A new QRS detector stress test combining temporal jitter and accuracy (JA) reveals significant performance differences amongst popular detectors

Bernd Porr[1*], Peter W Macfarlane[2]

1 School of Engineering, Biomedical Engineering, University of Glasgow, Glasgow, United Kingdom
2 Electrocardiology Group, Institute of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom

* bernd.porr@glasgow.ac.uk

Abstract

QRS detection within an ECG signal is the basis of virtually any further processing and any error caused by this detection will propagate to further processing stages. Despite this, standard benchmarks for QRS detection algorithms use large temporal error margins and noise-free ECG databases which overestimate their performance. The use of a fixed error margin ignores the temporal jitter between detection and ground truth measurements. This led to the development of a new performance measure which combines temporal Jitter with Accuracy (JA) which in turn reflects both missed beats and extra detections. Our new performance measure JA fully encompasses all the types of errors that can occur, equally weights them and provides a percentage value which allows direct comparison between QRS detection algorithms.

Introduction

An electrocardiogram (ECG) records the electrical activity of the heart [1]. In the normal heart, each cardiac cycle (heartbeat) consists of atrial activation which produces a P wave, and ventricular activation which generates a QRS complex. This is followed by atrial repolarisation (not usually seen on an ECG) and ventricular repolarisation giving rise to a T wave. In an abnormal recording, the P wave may not be present though other signs of atrial activity may be seen. In some abnormal cardiac cycles, the source of ventricular activation may be in the ventricles giving one or more QRS complexes of a different morphology compared to the normally conducted beat for that heart. In all cases, the most identifiable wave with the fastest rate of inscription is the QRS complex [2]. Detecting the QRS complex is the first step in calculating any further measures. Most QRS detection algorithms will use a distinctive R peak as may be found in one or two lead recording with carefully placed electrodes but if more leads are available, a function which combines leads and automatically produces a prominent R wave can be used [3].

When the presence of a QRS complex has been detected, further measures such as heart rate and heart rate variability (HRV) can be calculated. Heart rate can simply be calculated by taking the difference between two QRS detection time stamps, which is mathematically the same as finding the first derivative. However, derivatives amplify noise and thus any errors in the QRS detection will result in even larger errors as further

derivatives are taken. When calculating HRV, the 2nd derivative between successive QRS complexes is used, and thus any small errors in the QRS detection will compound into much larger errors in any subsequent HRV result [4]. As such, a heart beat detector algorithm that can accurately and precisely identify the QRS complexes in an ECG, is vital in obtaining a useful and representative HRV. Errors that can occur from poor QRS detector algorithms are: missed QRS complexes (false negatives), extra QRS complexes (false positives), and temporal inaccuracies (temporal jitter). To assess the abilities of these algorithms, we have to test and benchmark their performance. To carry out these assessments, an appropriate ECG database with correctly annotated QRS complexes has to be used. The most commonly used ECG reference database is the MIT-BIH arrhythmia database (MITDB) [5, 6]. This database consists of 48 half-hour long annotated ECG recordings from 47 subjects with a sampling rate of 360 Hz with an 11 bit resolution over a 10 mV range. This database has the shortcoming that it contains very few examples of motion artefacts. The work by Benitez, Kalidas and Chan [7–9] tried to highlight noise resilience using only small sections ($3 - 10$ s) of a few select records. As these noisy sections make up such a small proportion of the database, they have very little impact on the overall results. Of the current thirty-three available ECG databases on PhysioNet, only two aim to provide examples of noisy ECG signals. However, both have flaws which limit their usefulness. The MIT-BIH Noise Stress Test Database (NSTDB) [10] features synthesised noisy signals that are not representative of a realistic noisy recording. The other database, Motion Artefact Contaminated ECG Database [11], consists of 27 recordings of just a single subject standing and is not annotated, and so is therefore unsuitable for detector evaluation.

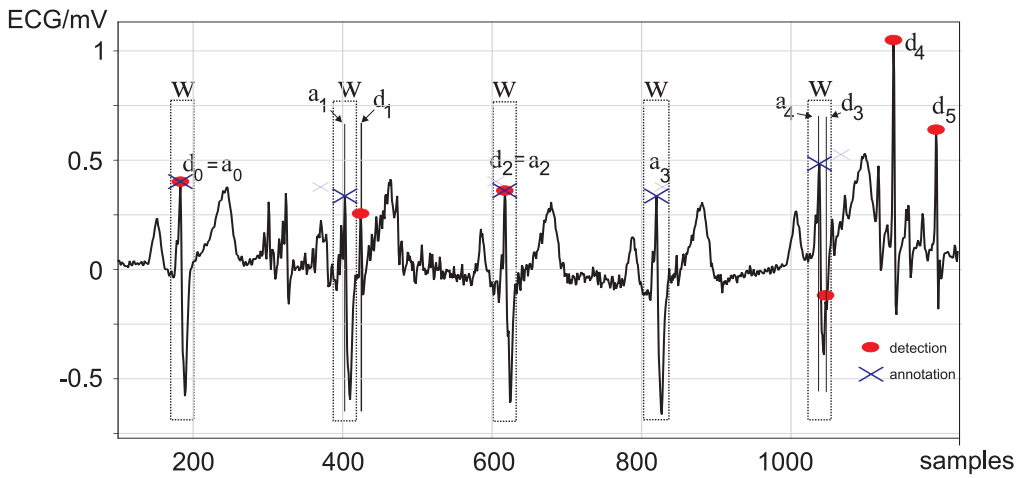


Fig 1. Examples of ECG detection errors. $d_{0...4}$ are sample numbers of detected QRS complexes. $a_{0...4}$ are sample numbers of the annotated QRS complex (ground truth). w is the temporal window used for the classical performance measures sensitivity, positive predictivity, false positive rate and overall accuracy.

Currently, the Association for the Advancement of Medical Instrumentation (AAMI) recommends evaluating detector performance on sensitivity (Se), positive predictivity (+P), false positive rate (FPR) and overall accuracy (Acc) [12]. All of these measures are based on the idea that a QRS detection is required to fall in a temporal window w which is illustrated in Fig. 1. a_l is a set of annotated sample positions where the actual detection has occurred. d_k are the sample positions where the detector has detected the QRS complex. For example in Fig. 1 the first QRS detection has been annotated at a_0 and the detection happens at exactly $d_0 = a_0$ the same sample and is classified as a true positive. The 2nd annotation a_1 has also a matching QRS complex d_1 but this falls

outside of the temporal window $|d_1 - a_1| > w$ so that it is a false positive. Of course there are also cases where there is an annotation as with a_3 but no matching QRS detection so that we have a false negative. There will also be detections such as d_4 and d_5 which are false positives. The standard approach is to count how many detections fall within the temporal window w and count these as "true positives" (TP). With the total number of detections $\text{card}(d_k)$ and annotations $\text{card}(a_l)$ we get:

$$TP = \text{card}(|a_l - a_k| < w) \tag{1}$$
$$FP = \text{card}(d_k) - TP \tag{2}$$
$$FN = \text{card}(a_l) - TP \tag{3}$$

where card() is the number of elements in a set which is called the cardinal number. For example if we have three annotations $\{a_0, a_1, a_2\}$ in a set then its cardinal number is: $\text{card}(\{a_0, a_1, a_2\}) = 3$. FN is the number of false negatives, and FP is the number of false positives where false negatives are the number of missed heartbeats, and false positives are essentially the number of extra detections.

$$SE = \frac{TP}{TP + FN} \tag{4}$$
$$+P = \frac{TP}{TP + FP} \tag{5}$$

where SE is sensitivity and $+P$ is positive predictivity.

The predefined window w of tolerance is widely set as ten times of the sampling interval as advised by Xie and Dubiel [13] in their software library for working with Physionet databases such as the MIT-BIH arrhythmia database [5, 6]:

$$w = \frac{10}{f_s} \tag{6}$$

For example, at a sampling rate of 250 samples/sec this gives a temporal window of 40 ms which means that any QRS complex within this window is classified as a true positive and outside of this window as a false negative. Many detector algorithms report high sensitivity scores due to these large windows w [12, 14], yet this is not representative of the actual detector ability across all applications, specifically HRV, as it disregards the temporal jitter. That omitting jitter is problematic can be seen in Fig. 1 when comparing the annotations a_1 and a_4. For a_1 the detection d_1 lies outside the window w and is classified as a false detection. However, for the annotation a_4 the detection d_3 lies within the window and is classified as a perfect true positive ignoring the temporal inaccuracy. Thus a resulting high sensitivity might not mean that the detector is perfect if it has a high jitter between detections falling within the window w. Gradl et al [15] proposed a possible measure to account for QRS detection inaccuracies using the MITDB. However, their measure uses the absolute deviation from each QRS complex which does not account for detectors that use different detection positions such as the R-S slope [16], and by using the MITDB, errors from inexact annotations are introduced and are not demonstrated against realistic noisy scenarios. Gradl et al [15] do not attempt to combine their QRS accuracy measure with any of the other possible errors that could occur during detection, and thus there is still no overarching benchmarking parameter proposed.

