Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Causal Peak Refinement Reduces Timing Error of a Classical R-Peak Detector Across Beat Morphologies

Sadekatul Shihab

241311094@vu.edu.bd

Varendra University https://orcid.org/0009-0004-3575-1426

Akib Ikbal

Varendra University https://orcid.org/0000-0003-4833-036X

Research Article

Keywords: ECG, R-peak detection, Pan-Tompkins, Simulink, Timing error, Added noise

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11728551/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Causal Peak Refinement Reduces Timing Error of a Classical R-Peak Detector Across Beat Morphologies

Sadekatul Shihab[1], Akib Ikbal[1]

[1] Varendra University, Department of Computer Science and Engineering, Rajshahi, Bangladesh

Corresponding author: Sadekatul Shihab, 241311094@vu.edu.bd

ORCID: Sadekatul Shihab 0009-0004-3575-1426; Akib Ikbal 0000-0003-4833-036X

Abstract

Purpose Classical ECG R-peak detectors are usually reported with a single aggregate score, which can hide failures on unusual beats and under noise. One such detector was evaluated by beat type, noise type and matching tolerance, and a refinement stage was tested.

Methods A simplified Pan-Tompkins-style detector and a causal peak-refinement stage, which reassigns each detection to the largest absolute-amplitude sample of the 5-15 Hz band-passed signal within ±50 ms, were implemented in Simulink. Nine MIT-BIH Arrhythmia Database records (first 120 s each) were used: five to choose settings and four held out. Noise from the MIT-BIH Noise Stress Test Database was added at 24, 12, 6 and 0 dB.

Results At ±50 ms the baseline's mean F1 was 92.5% (88.1% on a right bundle branch block record and 63.4% on a paced record), but it was 98.2% at ±75 ms, indicating mainly timing errors. Refinement reduced the mean absolute timing error from 25.8 ms to 6.0 ms and raised the ±50 ms mean F1 to 98.0%. Refined F1 was equal or higher on all four held-out records. Electrode motion at 0 dB was the hardest noise (tuning records: mean F1 87.3% baseline, 91.7% refined). Refinement did not help record 207, which contains ventricular flutter and unmatched premature ventricular contractions (F1 86.3%).

Conclusion Most errors under the ±50 ms matching criterion were timing errors, and a simple causal refinement stage reduced them. The results come from nine short records and one detector, and they do not rank detectors.

Keywords ECG; R-peak detection; Pan-Tompkins; Simulink; Timing error; Added noise

Introduction

The electrocardiogram (ECG) records the electrical activity of the heart. Each heartbeat produces a P wave, a QRS complex and a T wave, and the sharp R peak inside the QRS complex is the standard fiducial point for timing beats. Locating R peaks is the first step in most ECG analysis. Heart rate, heart rate variability (HRV), rhythm classification and arrhythmia screening all depend on it.

Classical R-peak detectors are built from hand-designed signal processing stages and need no training data. The best known is the algorithm of Pan and Tompkins [1], which band-pass filters the signal, differentiates and squares it, integrates it over a moving window and applies adaptive thresholds. Hamilton and Tompkins [2] studied the effect of individual decision rules on the MIT-BIH database. Review articles describe many other approaches, including wavelet, Hilbert-transform and morphological methods [3], and wavelet-based detectors remain an active line of work [4-6]. Classical detectors are cheap and easy to implement in embedded hardware, so they remain a common reference point.

Detectors are usually reported by a single aggregate score, such as sensitivity and positive predictivity over a whole database. This can hide failures. Filters and thresholds are tuned for the shape of a normal beat, so beats with other shapes (wide ventricular beats, bundle branch block complexes, paced beats) may be detected but localized away from the true R peak. Noise can also mask beats or create false ones, and detectors differ in how sensitive they are to different noise types [7]. A detection usually counts as correct if it lies within a fixed tolerance of the reference annotation. A detector that finds every beat but places it imprecisely can look poor under a strict tolerance and good under a loose one. A beat with no nearby detection is a detection failure, whereas a beat detected at the wrong time is a timing failure, and R-peak localization is known to matter for applications such as HRV analysis [8].

These issues have been studied before. Portet et al. [9] compared QRS detectors under combinations of beat morphology and clinical noise. Reklewski et al. [10] evaluated detectors with attention to temporal accuracy and noise, and showed that measured accuracy depends strongly on the detector time tolerance. Gradl et al. [11] proposed temporal correction of detected R-peaks as a step to improve QRS detection. Deep-learning R-peak detectors now also report high accuracy on MIT-BIH and on noisy recordings [12, 13]. No comparison with them is made in this study.

The present study applies these ideas to a classical detector implemented as a Simulink block diagram. The contributions of this study are:

1. An evaluation of a simplified Pan-Tompkins-style detector by beat type (normal beats, left bundle branch block (LBBB), right bundle branch block (RBBB), premature ventricular contraction (PVC) and paced beats), by noise type and level (baseline wander, electrode motion and muscle artifact at 24, 12, 6 and 0 dB signal-to-noise ratio (SNR)), and by matching tolerance.
2. A diagnosis, from detection offsets and from a comparison of ±50 ms and ±75 ms tolerances, that most of the baseline's errors on wide-QRS and paced beats are timing errors rather than the absence of a nearby detection.
3. A causal Simulink implementation of a peak-refinement stage in the spirit of temporal correction [11], with settings chosen on five tuning records and tested on four held-out records.

4. A timing-precision analysis and a failure analysis of the record where refinement does not help.

The results are limited to nine records of 120 s each and one baseline detector. They are reported as an evaluation of this specific detector, not as a ranking of detectors.

