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

Bernd Porr

School of Engineering, University of Glasgow, Glasgow, United Kingdom

bernd.porr@glasgow.ac.uk

Abstract

The R peak detection of 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 R peak 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 lead to the development of a new performance measure which combines temporal Jitter with Accuracy (JA) which in turn reflects the 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 algorithms.

1 Introduction

Detecting the presence of a heartbeat by recording an electrocardiogram (ECG) is the first step to calculating any further heartbeat parameters. Most heartbeat detection algorithms will use the distinctive R-peak of this waveform as their intended detection point [1–3]. When the presence of a heartbeat 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 R-peak timestamps, which is mathematically the same as finding the first derivative. However, derivatives amplify noise and thus any errors in the R-peak detection will result in even larger errors as further derivatives are taken. When calculating HRV the 2nd derivative between successive heart beat readings is used, and thus any small errors in the heart beat detection will compound into much larger errors in your subsequent HRV result [4]. As such, a heart beat detector algorithm that can accurately and precisely identify the R-peaks in an ECG, is vital in obtaining a useful and representative HRV. All the errors that can occur from poor R-peak detector algorithms are; missed heart beats (false negatives), extra heart beats (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 R-peaks 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 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 so is therefore unsuitable for detector evaluation.

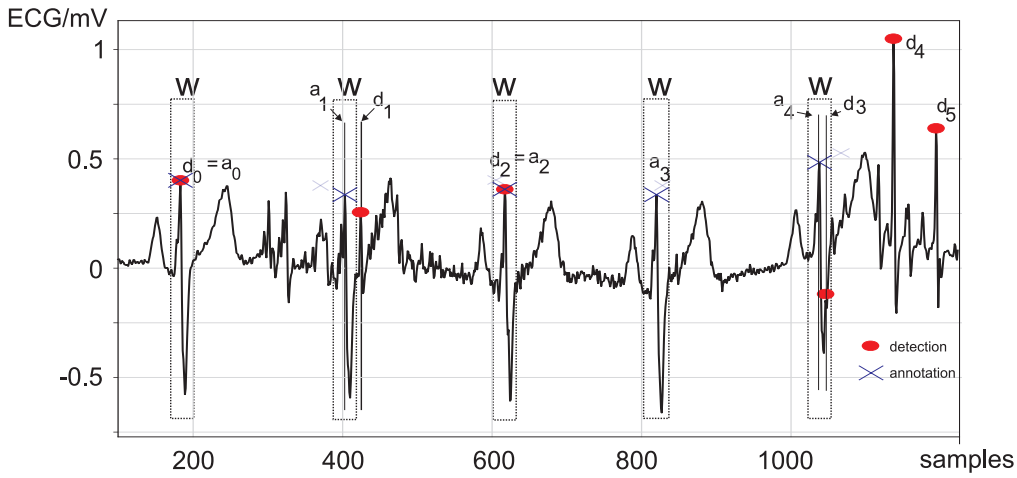

Fig 1. Examples of ECG detection errors. $d_{0...4}$ are sample numbers of detected ECG peaks. $a_{0...4}$ are sample numbers of the annotated ECG peaks (ground truth). w is the temporal window used for the classical performance measures sensitivity, positive predictivity, false positive rate and overall 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 if an R-peak falls in a temporal window w and is illustrated in Fig. 1. a_l are annotated sample positions where the actual R-peak has occurred. d_k are the sample positions where the detector has detected the ECG. For example in Fig. 1 the 1st R-peak 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 peak d_1 but this falls outside of the window $|d_1 - a_1| > w$ so that it's a false positive. Of course there are also cases where there is an annotation as with a_3 but no matching detection so that we have a false negative. There will be also 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:

$$TP = \text{card}(|a_l - a_k| < w) \quad (1)$$

$$FP = \text{card}(d_k) - TP \quad (2)$$

$$FN = \text{card}(a_l) - TP \quad (3)$$

where FN is the number of false negatives, and FP is the number of false positives. 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 one-tenth of the sampling rate 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]:

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

For example at a sampling rate of 250 Hz this gives a temporal window of 40 ms which means that any R-peak within this window are 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 detectors 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 but falling within the window w. [15] proposed a possible measure to account for R-peak inaccuracies using the MITDB. However, [15] uses absolute deviation from each R-peak 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. [15] does not attempt to combine their R-peak 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.

2 Methods

2.1 Glasgow University Database

The Glasgow University Database (GUDB) consists of two-minute 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

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]. The study was approved by the University of Glasgow ethics committee.