The combination of current benchmarking parameters not accounting for all error types, large tolerance windows masking temporal inaccuracies, inaccurately annotated ECG recordings, and a lack of realistic noisy recordings, has resulted in all of the most commonly used detector algorithms' performance measures being based on fallacious results under biased conditions. It is clear that the current ECG testing databases and

the current recommended performance parameters for benchmarking algorithms are insufficient.

This paper presents an ECG testing database [17] with precise annotations under realistic noise levels and a new benchmarking score (JA) that equally accounts for all error types, including temporal inaccuracies, and is application-independent to truly assess different detector algorithms.

Methods

Glasgow University Database

The Glasgow University Database (GUDB) consists of two-minute two lead ECG recordings from 25 subjects each performing five different tasks, for a total of 125 records. The tasks were chosen to be repeatable and representative of common, realistic scenarios. The tasks were as follows:

- sitting
- using a tablet to perform a maths test
- walking on a treadmill
- using a hand-bike
- jogging

Ethical approval was given by the ethics committee at the Institute of Neuroscience and Psychology, School of Psychology at the University of Glasgow, with reference 300180026. Prior to the experiment, participants were given an information sheet and were asked to give signed consent by signing two consent forms, one for the researchers and another for them to keep. Where the participant consented (24 out of the 25), a video synchronised to the data was recorded for each task. This video allows database users to see exactly how the movement was performed for each ECG recording and for any artefacts in the data to be identified. In addition, the acceleration of the torso was recorded. With the exception of the most heavily noise contaminated records, all ECGs were annotated with high precision beat locations. The participants were all over the age of 18 and had no known cardiovascular conditions. The database is available through the University of Glasgow's open access research data repository [17].

Equipment

The ECG signals were recorded using an Attys Bluetooth data acquisition board (Glasgow Neuro LTD, Glasgow). This board has a sampling rate of 250 samples/sec and a resolution of 24 bits over a range of ± 2.42 V. As this device is wireless, it increases electrical isolation and allows a moving subject to be recorded easily without the need of a cumbersome tether. The Attys features two differential analogue recording channels. Two separate Attys were used at the same time to record both Einthoven II & III and the difference between the chest leads $V_1 - V_2$. These two configurations represent a best and worst-case recording setup. This allows the impact of recording setup on signal noise to be investigated.

The best-case setup uses the first Attys mounted on an elastic electrode chest strap (Amazon, UK), connected with short cables zip tied together (Fig. 2A,B). This configuration minimises the effect of cable movement artefacts as much as possible and is worn tightly on the subject to prevent the electrodes from moving. As the chest strap is worn high around the chest, the electrodes are approximately in the same location as

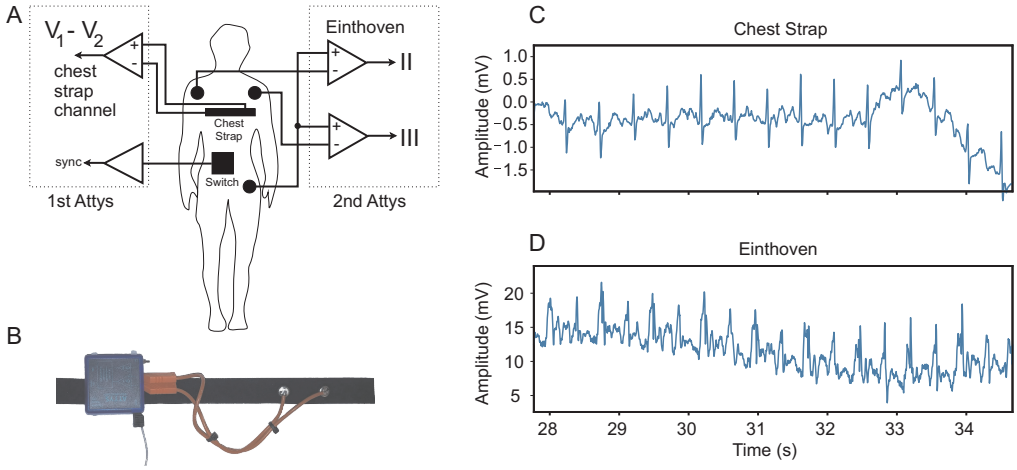


Fig 2. Experimental setup and example ECG traces. A: Wiring of the two wireless biosignal amplifiers (Attys). The first amplifier records $V_1 - V_2$ with a chest strap and the second amplifier Einthoven II & III with standard Ag/AgCl electrodes. B: Photo of the chest strap and the biosignal amplifier mounted on it. Signal comparison between C: chest strap and D: Einthoven II recording while the subject was jogging.

V_1 and V_2 in the standard six electrode chest configuration [18]. The left electrode (Pulse Medical, UK) on the chest strap is connected to the positive terminal of the differential amplifier and the right electrode is connected to the negative. The second channel of the Attys records the switch signal used to synchronise the data with the video. The switch (Fig. 2A) is worn on a belt around the waist and when switched produces an audible click and shorts channel two to ground. The circuit diagram configuration can be seen in Fig. 2A.

The worse-case configuration uses the second Attys connected to standard ECG electrodes (Pulse Medical, UK) with loose cables. The positive terminal of the first differential amplifier is connected to the left hip, and the negative terminal to the right shoulder. The Attys is put into Einthoven ECG mode where the positive terminal of the CH1 differential amplifier is connected internally to the positive terminal of CH2 amplifier input. This configuration allows two ECG signals to be recorded using only three cables. CH1 records Einthoven II between the left hip and right shoulder and CH2 records Einthoven III between the left hip and the left shoulder. The circuit diagram for this configuration can be seen in Figure 2A.

Having introduced the best and worse-case measurement situations, we show the corresponding raw signals in Fig. 2C, D when the subject is jogging. The chest strap recording remains largely noise free while the Einthoven signal has significant noise contamination.

Protocol

Before the experiment began, the participant was asked to read the information sheet and sign the consent form (see [17]), opting in or out of the video recording. The recording equipment was then connected to the participant and the experiment ran as follows:

1. 120 second ECG recording, sitting down
2. 120 second ECG recording, timed maths questions on a tablet
3. 120 second break

4. 120 second ECG recording, walking on a treadmill at 2 kph
5. 120 second break
6. 120 second ECG recording, using a hand bike
7. 120 second break
8. 120 second ECG recording, jogging on a treadmill at 7 kph
9. Electrodes and chest strap removed from participant

Post-processing

To annotate the data with heartbeat locations, a Python script was created which uses a Matplotlib [19] interactive plot. An ECG data file is loaded into the plot and run through a heartbeat detection algorithm (engzee segmenter from the BioSPPy library [20]) to provide an initial estimation of QRS locations. This estimation is then manually inspected to remove any false positives and add any missing QRS complexes. Where there was too much noise to reliably annotate the entire recording, no annotation file was made. Of the 125 recordings, 2 chest strap and 19 loose cable recordings were unable to be annotated. This mostly occurred in the jogging scenario. The annotation sample locations are saved to a .tsv file when the plot is closed. This is performed for both the chest strap ECG and the Einthoven II loose cable recording.

Detector Software Implementation

The eight algorithms chosen represent popular and well regarded QRS detectors, as well as a range of different detection techniques. The eight detector algorithms were: Pan & Tompkins [21], Elgendi et al [22], Kalidas & Tamil [8], Christov [23], Hamilton [24], the matched filter detector, the Engzee Mod detector [25] and the WQRS detector [26]. The main criterion for selection was that the algorithm could be implemented in a real time system. All of the algorithms were implemented in Python, and the code can be found at Porr et al 2023 [27].