Materials and Methods

Data

The MIT-BIH Arrhythmia Database [14] and the MIT-BIH Noise Stress Test Database (NSTDB) [15, 16], both distributed through PhysioNet [17], were used. The first channel of each record (modified limb lead II (MLII) in all nine records used here [18]) was analyzed at its native 360 Hz sampling rate and converted to millivolts using the gain (200 analog-to-digital converter (ADC) units/mV) and baseline (1024) in the record header. The first 120 s of each record was used.

Nine records were split into two sets. The tuning set (records 100, 107, 109, 118, 119) was used to choose the detector parameters and for the primary added-noise experiment. The held-out set (records 105, 200, 207, 217) was used for the final evaluation, and no parameter was changed after seeing these results.

Reference beat annotations were read with custom MATLAB routines for the WaveForm DataBase (WFDB) format. Beat types came from the annotation codes: normal, LBBB, RBBB, PVC and paced beats. Non-beat annotations, such as ventricular flutter waves, were not scored as beats. Annotations in the first 2 s (the detector's learning period) and in the last 150 samples (where the filter delay hides beats) were excluded from scoring, and detections in the first 2 s and in the final 150 samples were also discarded, so that detections and annotations were scored over the same interval. In the clean 120 s segments, only one detection (record 105, for both detectors) fell in the final 150 samples.

Baseline detector

The baseline is a simplified Pan-Tompkins-style detector [1], implemented as a fixed-step discrete Simulink model with a sample time of 1/360 s (Fig. 1). The ECG segment is loaded into the model with a From File block, and the signal then passes through four stages: a 201-tap finite impulse response (FIR) band-pass filter (5-15 Hz, designed with fir1); a five-point derivative filter, [1 2 0 −2 −1]·fs/8; squaring; and moving-window integration over 54 samples (150 ms).

A MATLAB Function block then applies an adaptive threshold to the integrated signal. During the first 2 s, the maximum of the integrated signal initializes the signal-peak estimate, and the threshold is set to 0.25 of that maximum. Afterwards, each local maximum is compared with the threshold. A peak above the threshold and more than 200 ms after the previous detection is accepted as a beat and updates the signal-peak estimate (weight 0.125). Other peaks update the noise-peak estimate with the same weight. The threshold is then set to the noise estimate plus 0.25 times the difference between the signal and noise estimates. The detector has no search-

back and no T-wave discrimination. The threshold constants are those of the standard formulation and were not tuned.

The filter chain delays detections by a constant amount. This delay (119 samples, 331 ms) was estimated as the median offset between detections and annotations on record 100 and subtracted once from all baseline detections in all experiments. The refined timestamps are computed in original-signal time by the refinement block (next subsection) and are not shifted again. Because the value was estimated empirically from annotations, it may include record-specific morphology effects, so the result on record 100 is not fully independent of this calibration. The value exceeds the 100-sample group delay of the band-pass filter because it covers the whole processing and decision chain.

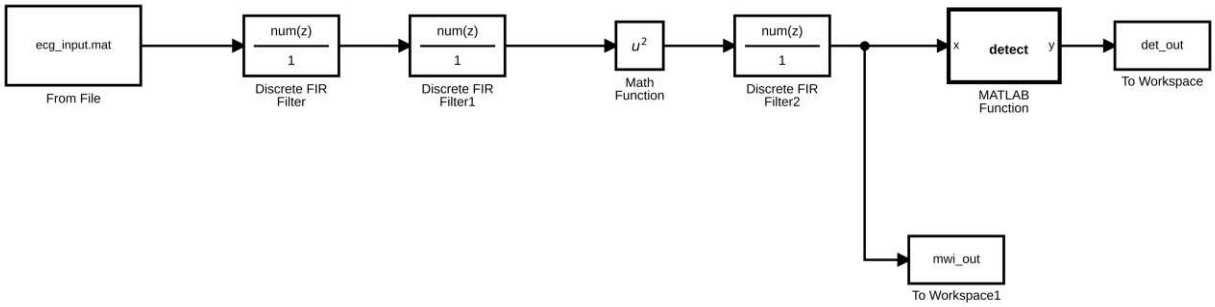


Fig. 1 Simulink model of the baseline detector: From File input block, FIR band-pass (5-15 Hz), five-point derivative, squaring, 150 ms moving-window integration, and a MATLAB Function block that applies the adaptive threshold and 200 ms refractory period; schematic redrawn from the saved model file

Proposed peak-refinement stage

The proposed detector adds a second MATLAB Function block to the model (Fig. 2). The block receives two inputs: the output of the 5-15 Hz band-pass stage and the detection flag from the baseline detector. It keeps the most recent 128 band-passed samples in a circular buffer. When a baseline flag arrives at sample n, the expected R-peak position in original-signal time is n − 119, because the flag is raised about 119 samples after the R peak. The block searches the 37 buffered samples that correspond to original-signal samples n − 137 to n − 101 (the 100-sample delay of the band-pass filter is removed in the index calculation), that is, ±18 samples (about ±50 ms) around n − 119. It assigns the refined timestamp to the sample with the largest absolute band-passed amplitude, and if several samples tie, the earliest one is kept. The block uses only samples that have already arrived, so it is causal, and the refined timestamp is output at the same time as the baseline flag, with no extra latency. Refinement changes only the timestamp of an existing detection; it cannot create or suppress a detection.