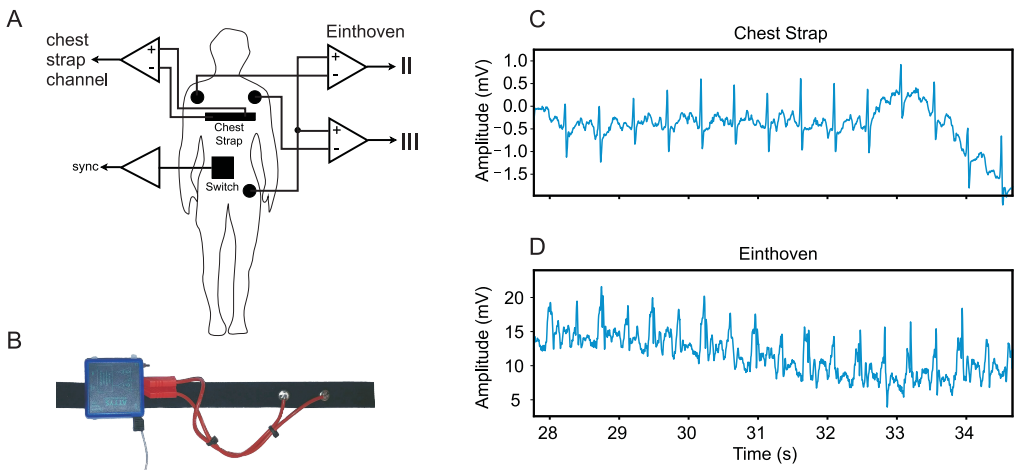

Fig 2. A) Wiring of the two two channel wireless biosignal amplifiers. B) Photo of the chest strap and wiring of the amplifier. The GND was placed on the shoulder. Comparison between chest strap (C) and Einthoven recording (D) while subject was jogging.

2.1.1 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 Hz and a resolution of 24-bit 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 analogue recording channels, one through an amplifier and the other through a differential amplifier. Two Attys were used at the same time to record different configurations representing 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 an 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 V1 and V2 in the standard six electrode chest configuration [18]. The left electrode on

the chest strap is connected to the positive terminal of the differential amplifier and the right electrode is connected to the negative. The GND terminal was connected to a silver chloride electrode (Pulse Medical, UK) on the right shoulder. 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 GND. 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 differential amplifier is connected to the left hip, the negative terminal to the right shoulder and the GND terminal to the left shoulder. The Attys is put into ECG mode where the positive terminal of the CH1 differential amplifier is connected internally to the 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 GND on 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 as shown 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.

2.1.2 Protocol

Before the experiment begins, the participant will read the information sheet and sign the consent form (see [17]), opting in or out of the video recording. The recording equipment is then connected to the participant and the experiment runs 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 will be removed from participant

2.1.3 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 ran through a heartbeat detection algorithm (engzee segmenter from the BioSPPy library [20]) to provide an initial estimation of R-peak locations. This estimation is then manually inspected to remove any false positives and add any missing R-peaks. 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.

2.2 Detector Software Implementation

The eight algorithms chosen represent popular and well regarded heartbeat detectors, as well as, a range of different detection techniques. The seven detector algorithms are; Pan and Tompkins [21], Elgendi et al. [22], Kalidas and Tamil [8], Christov [23], Hamilton [24], the matched filter detector, the Engzee Mod detector [25] and the WQRS detector [26]. The main criteria 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].

2.3 Evaluating Detectors

2.3.1 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 $w = f_s/10$ samples was used [13]. Note that all detectors will delay the signals to some extent, so the detected timestamps will always appear to be later. Thus, in order to allow sample precision evaluation, the median delay was subtracted from the true R peak timestamp. The median delay was calculated separately for every subject, activity, and measurement protocol, as it will be different each time.

The performance of the detector algorithms were compared using the current most commonly recommended performance parameter: sensitivity (Eq. 4). The sensitivities for sitting, jogging, 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.

2.3.2 JA

While the above approach simply assess the sensitivity performance when R-peaks occur either within a temporal window w or not, we are now presenting a new benchmarking approach 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): JA. In contrast to the traditional approach, no fixed temporal window is used but the accuracy is rather penalised more and more with increasing jitter. Fig. 3 demonstrates the working principle of the JA algorithm which is now explained step by step.