Evaluating Detectors

Traditional method: sensitivity

All of the GUDB recordings that had corresponding annotations were evaluated: 123 for the chest strap and 106 for the loose cable setups. The widely used detection tolerance of $w = 10/f_s$ seconds was used [13]. Note that all detectors will delay the signals to some extent, so the detected time stamps will always appear to be later. Thus, in order to allow sample precision evaluation, the median delay was subtracted from the true QRS detection time stamp and calculated separately for every subject, activity, and measurement protocol, as it was different each time.

The performance of the detector algorithms was compared using the current most commonly recommended performance parameter: sensitivity (Eq. 4). The sensitivities for sitting & jogging, Einthoven & chest strap and the different detectors were then tested against a threshold of 90% with the help of the one sample t-test. Valid p values need be based on at least 20 different sensitivity values. Alpha level was set to 0.05.

While the above approach simply assesses the sensitivity performance when QRS complexes occur either within a temporal window or not they ignore the temporal jitter between the annotations and the QRS detections. In the next section we are going to present a new measure which also takes into account the jitter.

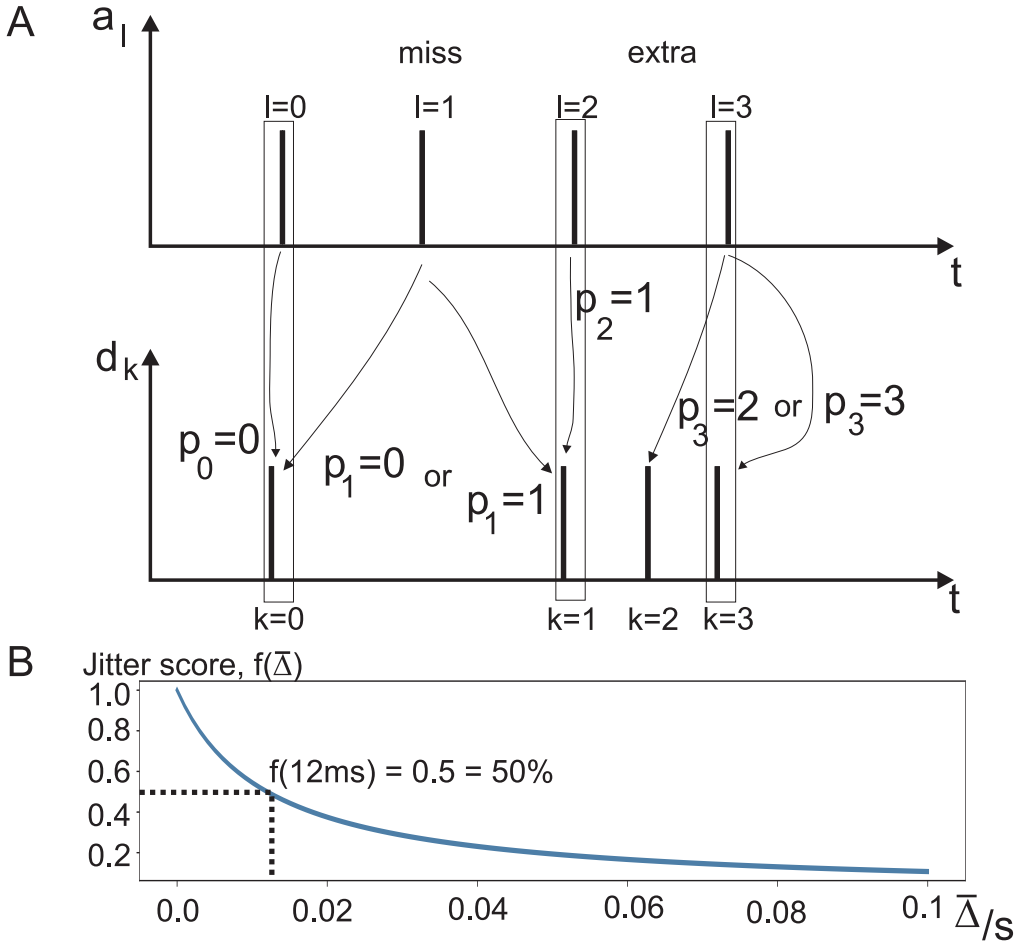


Fig 3. Graphical illustration of the JA algorithm. A: Time stamps a_l of the QRS annotations versus the time stamps of the actual QRS detections d_k. p_l links the index numbers of the annotations $l = 0, 1, 2, 3$ to the index numbers k of the detections d_k. B: Mapping function Eq. 10 which maps the average jitter $\bar{\Delta}$ to a score between zero and one which reaches 0.5 at a jitter of 12 ms.

JA

A new benchmarking approach JA is presented [28] which takes into account the temporal jitter (J) of the detection point and combines it with the standard statistical errors such as extra beats (false positives), and missed beats (false negatives) represented by the accuracy (A). In contrast to the traditional approach, no fixed temporal window is used but rather the accuracy is penalised more and more with increasing jitter.

Fig. 3 demonstrates the working principle of the JA algorithm which is now explained step by step. Its input is the sample points d_k of a QRS detector which are compared against the QRS detection annotations a_l. As a first step, we need to compensate for constant detection delays by subtracting the median delay between the detections d_k and annotations a_l from the individual detection timestamps:

$$\tilde{d_k} = d_k - \text{median}\left(\forall l, \min_k(|a_l - d_k|)\right) \tag{7}$$

where $\forall$ means "for all" with $l = 0 \ldots \text{card}(a_l) - 1$.

As a next step, we need to find now the mapping between the annotations and the detections to be able to identify true detections, missed QRS complexes and spuriously detected QRS complexes. In order to achieve this, we iterate through all annotation indices $l = 0 \ldots \mathrm{card}(a_l) - 1$ and find, for each annotation timestamp l, the index k within the detections $\tilde{d_k}$ which has the smallest temporal difference and store it in the set p_l:

$$p_l = \mathrm{argmin}_{k=0\ldots\mathrm{card}(\tilde{d_k})-1} |a_l - \tilde{d_k}| \tag{8}$$

where argmin outputs the index k where the minimum between $|a_l - \tilde{d_k}|$ occurs. Without any missing detections or spurious detections, the index of $p_l = l$ will be just a one-to-one mapping between annotations and detections. However, as illustrated for the annotation $l = 1$ in Fig. 3, there is no matching detection, and thus it is a missing beat. Depending if the annotation is closer to d_0 or d_1, the annotation will be falsely associated with one of them. This can be rectified by switching the perspective and then iterating through the p_l indices to check if there are duplicates. For example, here $p_0 = p_1 = 0$ both point to the same detection. However, clearly p_0 is closer to d_0 so that $p_1 = 0$ should be discarded. Similarly, if the annotation at $l = 1$ were closer to $k = 1$, then $p_1 = p_2 = 1$ is the duplicate position. Thus, formally we minimise the temporal difference between the detection indices while keeping the correspondence between QRS annotation and detection unique:

$$\Delta_m = \left\{ \frac{\min_k |a_l - d_{p_k}|}{fs} \text{ with unique } p_k \right\} \tag{9}$$

where Δ_m are now the jitter values of true detections, and the number of elements $\mathrm{card}(\Delta_m)$ are the number of the true positive detections.

As a next step, the jitter Δ_m needs to be evaluated and turned into a performance measure between 0 and 100%. We define the following mapping function as the jitter score:

$$f(\bar{\Delta}) = \frac{1}{1 + \frac{\bar{\Delta}}{12 \text{ ms}}} \tag{10}$$

where $\bar{\Delta}$ is the average over all individual jitter values Δ_m. This means that an average jitter of $\bar{\Delta} = 12$ ms results in a performance of 0.5 or 50 %.

Having calculated the average temporal jitter, we now need to calculate the accuracy. Here, the number of the individual jitter readings $\mathrm{card}(\Delta_m)$, being the number of the true positive (TP) detections, can be used to calculate the number of missed beats (false negatives = FN) and extra beats (false positives = FP):

$$TP = \mathrm{card}(\Delta_m) \tag{11}$$

$$FP = \mathrm{card}(d_k) - \mathrm{card}(\Delta_m) \tag{12}$$

$$FN = \mathrm{card}(a_l) - \mathrm{card}(\Delta_m) \tag{13}$$

Finally, we need to calculate the true negatives TN which can be estimated assuming that the maximum heartrate is $\mathrm{HR}_{\max} = 220$ beats per minute:

$$TN = \frac{N}{f_s} \frac{\mathrm{HR}_{\max}}{60\mathrm{bpm}} - (TP + FP + FN) \tag{14}$$

where N is the number of samples of the ECG. Having obtained the statistical parameters, we can now calculate the *accuracy* which is a performance measure for the both negative and positive detections and normalises against both positive and false detections.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

The accuracy A is a normalised measure between zero and one. As a final step, we can now combine the jitter score f (Eq. 10) and the accuracy (Eq. 15):

$$\mathrm{JA} = A \cdot f(\bar{\Delta}) \cdot 100\% \tag{16}$$

JA reaches 100% if the average jitter is zero ($\bar{\Delta} = 0$) and there are no false negative or false positive detections. Both increase of jitter and increase of false negative or false positive detections will decrease the value of JA

The JA performance values (Eq. 16) for sitting and jogging in combination with Einthoven, chest strap and the different detectors are then tested against a threshold of 90% with the help of the one sample t-test. Valid p-values need be based on at least 20 different sensitivity values. Alpha level is set to 0.05.