The refinement filter band and window length were chosen by a sweep on the tuning set. A 0.5-40 Hz and a 5-15 Hz filter and half-windows of 30, 50, 70 and 100 ms were compared, and 5-15

Hz with a 50 ms half-window was selected. The sweep was run in MATLAB with zero-phase filtering. The Simulink block, which is causal, reproduced the record-level F1 scores of that prototype on the tuning set. It also produced the same number of detections as the prototype on all five tuning records. At the detection level, 613 of the 727 timestamps (84%) were identical to those of the prototype. In records 100, 107, 109 and 119 the maximum difference was 1 sample (2.8 ms). In record 118, 25 timestamps differed by 16 to 18 samples (44 to 50 ms). An offline calculation that applied the same search rule to the causal FIR band-passed signal reproduced all 727 Simulink timestamps exactly, so these differences reflect the filter used (zero-phase Butterworth in the prototype, causal FIR in Simulink) and not an implementation error. In the 25 cases, the two largest deflections inside the search window differed in absolute amplitude by about 6% or less in the prototype signal, and the two filters ranked them differently. In 21 of the 25 cases the Simulink timestamp was 17 samples after the baseline timestamp, one sample inside the late edge of the search window, and in the other four it was within 2 samples of it. These Simulink timestamps lay 17 to 18 samples (47 to 50 ms) from the annotation, at the edge of the ±50 ms tolerance (18 samples), so the ±50 ms F1 of record 118 (99.30%) is sensitive to small changes; at ±75 ms it is 100%. It was not determined which waveform component of the right bundle branch block complex was selected.

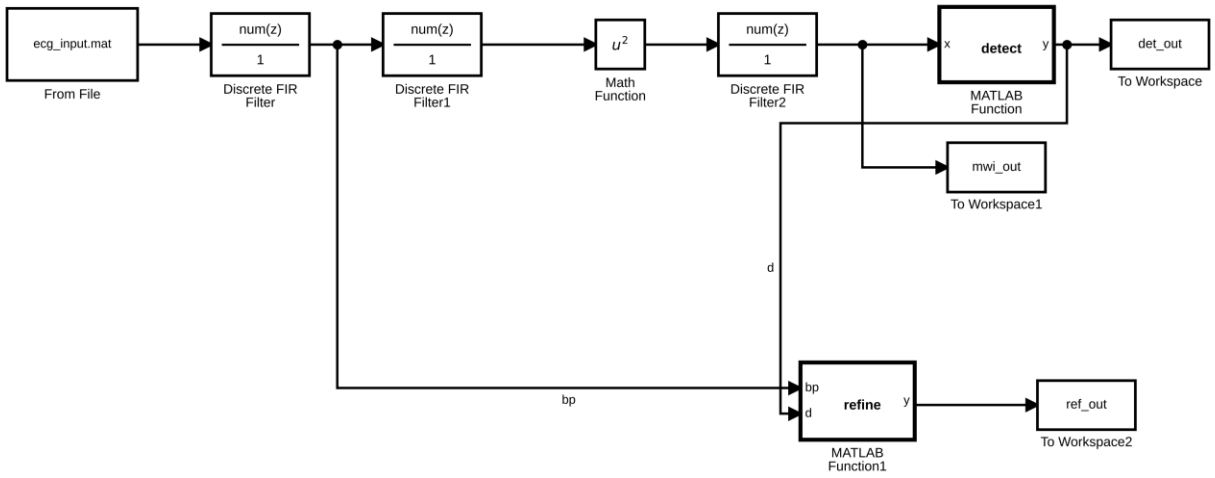


Fig. 2 Simulink model of the proposed detector. A second MATLAB Function block (refine) receives the band-passed signal (bp) and the baseline detection flag (d) and reassigns each detected beat to the largest absolute-amplitude sample within ±50 ms of the delay-corrected expected R-peak position (n − 119); a 128-sample buffer holds the recent band-passed samples; schematic redrawn from the saved model file

Noise protocol

Noise was added using the baseline wander (bw), electrode motion (em) and muscle artifact (ma) recordings of the NSTDB. Each ECG record received a different consecutive 120 s slice of the noise recording: slices 1 to 5 for the five tuning records and slices 6 to 9 for the four held-out records.

For each record, the noise slice n was made zero-mean and scaled by a factor α so that the noisy signal y = x + αn had the target SNR (Eq. 1):

$$\alpha = \sqrt{[\mathrm{var}(x) / (\mathrm{var}(n) \times 10^{SNR/10})]} \quad (1)$$

Here x is the 120 s ECG segment, n is the zero-mean noise slice, and var(·) is the variance over the analyzed 120 s interval. The target SNRs were 24, 12, 6 and 0 dB. The ECG mean was not removed because it does not affect the variance. The achieved SNR, computed from the generated noisy and clean signals, agreed with the target to within 2×10^{-12} dB; this follows from the scaling and is not an independent measurement. No clipping or saturation was applied, and 57 of the 4,665,600 samples in the 108 noisy signals (0.001%) exceeded ±5.12 mV, the nominal range of the original 11-bit recordings. Each condition used one fixed noise slice, so the experiments measure performance under these particular noise segments and do not estimate variability across repeated noise realizations. Both detectors processed identical noisy signals, and both experiments used the same two Simulink models as the clean-data experiments.

Evaluation

A detection was a true positive (TP) if it fell within a tolerance window of an unmatched annotation. Matching was one-to-one. Unmatched detections were false positives (FP) and unmatched annotations were false negatives (FN). Sensitivity (Se = TP/(TP+FN)), positive predictivity (PPV = TP/(TP+FP)) and F1 (the harmonic mean of Se and PPV) are reported. Unless stated otherwise, "mean F1" is the unweighted arithmetic mean of the record-level F1 values, not an F1 computed after pooling TP, FP and FN across records. Beat-by-beat comparison and record-by-record reporting follow the general approach of ANSI/AAMI EC57 [19]. Studies use different tolerance windows. The primary, strict criterion was ±50 ms, because timing matters for applications such as HRV, and a secondary criterion of ±75 ms was also used.