It's input are the sample points d_k of an r-peak detector which are compared against the r-peak annotations a_l. As a first step we need to calculate the median delay between the detections d_k and annotations a_l:

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

As a next step we loop through all annotation indices l and find for each l the index within the detections $\tilde{d_k}$ which has the smallest temporal difference:

$$p_l = \arg\min_k |a_l - \tilde{d_k}| \tag{8}$$

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 demonstrated for the annotation $l = 1$ there is no matching detection and, thus, it's 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 now loop through the p_l indices and check if there are duplicates. For example here $p_0 = p_1 = 0$

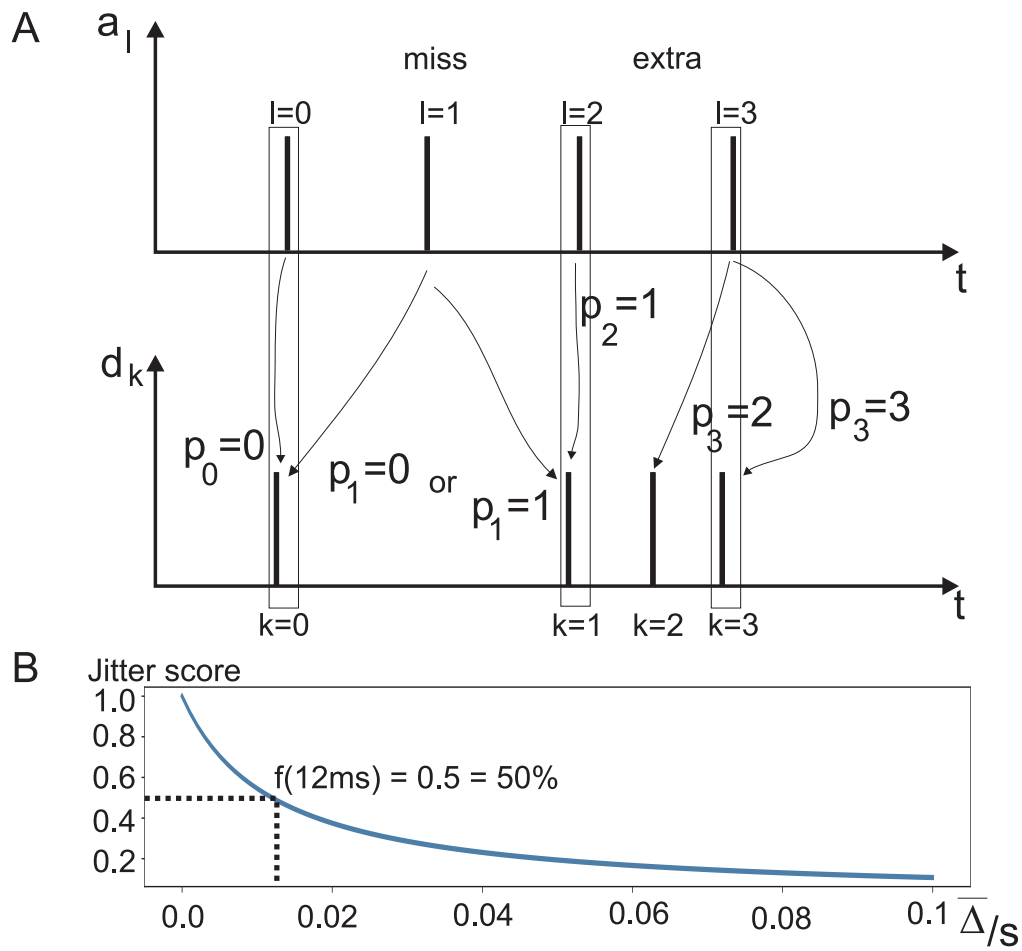

Fig 3. Graphical illustration of the JA algorithm

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 $p_1 = p_2 = 1$ were the duplicate position. Similarly an extra beat were generating two matches but again one has the lowest temporal difference. Thus, formally we minimise the temporal difference between the detection indices while keeping the result unique:

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

where Δ is now the jitter between true detection pairs. The number of elements of Δ are the true positive detections.

As a next step Δ needs to be normalised and evaluated.

$$\bar{\Delta} = \frac{\Delta}{12 \text{ ms}} \tag{10}$$

where $\bar{\Delta}$ is the normalised jitter. Since we have the goal to arrive at a performance measure between 0 and 100% we need to map the normalised jitter now with a mapping function which reflects perfect operation, perfect for HRV and failure not to detect anything. We define the following mapping function

$$f(\bar{\Delta}) = \frac{1}{1 + \bar{\Delta}} \tag{11}$$

which means that at a jitter of 12 ms we have a performance of 0.5 or 50 %.