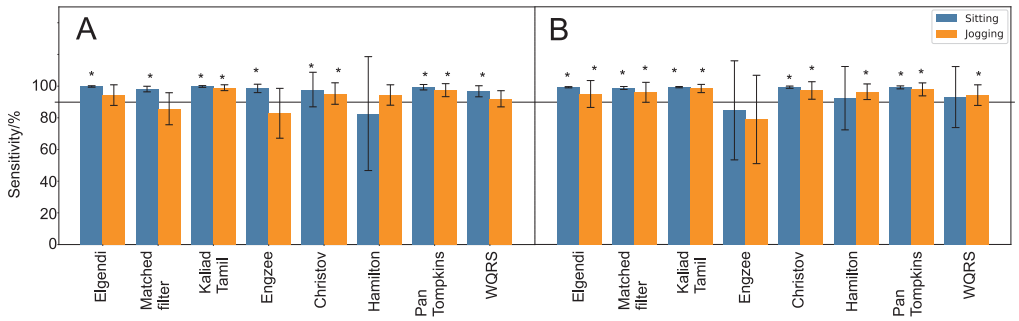


Fig 4. Classical sensitivity values. A: Einthoven, B: Chest strap. Significantly above 90% indicated with an asterisk.

Results

Traditional method: Sensitivity

Fig. 4 shows sensitivity values for the different detectors as calculated with Eq. 4. The asterisk indicates if the sensitivity of a detector is significantly greater than 90%. It can be seen that 22 of the 32 sensitivity values are above 90%. Panel A shows the results for the standard Einthoven II leads for sitting and jogging. Let us first focus on sitting with the Einthoven leads. This is a standard condition comparable to other noise free databases such as the MITDB with virtually no artefacts. Except for the Hamilton detector, all other detectors reach sensitivity values above 90%. Jogging creates movement artefacts through cable movements and muscle noise (see Fig. 2) which then cause sensitivity values for more detectors to drop below 90% such as Elgendi, matched filter, EngZee, Hamilton and WQRS. Fig. 4B shows the same detectors for sitting and jogging but with a chest strap. This reduces the movement artefacts and muscle noise but at the same time, the ECG measured at the chest electrodes has a different shape compared to the standard Einthoven leads. Still, the sensitivity values are all around the 90% threshold and it clearly makes it hard to decide which detector to choose. Overall the large detection error margin between detector timestamp and ground truth results in very high sensitivity readings. Even in the worst case scenario of recording Einthoven II while jogging, the detector performance only marginally drops.

JA

Fig. 5 shows the JA score for the different detectors as calculated with Eq. 16. Overall it is apparent that the introduction of jitter to the performance measure causes a much

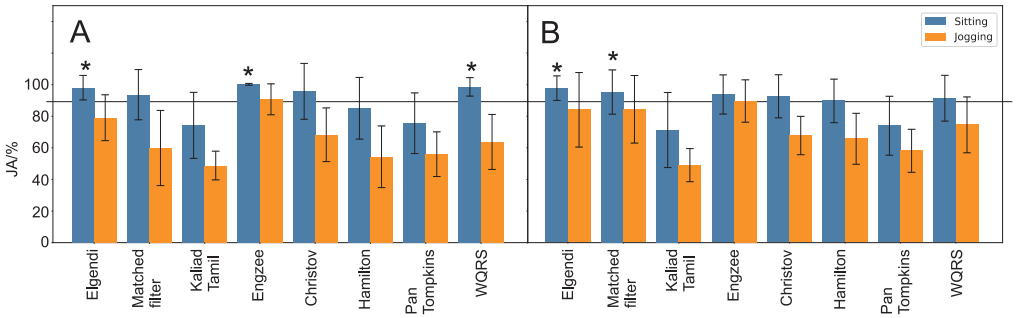


Fig 5. JA values. A: Einthoven, B: Chest strap. Significantly above 90% indicated with a "*".

larger variation of the the JA values than the sensitivity above (Fig. 5). Let us start again with the Einthoven leads while sitting which has the best signal to noise ratio. Now, only the EngZee, the Elgendi and the WQRS detector reach JA values significantly above 90%. When using a chest strap instead, the Elgendi detector and the matched filter reaches JA values above 90%. This shows that the Elgendi detector is robust against different ECG morphologies and keeps its excellent performance irrespective of lead configurations.

When jogging, the signal to noise ratio of the ECG signal drops because of EMG and movement artefacts. Here, clearly the JA score drops which pushes all detectors below the 90% threshold. Note that jogging is the worst case scenario and that the other experiments introduce gradually more and more noise.

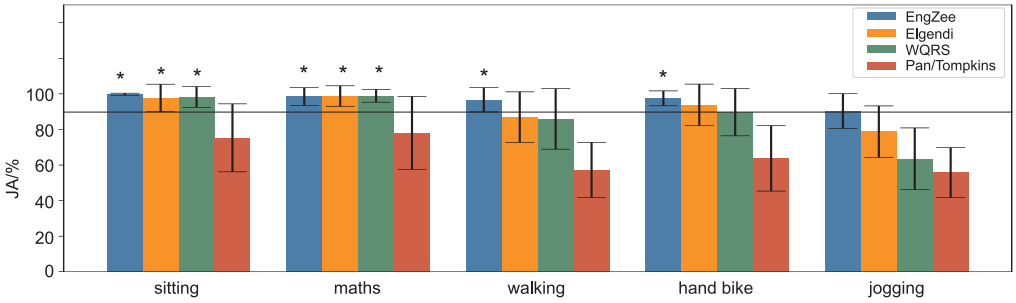


Fig 6. Comparing detector performance against increasing noise. The JA benchmark value was calculated for the different activities of sitting, doing a maths test, walking, operating a hand-bike and jogging. The 4 detectors EngZee, Elgendi, WQRS and Pan& Tompkins have been benchmarked. If a JA value is significantly above 90% it is indicated with an asterisk.

Fig. 6 benchmarks the detectors against increasing noise levels imposed on the ECG. We have taken the 3 detectors EngZee, Elgendi and the WQRS which had excellent performance in the noise free condition (Fig. 5) and subjected them to noise. In addition, we have taken the worst performer, namely the the Pan & Tompkins detector. The maths test introduces small and irregular noise through movement artefacts and muscle noise. We note that all three contemporary detectors EngZee, Elgendi and WQRS perform identically here. As mentioned before, the EngZee detector does a final maximum detection of the QRS-complex which is no longer advantageous in the presence of noise, possibly because of multiple peaks introduced by high frequency EMG noise. Nevertheless, all three detectors (excluding Pan & Tompkins) are significantly above 90%. During walking and while operating a handbike, the performance of the Elgendi and WQRS detectors drops significantly below 90%. Both activities generate

stronger artefacts. The handbike generates more upper body artefacts while walking generates artefacts from muscles in the whole body. Jogging is the worst case scenario where sometimes, even through visual inspection, QRS complexes cannot be identified at all. Here, all detectors are below the 90% mark and thus not reliable enough, certainly for high precision tasks such as heartrate variability. The EngZee detector is the clear winner for noisy ECGs while for low-noise tasks such as sitting in a chair, EngZee, Elgendi and the WQRS perform equally well, producing JA values above 90%. Note that this means that the accuracy was certainly above 90% and that the jitter has not downgraded performance below 90%.

Table 1. Original cited sensitivities with unspecified tolerance, in comparison to the new overall JA Benchmark scores for each detector using GUDB.