Sensitivity was computed separately for each beat type. A beat counted as detected if any detection lay within ±50 ms of it. PPV cannot be split by beat type because a false positive has no annotated type, so PPV is reported per record only.

For timing precision, each annotated beat matched by both detectors within ±150 ms gave a signed offset (detection minus annotation) in ms, summarized by mean, standard deviation (SD) and mean absolute error (MAE). The ±150 ms matching window is wider than the detection tolerance so that offsets larger than ±50 ms are included. Beats without a detection within ±150 ms for either detector are excluded from this analysis. The primary comparison is the record-level MAE of each detector. The paired record-level differences were summarized by their median, interquartile range (MATLAB percentile method) and range, tested with an exact paired signed-rank test, and given a 95% percentile bootstrap confidence interval for the median (10,000 resamples of the nine records). A beat-level paired Wilcoxon signed-rank test on the absolute offsets was also calculated as supplementary descriptive information only, because beats within one record are correlated and are not independent units.

Implementation and use of AI tools

All experiments used MATLAB R2016a with Simulink 8.7, the Signal Processing Toolbox and the DSP System Toolbox, on Windows 11. Each model run used a fixed-step solver with a step size of 1/360 s (automatic solver selection, which resolves to the discrete solver because the models contain no continuous states).

An AI assistant (Claude, Anthropic) was used to help write the MATLAB and Simulink code, to support the data analysis and literature search, and to draft the text. The authors ran all experiments, checked the results and the references, and reviewed and edited the text, and they take full responsibility for the content. The AI assistant is not an author.

Results

Detection accuracy on the tuning set

Table 1 gives the clean-data results at the ±50 ms tolerance. The baseline detector scored a mean F1 of 97.1%. Records 100 and 107 were scored perfectly, but record 118 (right bundle branch block) reached only 88.1%, with 17 false positives and 17 false negatives. The refined detector raised the mean F1 to 99.9%. Record 118 improved to 99.3% and every other record scored 100%.

Table 1 Clean-data F1 (%) on the tuning set, ±50 ms tolerance

Record	Baseline	Refined
100	100.00	100.00
107	100.00	100.00
109	98.27	100.00
118	88.11	99.30
119	99.22	100.00
Mean	**97.12**	**99.86**

Table 2 gives sensitivity by beat type. The baseline's weakest type was RBBB (89.1%, 122 of 137 beats), followed by PVC (93.5%, 29 of 31) and LBBB (98.2%). The refined detector detected 99.3% of RBBB beats (136 of 137) and all LBBB and PVC beats. The PVC row has only 31 beats and should not be over-interpreted.

Table 2 Sensitivity (%) by beat type, tuning set, ±50 ms

Beat type	Beats	Baseline	Refined
Normal	247	100.00	100.00

Beat type	Beats	Baseline	Refined
LBBB	168	98.21	100.00
RBBB	137	89.05	99.27
PVC	31	93.55	100.00
Paced	137	100.00	100.00

A diagnostic on record 118 showed that no annotated beat lacked a detection within 100 samples (0.28 s). The detection offsets fell into three clusters, at about −47, 0 and +56 ms, and the outer clusters lie at or beyond the ±50 ms tolerance. The loss of accuracy on this record is therefore consistent with timing error rather than the absence of a nearby detection.

Table 3 shows the refinement sweep. With the 0.5-40 Hz filter, F1 on the paced record 107 fell as the window grew (99.3, 81.9, 65.9 and 55.8% for 30, 50, 70 and 100 ms). With the 5-15 Hz filter, F1 on record 107 stayed at 99.3-100% for all windows. A 30 ms window was too short for record 118 (98.6% at 5-15 Hz and 90.2% at 0.5-40 Hz). The 5-15 Hz band with a 50 ms half-window was therefore used, and windows of 70 and 100 ms gave identical results. These settings were chosen on the same records used to evaluate them, so these scores are optimistic. The held-out evaluation addresses this.

Table 3 F1 (%) of the refinement variants on the tuning set (±50 ms tolerance), for half-windows of 30 / 50 / 70 / 100 ms

Record	Baseline	0.5-40 Hz	5-15 Hz
100	100.00	100 / 100 / 100 / 100	100 / 100 / 100 / 100
107	100.00	99.3 / 81.9 / 65.9 / 55.8	99.3 / 100 / 100 / 100
109	98.27	98.3 / 98.3 / 98.3 / 98.3	98.3 / 100 / 100 / 100
118	88.11	90.2 / 98.6 / 98.6 / 98.6	98.6 / 99.3 / 99.3 / 99.3
119	99.22	100 / 100 / 100 / 100	99.2 / 100 / 100 / 100

Held-out evaluation

On records 105, 200, 207 and 217, with all settings fixed in advance, refined F1 was equal to or higher than baseline F1 in each record (Table 4). Mean F1 rose from 86.7% to 95.7%. Nearly all of this gain came from record 217, where the baseline produced 50 false positives and 53 false negatives (F1 63.4%) and the refined detector produced no false positives and 3 false negatives (F1 98.9%). Records 105 and 207 were unchanged and record 200 improved by 0.6 points.

