patient independent heartbeat classification in comparison with some benchmark methods. *Computers in Biology and Medicine* 41(6):411–419 DOI 10.1016/j.compbiomed.2011.04.003.
- Nezamabadi K, Sardaripour N, Haghi B, Forouzanfar M. 2023.** Unsupervised ECG analysis: a review. *IEEE Reviews in Biomedical Engineering* 16:208–224 DOI 10.1109/RBME.2022.3154893.
- Nuthalapati S, Goduguluri VY, Rahman MZU, Prasanna CS, Divvela SY, Cheemakurty K. 2019.** Artifact elimination in cardiac signals using through circular leaky adaptive algorithms for remote patient care monitoring. *Indian Journal of Public Health Research & Development* 10(11):2350 DOI 10.5958/0976-5506.2019.03956.1.
- Oktivasari P, Hasyim M, Amy HS, Freddy H, Suprijadi. 2019.** A simple real-time system for detection of normal and myocardial ischemia in the ST segment and *t* wave ECG signal. In: *Proceedings of ICOIACT 2019—the 7th International Conference on Information and Communications Technology*. Piscataway: IEEE, 327–331.
- Pérez-Riera AR, Barbosa-Barros R, Daminello-Raimundo R, de Abreu LC. 2018.** Main artifacts in electrocardiography. *Annals of Noninvasive Electrocardiology* 23(2):e12494 DOI 10.1111/anec.12494.
- Petelczyc M, Gieraltowski JJ, Żogała-Siudem B, Siudem G. 2020.** Impact of observational error on heart rate variability analysis. *Heliyon* 6(5):e03984 DOI 10.1016/j.heliyon.2020.e03984.
- PhysioNet. 2023.** The research resource for complex physiologic signals. Available at <https://www.physionet.org/>.
- PhysioZoo. 2023.** Continuous physiological time series analysis from human and other mammals. Available at <https://physiozoo.com/>.
- pypi ecg-quality. 2023.** Library that classifies quality of ECG signal using deep learning methods. Available at <https://pypi.org/project/ecg-quality/>.
- pypi ecghelper. 2023.** Tools to load and process electrocardiogram data. Available at <https://pypi.org/project/ecghelper/>.
- pypi ecgtools. 2024.** ESM catalog generation utilities. Available at <https://pypi.org/project/ecgtools/>.
- pypi ndx-ecg. 2023.** This extension is developed to extend NWB data standards to incorporate ECG recordings. Available at <https://pypi.org/project/ndx-ecg/>.
- PYPL. 2023.** PopularitY of programming language. Available at <https://pypl.github.io/PYPL.html>.
- Registry of Research Data Repositories. 2023.** Registry of research data repositories. Available at <https://www.re3data.org/>.
- Rougier NP, Droettboom M, Bourne PE. 2014.** Ten simple rules for better figures. *PLOS Computational Biology* 10(9):e1003833 DOI 10.1371/journal.pcbi.1003833.

- Rudin C.** 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**(5):206–215
DOI [10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x).
- Salman MN, Rao PT, Rahman MZU.** 2018. Novel logarithmic reference free adaptive signal enhancers for ECG analysis of wireless cardiac care monitoring systems. *IEEE Access* **6**:46382–46395 DOI [10.1109/ACCESS.2018.2866303](https://doi.org/10.1109/ACCESS.2018.2866303).
- Sameni R, Clifford GD.** 2010. A review of fetal ECG signal processing; issues and promising directions. *The Open Pacing, Electrophysiology & Therapy Journal* **3**:4
DOI [10.2174/1876536X01003010004](https://doi.org/10.2174/1876536X01003010004).
- Scimago Journal Ranking.** 2023. Health informatics open access journals. Available at <https://www.scimagojr.com/journalrank.php?openaccess=true&type=j&category=2718>.
- Seeuws N, De Vos M, Bertrand A.** 2021. Electrocardiogram quality assessment using unsupervised deep learning. *IEEE Transactions on Biomedical Engineering* **69**(2):882–893
DOI [10.1109/TBME.2021.3108621](https://doi.org/10.1109/TBME.2021.3108621).
- Silva HPD, Lourenço A, Fred A, Raposo N, Aires-de Sousa M.** 2014. Check your biosignals here: a new dataset for off-the-person ECG biometrics. *Computer Methods and Programs in Biomedicine* **113**(2):503–514 DOI [10.1016/j.cmpb.2013.11.017](https://doi.org/10.1016/j.cmpb.2013.11.017).
- Sörnmo L.** 1993. Time-varying digital filtering of ECG baseline wander. *Medical and Biological Engineering and Computing* **31**(5):503–508 DOI [10.1007/BF02441986](https://doi.org/10.1007/BF02441986).
- Sörnmo L, Laguna P.** 2006. Electrocardiogram (ECG) signal processing. In: *Wiley encyclopedia of biomedical engineering*. Berlin, Germany: Springer, 1–16.
- Soria ML, Martinez JP.** 2007. An ECG classification model based on multilead wavelet transform features. In: *Proceedings of CinC 2007—the 2007 Computers in Cardiology Conference*. 105–108.
- Sprang M, Andrade-Navarro MA, Fontaine JF.** 2022. Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality. *BMC Bioinformatics* **23**(6):1–15 DOI [10.1186/s12859-022-04775-y](https://doi.org/10.1186/s12859-022-04775-y).