	Original cited sensitivity with unspecified tolerance (%)	Sensitivity of Einthoven II, sitting, based on GUDB and $w = 10/f_s$	JA score of Einthoven II, sitting, based on GUDB
Elgendi et al. [29]	92.66 ... 98.31	99.4 ± 0.4	97.7 ± 7.7
Kalidas & Tamil [30]	99.88	99.4 ± 0.6	73.9 ± 20.8
Engzee [25]	96.50	98.2 ± 2.7	99.8 ± 0.6
Christov [31]	84.5 ... 96.5	97.4 ± 10.9	95.4 ± 17.6
Hamilton [31]	99.74 ... 99.80	82.3 ± 35.8	84.7 ± 19.4
Pan & Tompkins [21]	99.30	98.9 ± 1.7	75.3 ± 19.1
Matched filter	N/A	97.7 ± 1.8	93.3 ± 15.9
WQRS [26]	99.65	96.3 ± 3.5	98.2 ± 5.8

Discussion

This work was undertaken in response to the apparent lack of research into how realistic movement noise affects the performance of ECG QRS detection algorithms. To provide a dataset for evaluating the detectors and a resource for future research, an open access database of ECG recordings was created [17]. This database consists of 125 recordings of realistic scenarios, the majority of which have been videoed and annotated with high precision. Eight real-time QRS detection algorithms were implemented in Python to be evaluated, representing a range of popular techniques [27].

Depending on which ECG lead is recorded the QRS complex can have a dominant positive (upright) appearance (an R wave) or the opposite, a Q or S wave. Databases used to benchmark the detection of QRS complexes generally have dominant positive R waves. In 12 lead electrocardiography [3], it is customary to use a function that combines several leads into a single positively constrained signal, effectively a dominant positive R wave. Hence the work discussed in this article is of relevance to QRS detection of any form of electrocardiogram and not simply those from the databases used in this paper.

Pan & Tompkins [21] were one of the first to develop a real-time QRS detection algorithm which was written in assembly language on a Z80 microprocessor to be able to cope with the filtering demands. The performance was analysed by playing tapes of ECG recordings from the MIT/BIH database which were turned into digital signals, processed in the Z80 processor and then turned back into analogue signals for comparison with the annotations. Coincidence between the detected QRS complexes and the annotations was done by visual inspection so that a precise jitter tolerance could not be given. Overall, Pan & Tompkins [21] reported a sensitivity of 99.3 % (Table. 1)

which we have confirmed closely. However, our JA benchmark produces a value of 75.3 ± 19.1 which is not only significantly below the 90% margin but also has a huge standard deviation. The Pan & Tompkins algorithm is the oldest detector benchmarked here and should only be seen in a historical context, but not for any new applications.

Instead of bandpass filtering the ECG, Kalidas & Tamil [8] employed the stationary wavelet transform to filter the ECG which resulted in a high sensitivity in their paper of 99.88 % (Table. 1) which is reproduced by our work here. However, also the cited sensitivities of their competitors are in the 99 % band and thus render any comparison useless. Neither a statistical significance test nor the detection tolerance is published. However, given their 99.88 % sensitivity, it suggests that again a large temporal jitter was permitted. However, the exact margin is not mentioned in the text. Indeed, our JA benchmark is one of the lowest and comparable to the original Pan & Tompkins detector. This comes as no surprise because also Kalidas & Tamil used the original Pan & Tompkins detection but preceded by a wavelet filter instead of a bandpass filter. A wavelet filter is essentially a bandpass filter so that the overall performance is very similar to the Pan & Tompkins detector.

Hamilton [24] had in mind an open source QRS detector with a high sensitivity which is a further development of the detector proposed by Pan & Tompkins [21]. In order to be able to run it on a microcontroller, Hamilton [24] presented a stripped down version of their full detector implementation, omitting certain QRS detection rules. For their original detector, they report a sensitivity of 99.74 % and for the microcontroller version 99.80 % (Table. 1). The sensitivities were not compared to other detectors nor has a statistical analysis has been performed. The authors state that the detectors perform comparably. The jitter tolerance is not mentioned in the paper. However, we report a sensitivity of just 82.3 ± 35.8 % which falls substantially short of the claimed 99.74 % and has a huge standard deviation. This again shows that the Pan & Tompkins detection algorithm cannot be much improved beyond its original performance.

Instead of improving the detection algorithm, the work by Elgendi et al [22] aimed to optimise the bandpass filter frequencies used for the QRS detection. Different frequency bands were benchmarked and the $8 - 20$ Hz band was chosen to be optimal, having the highest sensitivity. However, all originally cited sensitivities, even the sub-optimal ones, vary only between 92.66% and 98.31% (Table. 1) where some deviate less than one percent. There is no statistical analysis in the paper stating which band is significantly better than another. A temporal tolerance is not given. The lack of statistical tests in the original paper and the omission of the tolerance render the recommendations towards cut-off frequencies questionable. While their own analysis is inconclusive, our own JA benchmark (Table. 1) places them into the category of the best performers and is certainly an algorithm which could be used in production.

Christov [23] starts off not just with Einthoven I but creates a complex lead combining the derivatives from all standard 12 leads. This is then sent through various thresholds to detect the QRS complexes. They achieved a sensitivity of 99.96 % at a tolerance of 60 ms (Table. 1). The article has no discussion section and thus does not compare its results to other approaches and has no statistical evaluations. In the conclusion section, the author claims that: "The statistical indices are higher than, or comparable to those, cited in the scientific literature." [23]. This is not confirmed for our JA measure where even when sitting its value is significantly below 90% with also a very high standard deviation.

Similar to Christov [23], taking the derivative of the ECG signal is the central idea of Englese et al [32] which is then turned into a real-time version by Lourenco et al [25] using an adaptive threshold. The paper then compares the real-time version with the previous off-line version and the detector by Christov [23] which requires a fixed threshold. This paper does not use the MIT-DB database because the authors recorded

their own ECGs. They also devised a different performance measure: the deviation from the mean RR interval needs to be less than 10 % for individual RR pairs or in other words, the heart rate needs to be within normal limits of the resting heart rate variability [33]. Their online algorithms lead to average valid RR intervals between 84.5 % and 96.5 % (Table. 1) depending on on- or offline algorithms, use of electrodes, the algorithm itself and filtering. These high readings are expected because their criterion will most likely detect only crude deviations from RR intervals, for example a missed beat, which then results in twice the RR interval being measured or an additional spurious detection, which results in a very short RR interval. Even more important is that this measure does not compare against the ground truth of RR intervals and only looks at the self consistency of the RR intervals, which might have been wrong in relation to the annotations in the first place. There has been no statistical test to determine which of these results differ significantly but one could have been easily performed with the reasonably large number of subjects available ($N = 62$). Our JA benchmark produces a reasonably high JA value but with a large standard deviation.

Overall, virtually every paper reporting very high sensitivities of 98 % or more is not helpful at all and has only been possible because of high temporal tolerances of probably 100 ms or even more. In addition, the sole use of the MIT-BIH database with mostly artefact-free ECGs again overestimates the performance of these algorithms. These factors combined grossly overestimate the performances of the algorithms, thus making a comparison impossible. Our new JA benchmark, taking into account both the the all-or-nothing errors such as missed beats and the continuous jitter, produces values which allow meaningful comparisons between detectors. We conclude that EngZee, Elgendi and the WQRS are the best detectors available.

Conflict of Interest Statement

B.P. is CEO of Glasgow Neuro LTD which manufactures the Attys DAQ board.

References

1. Stroobandt RX, Barold SS, Sinnaeve AF. ECG from Basics to Essentials: Step by Step. John Wiley & Sons; 2016.

2. Davies A, Scott A. ECG Basics. In: Starting to Read ECGs. Springer; 2014. p. 19–33.

3. Macfarlane PW. Lead Systems. In: Macfarlane PW, van Oosterom A, Pahlm O, Kligfield P, Janse M, Camm J, editors. Comprehensive electrocardiology: theory and practice in health and disease. Springer; 2010.

4. Petelczyc M, Gierałtowski JJ, Żogała-Siudem B, Siudem G. Impact of observational error on heart rate variability analysis. Heliyon. 2020;6(5):e03984. doi:10.1016/j.heliyon.2020.e03984.

5. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):E215–20.

6. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society. 2001;20(3):45–50.

7. Benitez DS, Gaydecki PA, Zaidi A, Fitzpatrick AP. A new QRS detection algorithm based on the Hilbert transform. Computers in Cardiology. 2000;27:379–382. doi:10.1109/CIC.2000.898536.