Table 4 Held-out results, ±50 ms tolerance

Record	Baseline TP / FP / FN	Baseline F1 (%)	Refined TP / FP / FN	Refined F1 (%)
105	162 / 0 / 0	100.00	162 / 0 / 0	100.00
200	161 / 5 / 5	96.99	162 / 4 / 4	97.59
207	101 / 25 / 7	86.32	101 / 25 / 7	86.32
217	89 / 50 / 53	63.35	139 / 0 / 3	98.93
Mean		**86.66**		**95.71**

Sensitivity by beat type (Table 5) shows the same pattern. Paced-beat sensitivity rose from 60.3% to 100%, which is consistent with the baseline's problem on paced beats being one of timing. PVC sensitivity was about 90% for both detectors (89.3% vs 90.2%), so refinement did not help there. The RBBB and LBBB counts in this set are small (28 and 44 beats).

Table 5 Sensitivity (%) by beat type on the held-out set (pooled, ±50 ms)

Beat type	Beats	Baseline	Refined
Normal	260	100.00	100.00
LBBB	44	100.00	100.00
RBBB	28	100.00	100.00
PVC	112	89.29	90.18
Paced	126	60.32	100.00

Effect of the tolerance window

Table 6 shows F1 at both tolerances for all nine records. At ±50 ms the mean F1 was 92.5% for the baseline and 98.0% for the refined detector. At ±75 ms the baseline mean rose to 98.2% and the refined mean to 98.4%, so the advantage almost disappeared. The baseline recovered on record 118 (88.1% to 99.3%) and record 217 (63.4% to 98.9%). Record 207 stayed at 86.3% under every setting. The baseline's low scores on records 118 and 217 are therefore mainly localization errors of roughly 50 to 75 ms, not the absence of a nearby detection, and the benefit of refinement depends on how strict the tolerance is.

Table 6 F1 (%) at two tolerances, all nine records

Record	±50 ms Baseline	±50 ms Refined	±75 ms Baseline	±75 ms Refined
100	100.00	100.00	100.00	100.00

Record	±50 ms Baseline	±50 ms Refined	±75 ms Baseline	±75 ms Refined
107	100.00	100.00	100.00	100.00
109	98.27	100.00	100.00	100.00
118	88.11	99.30	99.30	100.00
119	99.22	100.00	100.00	100.00
105	100.00	100.00	100.00	100.00
200	96.99	97.59	99.40	100.00
207	86.32	86.32	86.32	86.32
217	63.35	98.93	98.93	98.93
Mean	**92.47**	**98.02**	**98.22**	**98.36**

Timing precision

Fig. 3 and Table 7 summarize the detection offsets for the 1,295 beats matched by both detectors within ±150 ms. The baseline offsets form several clusters (near −45, 0 and +30 ms), with mean −8.9 ms, SD 30.9 ms and MAE 25.8 ms. The refined offsets form one narrow peak near zero, with mean +4.3 ms, SD 10.1 ms and MAE 6.0 ms. The refined MAE was lower than the baseline MAE in all nine records.

Precision was highest for normal beats (refined SD 1.6 ms) and LBBB beats (2.0 ms). It was lower for paced beats (8.6 ms), PVC beats (13.6 ms) and RBBB beats (20.6 ms). Record 118, the RBBB record, had a refined SD of 22.9 ms. A small secondary cluster near +50 ms is visible in the refined histogram. In record 118, 25 of 143 refined timestamps lay 47 to 50 ms after the annotation (see Materials and Methods), which contributes to the larger SD and the +13.1 ms mean offset of that record. The cluster was not examined in other records. The refined offsets also have a small positive bias (+4.3 ms overall, +9.8 ms in record 107). Part of this may reflect the annotators' convention for placing the beat mark rather than detector error, but this was not tested.

The refined MAE was lower than the baseline MAE in all 9 of 9 records (Table 7). The record-level reductions ranged from 5.3 to 29.6 ms, with a median of 20.8 ms (interquartile range 16.5 to 23.3 ms). A bootstrap over the nine records gave a 95% confidence interval of 12.8 to 28.2 ms for the median reduction, which is wide because only nine records were available. Because all nine differences have the same sign, an exact two-sided paired signed-rank test gives $p = 0.0039$, which is limited evidence given $n = 9$. A beat-level Wilcoxon test gave $p = 4.85 \times 10^{-151}$, but beats within a record are correlated, so this value is supplementary and is not used as the primary inference.