So far we have only dealt with the jitter but the number of elements of the detection pairs card$\bar{\Delta}$ = true positives (TP) can now be used to calculate the number of missed beats (false negatives = FN) and extra beats (false positives = FP):

$$TP = \text{card}(\bar{\Delta}) \tag{12}$$
$$FP = \text{card}(d_k) - \text{card}(\bar{\Delta}) \tag{13}$$
$$FN = \text{card}(a_l) - \text{card}(\bar{\Delta}) \tag{14}$$

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

$$TN = \frac{N}{f_s}\frac{\text{HR}_{\text{max}}}{60\text{bpm}} - (TP + FP + FN) \tag{15}$$

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

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

The accuracy A is now a normalised measure between zero and one. As a final step we can now combine the jitter and the accuracy:

$$\text{JA} = A \cdot f(\bar{\Delta}) \cdot 100\% \tag{17}$$

The JA reaches 100% if the jitter is zero 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 the JA

The JA performance values (Eq. 17) for sitting, jogging, 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.

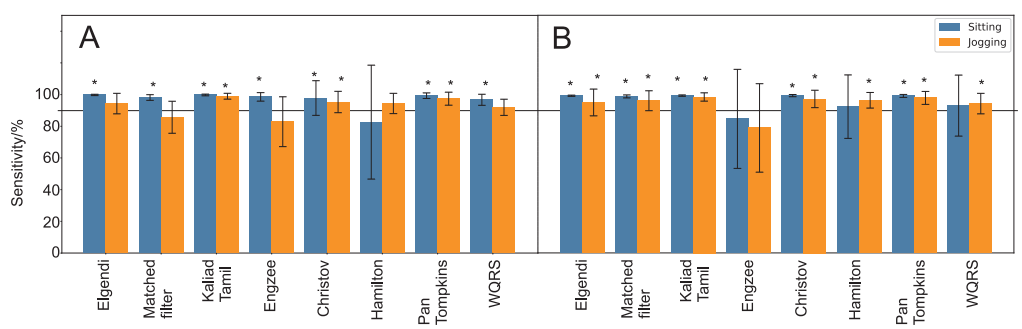

Fig 4. Classical sensitivity values. A: Einthoven, B: Chest strap. Significantly above 90% indicated with a "*".

3 Results

3.1 Traditional method: Sensitivity

Fig. 4 shows sensitivity values for the different detectors as calculated with Eq. 4. The "*" indicates if the sensitivity of a detector is significantly greater than 90%. Overall we

note 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 causes 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.

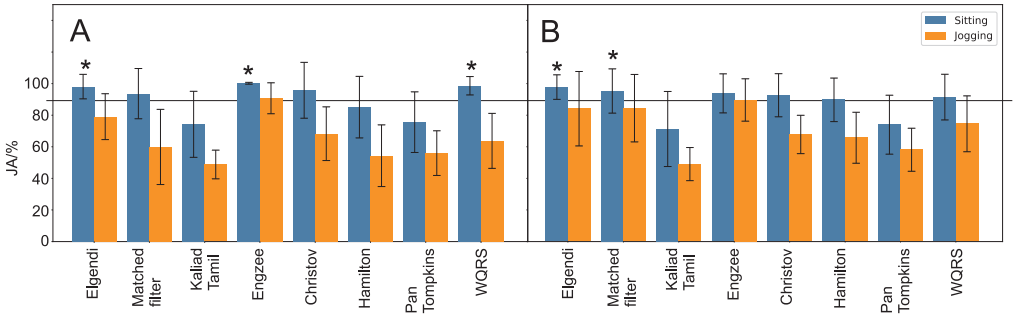

Fig 5. JA values. A: Einthoven, B: Chest strap. Significantly above 90% indicated with a "*".

3.2 JA

Fig. 5 shows the JA score for the different detectors as calculated with Eq. 17. Overall it is apparent that the introduction of jitter to the performance measure causes a much larger variation of the the JA values than the sensitivity above. Let's 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 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 shapes and keeps its excellent performance.

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 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: the Pan & Tompkins detector. The math test introduces small and irregular noise through movement artefacts and muscle noise. We note that all three contemporary detectors EngZee, Elgendi and WQRS performing identically here. As mentioned before the EngZee detector does a final maximum detection of the R-peaks which is no longer advantageous in the presence of noise possibly because of multiple peaks introduced by high frequency EMG noise. Still all