8. Kalidas V, Tamil L. Real-time QRS detector using Stationary Wavelet Transform for Automated ECG Analysis. In: 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE; 2017. p. 457–461. Available from: `http://ieeexplore.ieee.org/document/8251332/`.

9. Chan HL, Wang FT, Lee YS, Wang CL. Heartbeat Detection Using Oscillatory Envelope Pattern in Noisy Electrocardiogram. Computing in Cardiology. 2017; p. 1–4. doi:10.22489/CinC.2017.086-353.

10. Moody G, Muldrow W, Mark R. A noise stress test for arrhythmia detectors. Computers in Cardiology. 1984;11:381–384.

11. Behravan V, Glover NE, Farry R, Chiang PY, Shoaib M. Rate-adaptive compressed-sensing and sparsity variance of biomedical signals. In: 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE; 2015. p. 1–6.

12. Luz EJdS, Schwartz WR, Cámara-Chávez G, Menotti D. ECG-based heartbeat classification for arrhythmia detection: A survey. Computer methods and programs in biomedicine. 2016;127:144–164. doi:10.1016/j.cmpb.2015.12.008.

13. Xie C, Dubiel J. WFDB Python; 2018. Available from: `https://github.com/MIT-LCP/wfdb-python`.

14. Arzeno NM, Deng ZD, Poon CS. Analysis of first-derivative based QRS detection algorithms. IEEE Transactions on Biomedical Engineering. 2008;55(2):478–484. doi:10.1109/TBME.2007.912658.

15. Gradl S, Leutheuser H, Elgendi M, Lang N, Eskofier BM. Temporal correction of detected R-peaks in ECG signals: A crucial step to improve QRS detection algorithms. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2015. p. 522–525.

16. Liu F, Liu C, Jiang X, Zhang Z, Zhang Y, Li J, et al. Performance analysis of ten common QRS detectors on different ECG application cases. Journal of healthcare engineering. 2018;2018. doi:10.1155/2018/9050812.

17. Howell L, Porr B. High precision ECG Database with annotated R peaks, recorded and filmed under realistic conditions; 2018. Available from: `http://researchdata.gla.ac.uk/716/`.

18. Macfarlane PW, Coleman EN. Resting 12-lead ECG electrode placement and associated problems. The Society for Cardiological Science and Technology; 1995. Available from: `http://www.scst.org.uk/resources/RESTING_12.pdf`.

19. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007;9(3):90–95. doi:10.1109/MCSE.2007.55.

20. Carreiras C, Alves A, Lourenço A, Canento F, Silva H, Fred A. BioSPPy - Biosignal Processing in Python; 2018. Available from: `https://github.com/PIA-Group/BioSPPy/`.

21. Pan J, Tompkins WJ. A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering. 1985;BME-32(3):230–236. doi:10.1109/TBME.1985.325532.

22. Elgendi M, Jonkman M, Deboer F. Frequency Bands Effects on QRS Detection. In: Biomedical Engineering Systems and Technologies. Springer; 2010. p. 428–431. Available from: `http://www.elgendi.net/papers/QRS_Bands_final.pdf`.

23. Christov II. Real time electrocardiogram QRS detection using combined adaptive threshold. BioMedical Engineering OnLine. 2004;3(1):28. doi:10.1186/1475-925X-3-28.

24. Hamilton P. Open source ECG analysis. In: Computers in Cardiology. Memphis, TN, USA: IEEE; 2002. p. 101–104. Available from: `http://ieeexplore.ieee.org/document/1166717/`.

25. Lourenço A, Silva H, Leite P, Lourenço R, Fred AL. Real Time Electrocardiogram Segmentation for Finger based ECG Biometrics. In: Biosignals; 2012. p. 49–54.

26. Zong W, Moody GB, Jiang D. A robust open-source algorithm to detect onset and duration of QRS complexes. In: Computers in Cardiology, 2003; 2003. p. 737–740.

27. Porr B, Howell L, Stournaras I, Nir Y. Popular ECG R-peak detectors written in python; 2023. Available from: `https://doi.org/10.5281/zenodo.7652725`.

28. Porr B, Forsyth E. QRS detector stress test combining temporal Jitter and Accuracy (JA); 2023. Available from: `https://doi.org/10.5281/zenodo.7653718`.

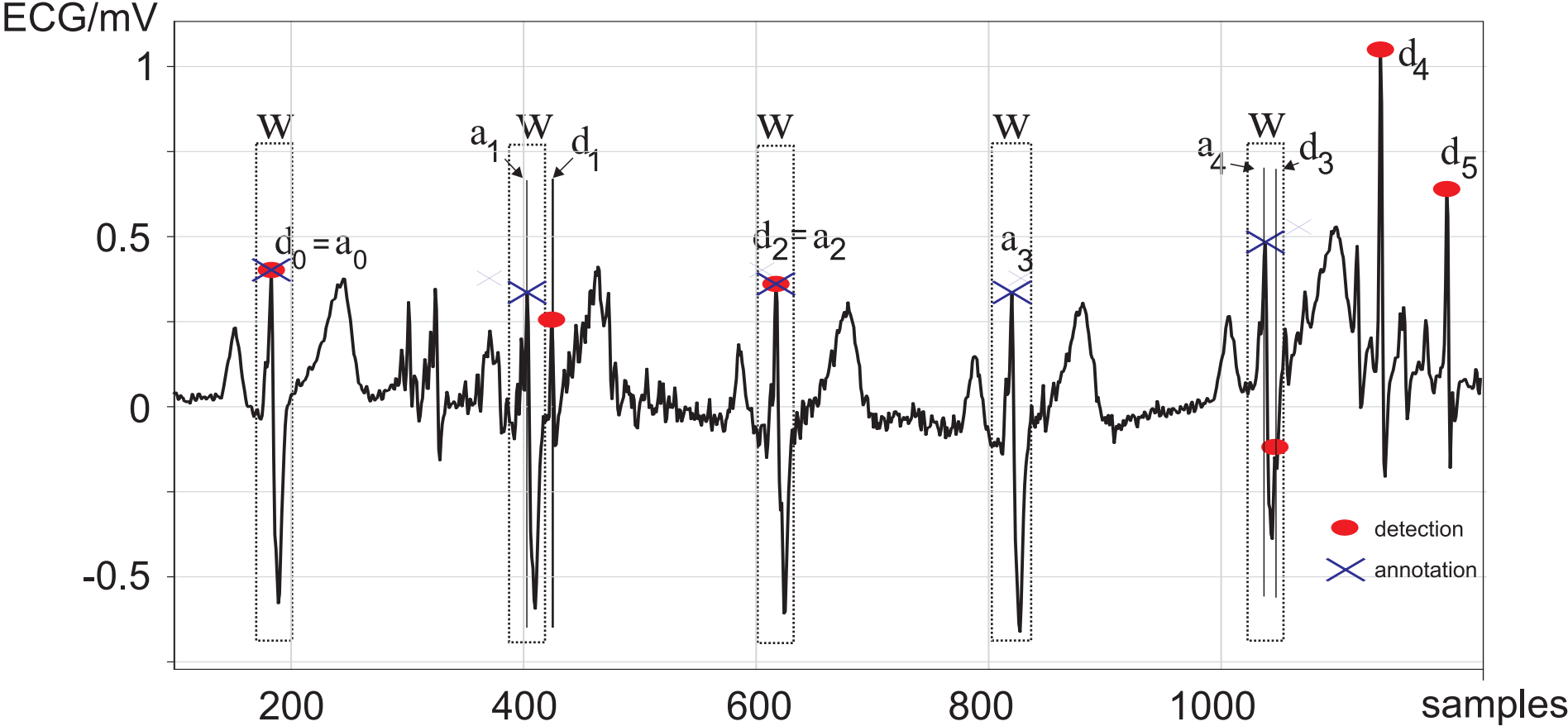
29. Elgendi M, Jonkman M, De Boer F. Frequency Bands Effects on QRS Detection. Biosignals. 2010;2003:2002. doi:10.5220/0002742704280431.