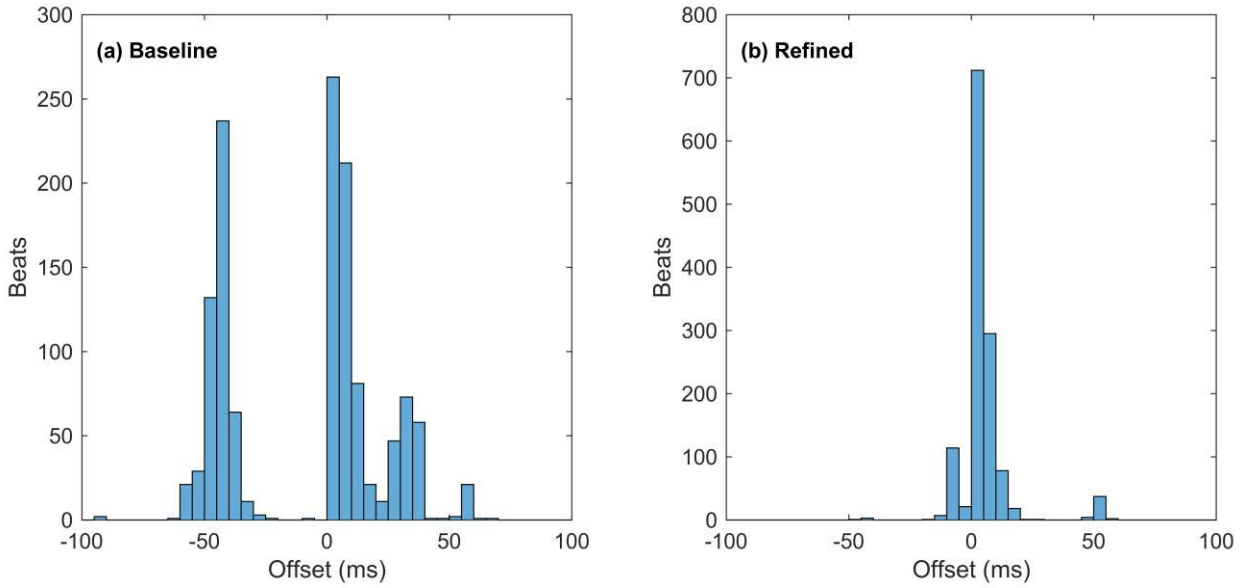


Fig. 3 Detection offset (detection minus annotation, ms) for 1,295 beats matched by both detectors within ±150 ms: (a) baseline and (b) refined detector. Nine records pooled

Table 7 Offset (detection minus annotation, ms) for beats matched by both detectors within ±150 ms

Group	n	Baseline mean / SD / MAE	Refined mean / SD / MAE
100	145	−5.6 / 17.4 / 8.5	3.2 / 1.0 / 3.2
107	138	−3.5 / 31.8 / 31.2	9.8 / 1.9 / 9.8
109	173	−10.3 / 27.4 / 22.7	1.9 / 1.6 / 1.9
118	143	−20.6 / 36.4 / 35.8	13.1 / 22.9 / 14.1
119	128	−22.8 / 29.5 / 34.0	5.5 / 5.8 / 5.8
105	162	−3.8 / 35.9 / 33.9	4.3 / 1.6 / 4.3
200	166	0.9 / 29.6 / 22.5	4.2 / 9.3 / 4.8
207	101	1.3 / 23.5 / 17.1	2.3 / 7.1 / 4.3
217	139	−15.7 / 30.2 / 25.3	−5.8 / 4.1 / 6.5
Normal	507	−9.2 / 30.2 / 24.9	3.2 / 1.6 / 3.2
LBBB	212	−7.4 / 26.3 / 21.0	2.5 / 2.0 / 2.5
RBBB	165	−16.8 / 35.4 / 32.6	11.0 / 20.6 / 11.3

Group	n	Baseline mean / SD / MAE	Refined mean / SD / MAE
PVC	136	1.5 / 28.9 / 22.2	6.3 / 13.6 / 9.5
Paced	263	−10.3 / 32.1 / 29.2	2.0 / 8.6 / 8.5
All beats	**1295**	**−8.9 / 30.9 / 25.8**	**4.3 / 10.1 / 6.0**

Performance under added noise: tuning records

Table 8 and Fig. 4 give the mean F1 over the five tuning records after noise was added, using the Simulink models. Baseline wander had little effect on either detector (baseline 96.8-97.4%, refined 99.7-99.9%). Electrode motion was the most damaging noise type. At 0 dB it reduced the mean F1 to 87.3% for the baseline and 91.7% for the refined detector. Muscle artifact at 0 dB gave 91.0% and 97.2%. At 24 dB the scores were close to the clean-data values (97.1% and 99.9%), so most of the difference between detectors at high SNR is the timing effect already described, not a noise effect. The gains specific to heavy noise were 4.4 points for electrode motion and 6.2 points for muscle artifact at 0 dB. Refinement cannot add detections for beats that have no nearby detection, and the 8.3-point gap to 100% that remains for electrode motion at 0 dB is not a timing effect. Its cause was not diagnosed.

Table 8 Mean F1 (%) over the five tuning records with added noise (baseline → refined)

Noise	24 dB	12 dB	6 dB	0 dB
bw	97.1 → 99.9	97.1 → 99.7	97.4 → 99.7	96.8 → 99.7
em	96.6 → 99.7	95.4 → 98.5	94.5 → 98.0	87.3 → 91.7
ma	96.7 → 99.7	95.5 → 99.4	96.4 → 99.0	91.0 → 97.2