- Strothoff N, Mehari T, Nagel C, Aston PJ, Sundar A, Graff C, Kanters JK, Haverkamp W, Dössel O, Loewe A, Bär M, Schaeffter T.** 2023. PTB-XL+, a comprehensive electrocardiographic feature dataset. *Scientific Data* **10**:279 DOI [10.1038/s41597-023-02153-8](https://doi.org/10.1038/s41597-023-02153-8).
- Sulthana A, Rahman MZU, Mirza SS.** 2018. An efficient Kalman noise canceller for cardiac signal analysis in modern telecardiology systems. *IEEE Access* **6**:34616–34630
DOI [10.1109/ACCESS.2018.2848201](https://doi.org/10.1109/ACCESS.2018.2848201).
- The Comprehensive R Archive Network.** 2023. RHRV: heart rate variability analysis of ECG data. Available at <https://cran.r-project.org/web/packages/RHRV/index.html>.
- Tibshirani R, Walther G, Hastie T.** 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2):411–423 DOI [10.1111/1467-9868.00293](https://doi.org/10.1111/1467-9868.00293).
- TIOBE.** 2023. TIOBE index for November 2023. Available at <https://www.tiobe.com/tiobe-index/>.
- Tobón VDP, Falk TH, Maier M.** 2014. MS-QI: a modulation spectrum-based ECG quality index for telehealth applications. *IEEE Transactions on Biomedical Engineering* **63**(8):1613–1622
DOI [10.1109/TBME.2014.2355135](https://doi.org/10.1109/TBME.2014.2355135).
- University of California Irvine.** 2023. Machine learning repository. Available at <https://archive.ics.uci.edu/>.
- Van Gent P, Farah H, Van Nes N, Van Arem B.** 2019. Heartpy: a novel heart rate algorithm for the analysis of noisy signals. *Transportation Research Part F: Traffic Psychology and Behaviour* **66**:368–378 DOI [10.1016/j.trf.2019.09.015](https://doi.org/10.1016/j.trf.2019.09.015).

- Van Gent P, Farah H, Nes N, Van Arem B. 2018.** Heart rate analysis for human factors: development and validation of an open source toolkit for noisy naturalistic heart rate data. In: *Proceedings of the 6th HUMANIST Conference*, 173–178.
- Vandecasteele K, De Cooman T, Gu Y, Cleeren E, Claes K, van Paesschen W, van Huffel S, Hunyadi B. 2017.** Automated epileptic seizure detection based on wearable ECG and PPG in a hospital environment. *Sensors* **17(10)**:2338 DOI [10.3390/s17102338](https://doi.org/10.3390/s17102338).
- Varon C, van Huffel S, Suykens J. 2015.** Mining the ECG: algorithms and applications. Technical report, Katholieke Universiteit, Leuven, Belgium.
- Villa A, Deviaene M, Willems R, Van Huffel S, Varon C. 2019.** Are we training our heartbeat classification algorithms properly? In: *Proceedings of EMBC, 2019—the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 6363–6366.
- Wagner P, Strodthoff N, Boussejot R-D, Kreiseler D, Lunze FI, Samek W, Schaeffter T. 2020.** PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data* **7(1)**:154 DOI [10.1038/s41597-020-0495-6](https://doi.org/10.1038/s41597-020-0495-6).
- Wang Y, Zhou G, Yang C. 2022.** Interpatient heartbeat classification using modified residual attention network with two-phase training and assistant decision. *IEEE Transactions on Instrumentation and Measurement* **72**:1–15 DOI [10.1109/TIM.2022.3232646](https://doi.org/10.1109/TIM.2022.3232646).
- Zenodo. 2023.** Research, shared. Available at <https://www.zenodo.org>.
- Zhang D. 2006.** Wavelet approach for ECG baseline wander correction and noise reduction. In: *Proceedings of IEEE EMBC, 2005—the 27th IEEE Engineering in Medicine and Biology Annual Conference*. Piscataway: IEEE, 1212–1215.
- Zhang Z, Dong J, Luo X, Choi KS, Wu X. 2014.** Heartbeat classification using disease-specific feature selection. *Computers in Biology and Medicine* **46(1)**:79–89 DOI [10.1016/j.compbiomed.2013.11.019](https://doi.org/10.1016/j.compbiomed.2013.11.019).
- Zhang Q, Zahed C, Nathan V, Hall DA, Jafari R. 2015.** An ECG dataset representing real-world signal characteristics for wearable computers. In: *Proceedings of IEEE BioCAS, 2015—the 2015 IEEE Biomedical Circuits and Systems Conference*. Piscataway: IEEE, 1–4.
- Zheng J, Chu H, Struppa D, Zhang J, Yacoub SM, El-Askary H, Chang A, Ehwerhemuepha L, Abudayyeh I, Barrett A, Fu G, Yao H, Li D, Guo H, Rakovski C. 2020.** Optimal multi-stage arrhythmia classification approach. *Scientific Reports* **10**:2898 DOI [10.1038/s41598-020-59821-7](https://doi.org/10.1038/s41598-020-59821-7).
- Zheng J, Fu G, Anderson K, Chu H, Rakovski C. 2020.** A 12-lead ECG database to identify origins of idiopathic ventricular arrhythmia containing 334 patients. *Scientific Data* **7**:98 DOI [10.1038/s41597-020-0440-8](https://doi.org/10.1038/s41597-020-0440-8).
- Zong W, Moody GB, Jiang D. 2003.** A robust open-source algorithm to detect onset and duration of QRS complexes. In: *Computers in Cardiology*. Piscataway: IEEE, 737–740.