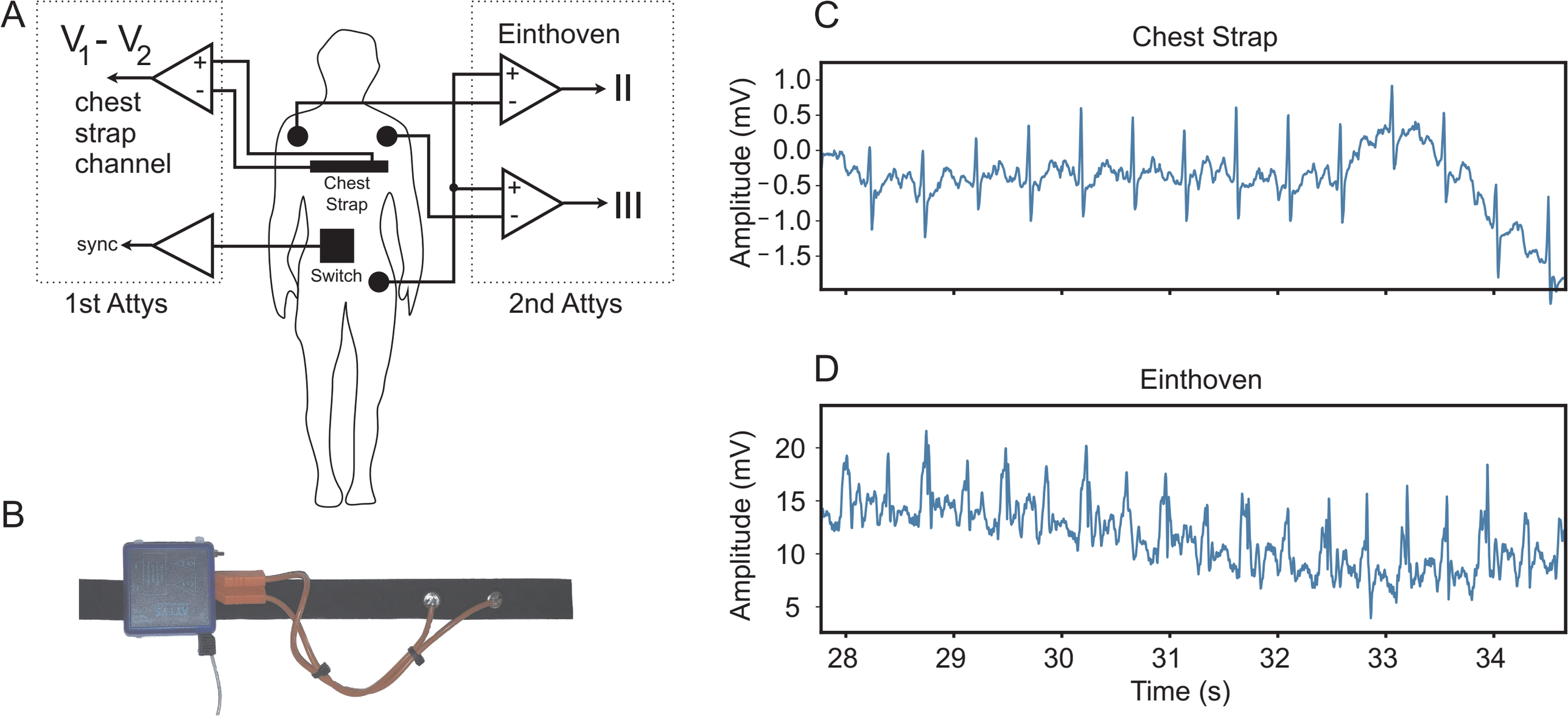
30. Kalidas V, Tamil L. Real-time QRS detector using stationary wavelet transform for automated ECG analysis. In: 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE; 2017. p. 457–461.

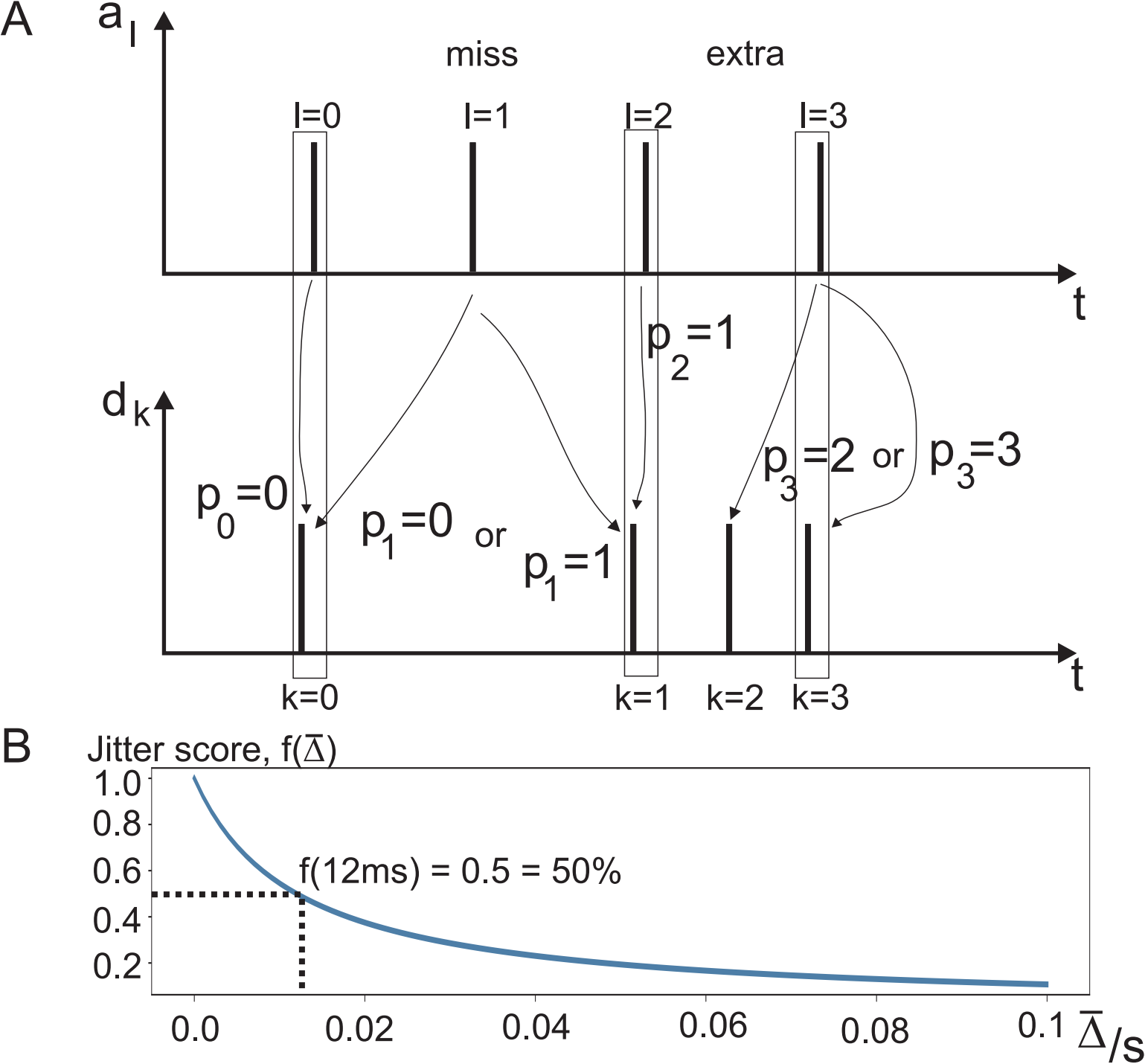
31. Christov II. Real time electrocardiogram QRS detection using combined adaptive threshold. Biomedical engineering online. 2004;3(1):1–9. doi:10.1186/1475-925X-3-28.