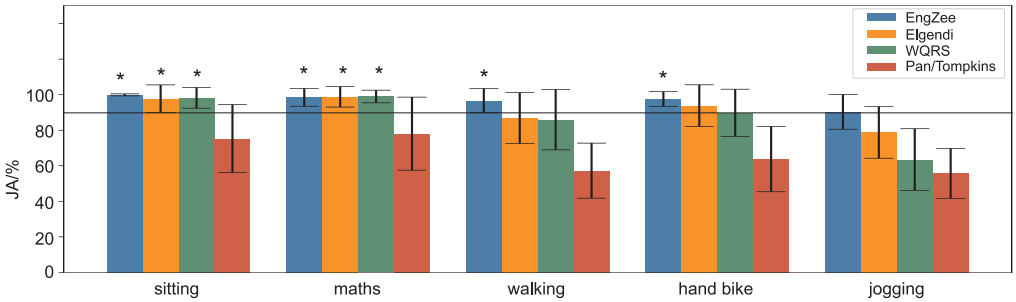

Fig 6. Comparing detector performance against increasing noise. The JA benchmark value was calculated for the different activities of sitting, doing a math test, walking, operating a hand-bike and jogging. The 4 detectors EngZee, Elgendi, WQRS and Pan& Tompkins have been benchmarked. If a JA value is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

three detectors (excluding Pan & Tompkins) are significantly above 90%. During walking and while operating a handbike the performance of the Elgendi and WQRS detectors drop significantly below 90%. Both activities generate stronger artefacts where the handbike generates more upper body artefacts while walking generates artefacts form muscles in the whole body. Jogging is the worst case scenario where sometimes even through visual inspection R-peaks cannot be identified at all. Here, all detectors are below the 90% mark and thus not reliable enough certainly for high precision tasks as heartrate variability. The EngZee detector is the clear winner for noisy ECGs while for low-noise tasks sitting in a chair EngZee, Elgendi and the WQRS perform equally well at JA values above 90%. Note that this means that the accuracy was certainly above 90% and that the jitter hasn't downgraded them 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, sitting based on GUDB and $w = f_s/10$	Overall JA Benchmark score Einthoven II, sitting
Elgendi et al. [28]	92.66 ... 98.31	99.4 ± 0.4	97.7 ± 7.7
Kalidas and Tamil [29]	99.88	99.4 ± 0.6	73.9 ± 20.8
Engzee [30]	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 and Tompkins [32]	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

4 Discussion

This work was undertaken in response to the apparent lack of research into how realistic movement noise effects the performance of ECG heartbeat 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 heartbeat detection algorithms were implemented in Python to be evaluated, representing a range of popular techniques [27].

Pan & Tompkins [21] were one of the first who developed 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 to the annotations. Coincidence between the detected R peak and the annotations was done by visual inspection so that a precise jitter tolerance couldn't be given. Overall [21] reports a sensitivity of 99.3 % (Table. 1) which we've 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 is the oldest detector benchmarked here and should only be seen in the historical context but not for any new applications.

Instead of bandpass filtering the ECG Kalidas [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 the 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 and Tompkins detector. This comes at no surprise as also Kalidas use the original Pan and Tompkins detection but preceded by a wavelet filter instead of a bandpass. 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 R peak detector with a high sensitivity which is a further development of the detector proposed by [21]. In order to be able to run it on a microcontroller, [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 one of 99.80 % (Table. 1). The sensitivities were not compared to other detectors nor 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 [22] aimed to optimise the bandpass filter frequencies used for the R-peak 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 one. A temporal tolerance is not given. 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 R peaks. 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 by Englese [33] which is then turned into a real-time version by Lourenco [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 heartrate needs to be within normal limits of the resting heartrate variability [34]. 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 length 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 1st place. There is no statistical test which of these results differ significantly but could have been easily performed with the reasonably large number of subjects ($N = 62$). Our JA benchmark produces a reasonably high JA value but with a large standard deviation.

Overall, with virtually every paper reporting very high sensitivities of 98 % or more is not helpful at all and have 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 overestimate the performance of these algorithms. These factors combined grossly overestimate the performances of the algorithms and 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 in that order are the best detectors available.

Acknowledgments

We thank Peter Macfarlane for constructive feedback on the Manuscript.

Conflict of Interest Statement

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

References

1. Niebauer J, Ashley E. Cardiology explained. Remedica; 2004.
2. Davies A, Scott A. ECG Basics. In: Starting to Read ECGs. Springer; 2014. p. 19–33.
3. Stroobandt RX, Barold SS, Sinnaeve AF. ECG from Basics to Essentials: Step by Step. John Wiley & Sons; 2016.

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 A. Real Time Electrocardiogram Segmentation for Finger Based ECG Biometrics. In: Biosignals 2012; 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. Elgendi M, Jonkman M, De Boer F. Frequency Bands Effects on QRS Detection. BIOSIGNALS. 2010;2003:2002. doi:10.5220/0002742704280431.

29. 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.

30. 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.

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. Pan J, Tompkins WJ. A real-time QRS detection algorithm. IEEE transactions on biomedical engineering. 1985;(3):230–236. doi:10.1109/TBME.1985.325532.

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

34. 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.