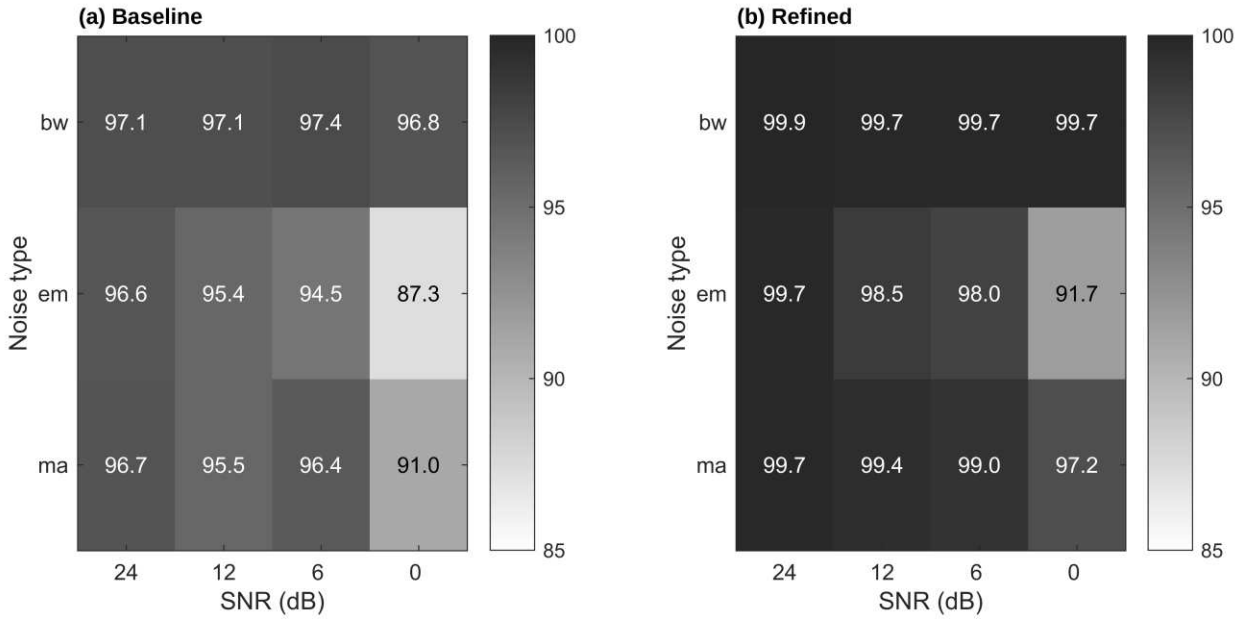


Fig. 4 Mean F1 (%) over the five tuning records with added baseline wander (bw), electrode motion (em) and muscle artifact (ma) noise at 24, 12, 6 and 0 dB SNR: (a) baseline and (b) refined detector. Darker cells show higher F1, and the mean F1 is printed in each cell

Performance under added noise: held-out records

Table 9 repeats the noise experiment on the four held-out records, using noise slices not used for the tuning records. The refined detector scored higher than the baseline in all 12 conditions, by 8.7 to 14.6 points of mean F1. Baseline wander again had little effect. Electrode motion at 0 dB was again the hardest condition (baseline 74.1%, refined 86.6%), followed by muscle artifact at 0 dB (80.3% and 93.4%). Absolute scores are lower than on the tuning records because record 207 (spurious detections during the high-amplitude episode and unmatched PVCs) and record 217 (paced beats, baseline F1 63.4% on clean data) are included. Most of the difference at 24 dB reflects the timing effect seen on clean data, not a noise effect.

Table 9 Mean F1 (%) over the four held-out records with added noise (baseline → refined)

Noise	24 dB	12 dB	6 dB	0 dB
bw	86.5 → 95.7	86.3 → 95.7	86.2 → 95.7	86.0 → 96.0
em	87.0 → 95.7	82.9 → 95.7	80.5 → 95.1	74.1 → 86.6
ma	85.7 → 95.6	85.7 → 95.8	84.7 → 96.1	80.3 → 93.4

Failure analysis: record 207

Record 207 scored 86.3% under both detectors and both tolerances. The refined detector produced 126 detections for 108 scored beats. Of the 25 false positives, none had an annotated

beat within 0.42 s, and 24 fell between 40 and 60 s, where the signal contains a high-amplitude oscillation (Fig. 5). The PhysioNet notes for this record describe it as extremely difficult, with multiform PVCs, left and right bundle branch block patterns, and several episodes of ventricular flutter [18]. The listed points of interest include ventricular flutter starting at 0:33 and continuing past 0:57 [18], which overlaps the observed burst. The burst is therefore regarded as probably a flutter episode, but this was not checked sample by sample. Because flutter waves are non-beat annotations, detections inside the episode count as false positives under the scoring used here.

Of the 7 false negatives, all were PVCs (7 of the 36 PVCs in this segment). Four had no detection within 0.42 s, and three had a detection more than 50 ms away. The unmatched PVCs were spread across the record, not clustered in the noisy segment. Record 207 therefore has two separate problems, spurious detections in a flutter-like burst and unmatched PVCs, and neither can be corrected by refinement.