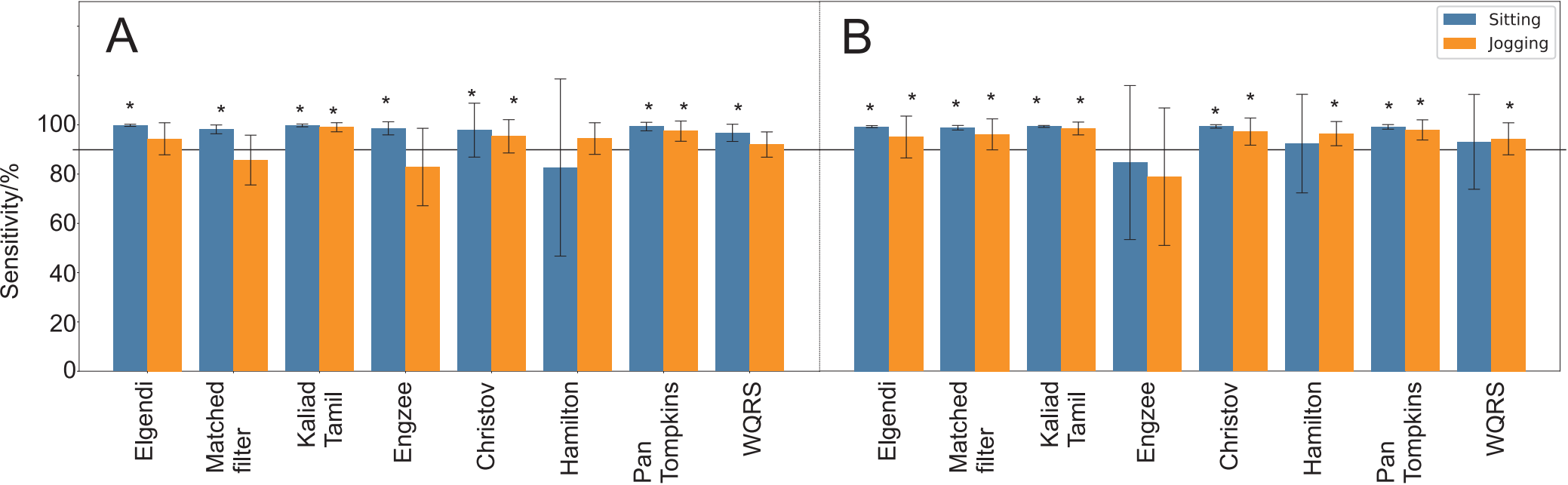
32. Englese W, Zeelenberg C. A single scan algorithm for QRS-detection and feature extraction. Computers in Cardiology. 1979;6:37–42.

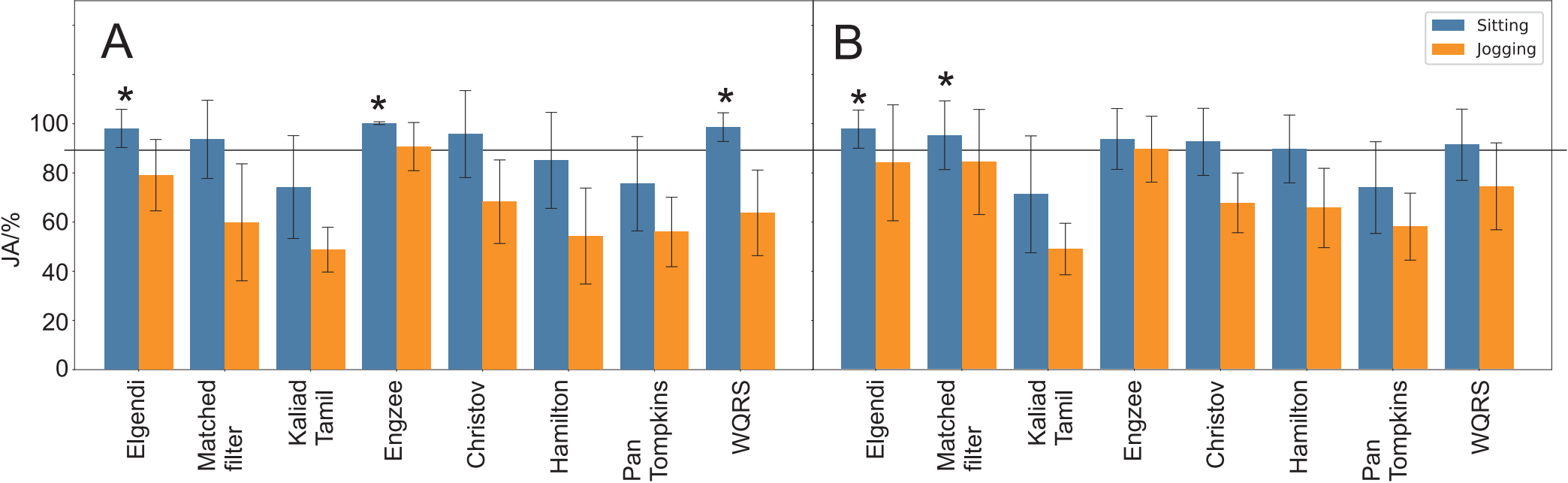
33. Shaffer F, Ginsberg JP. An Overview of Heart Rate Variability Metrics and Norms. Frontiers in public health. 2017;5:258. doi:10.3389/fpubh.2017.00258.

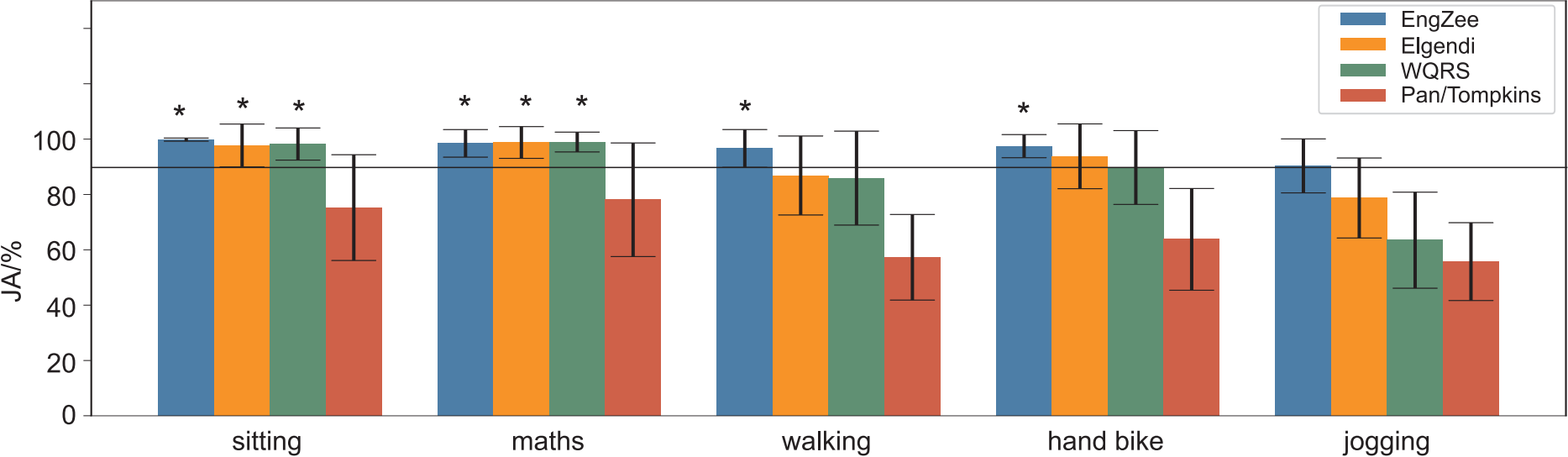

ECG/mV
1
0.5
0
-0.5
W
W
W
W
W
a_1
d_1
$d_0 = a_0$
$d_2 = a_2$
a_3
a_4
d_3
d_4
d_5
detection
annotation
200
400
600
800
1000
samples


A
$V_1 - V_2$
chest
strap
channel
sync
1st Attys
Chest
Strap
Switch
Einthoven
II
III
2nd Attys
B
C
Chest Strap
Amplitude (mV)
1.0
0.5
0.0
−0.5
−1.0
−1.5
D
Einthoven
Amplitude (mV)
20
15
10
5
28
29
30
31
32
33
34
Time (s)


A
a_l
miss
extra
l=0
l=1
l=2
l=3
t
$p_2=1$
d_k
$p_3=2$ or $p_3=3$
$p_0=0$
$p_1=0$ or $p_1=1$
t
k=0
k=1
k=2
k=3
B
Jitter score, $f(\bar{\Delta})$
1.0
0.8
0.6
0.4
0.2
f(12ms) = 0.5 = 50%
0.0
0.02
0.04
0.06
0.08
0.1
$\bar{\Delta}$/s


A
B
Sitting
Jogging
Sensitivity/%
0
20
40
60
80
100
Elgendi
Matched filter
Kaliad Tamil
Engzee
Christov
Hamilton
Pan Tompkins
WQRS


A
B
Sitting
Jogging
JA/%
0
20
40
60
80
100
*
Elgendi
Matched filter
Kaliad Tamil
Engzee
Christov
Hamilton
Pan Tompkins
WQRS


EngZee
Elgendi
WQRS
Pan/Tompkins
JA/%
0
20
40
60
80
100
*
sitting
maths
walking
hand bike
jogging