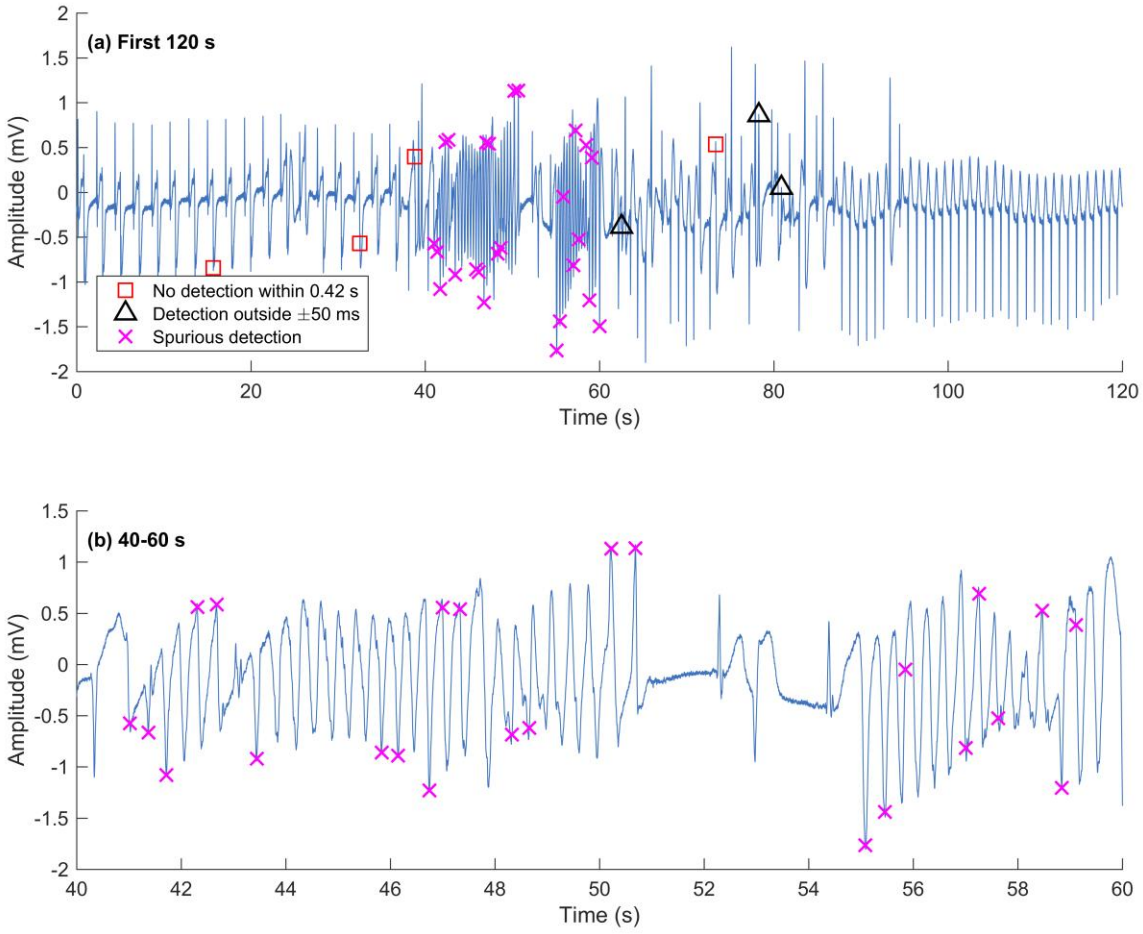


Fig. 5 Record 207 with the refined detector's errors: (a) first 120 s and (b) enlarged view of 40–60 s. Red squares mark annotated beats with no detection within 0.42 s (n = 4), black triangles mark annotated beats whose nearest detection lies outside ±50 ms (n = 3), and magenta crosses mark detections with no annotated beat within ±50 ms (n = 25)

Discussion

Main findings

Except for record 207, the baseline produced a detection near most annotated beats on clean data, but it often localized detections away from the annotated R peak on wide-QRS, paced and bundle-branch beats. Three results support this reading. On record 118 no annotated beat lacked a detection within 100 samples, and the offsets fell into clusters at about −47, 0 and +56 ms. At ±75 ms the baseline's mean F1 rose from 92.5% to 98.2%, so most of its errors under the ±50 ms matching criterion were localization errors of roughly 50 to 75 ms. The baseline's MAE across 1,295 matched beats was 25.8 ms.

The refinement stage cut this MAE to 6.0 ms and raised the ±50 ms mean F1 from 92.5% to 98.0% over nine records. On the held-out records, where no setting was changed after seeing the results, refined F1 was equal to or higher than baseline F1 in each of the four prespecified records. Most of the gain came from record 217 and from paced beats (sensitivity 60.3% to 100%). The refinement is therefore described as improving timing precision. It is not claimed that refinement improves beat detection in the sense of finding more beats. At ±75 ms the two detectors are nearly equal.

Relation to earlier work

The findings agree in direction with earlier studies. Reklewski et al. [10] reported that detection accuracy depends strongly on the temporal tolerance, and the comparison of ±50 ms and ±75 ms in this study shows the same dependence for this detector. Portet et al. [9] showed that detector performance varies with beat morphology and noise type, and the beat-type and noise results of this study are in line with this. The NSTDB documentation describes electrode motion artifact as generally the most troublesome noise type [16], and it was the hardest condition in the present experiments. Gradl et al. [11] proposed temporal correction of detected R-peaks. The refinement stage used here is a simple causal variant of this idea built from Simulink blocks, and it was not compared with their method. What this study adds is a documented Simulink implementation, an offset analysis that separates timing errors from the absence of a nearby detection, and a held-out test, all for one detector and nine short records.

Why the evaluation criterion matters

A single F1 at a single tolerance would have told two different stories. Under ±50 ms the baseline looks poor on records 118 and 217. Under ±75 ms it looks nearly perfect. Papers that report only one tolerance, or only a database-wide figure, cannot show which kind of error a detector makes. It is suggested that detector evaluations report at least two tolerances and, where timing matters, a signed offset in milliseconds. The right tolerance depends on the application. For heart rate alone, 75 ms may be acceptable. For HRV analysis, where R-peak localization affects the results [8], timing errors of 25 ms (the baseline MAE) may be too large, though the effect on HRV was not measured directly.

Interpreting the refinement

With the wide 0.5-40 Hz band, F1 on the paced record 107 fell as the search window grew, while with 5-15 Hz it stayed near 100%. A plausible explanation is that the wider band keeps large deflections other than the R peak, such as pacing artifacts or T waves, which the largest-absolute-value rule can select. This was not verified by inspecting individual beats, so it remains a hypothesis. It does suggest that the choice of search signal matters more than the window length.

The refinement is a fixed rule applied to every beat. It does not adapt to beat morphology or noise level, and the stratified results show where a fixed rule is not enough. Its timing was tight for normal and LBBB beats (SD 1.6 and 2.0 ms) but looser for paced (8.6 ms), PVC (13.6 ms) and RBBB beats (20.6 ms). The small secondary cluster near +50 ms in the refined histogram includes 25 beats of the right bundle branch block record 118 whose refined timestamps lay 47 to 50 ms after the annotation, but the cluster was not traced in other records.

Limits of refinement

Refinement moves detections. It cannot add a detection where none exists or remove a false one. In record 207 the 25 spurious detections lie in a high-amplitude burst (40 to 60 s) that overlaps a documented ventricular flutter episode [18], and the 7 unmatched annotations are all PVCs. A signal-quality check that suppresses detections in such segments, and a threshold rule that handles PVCs better, are possible remedies that were not tested. Electrode-motion noise at 0 dB also remained difficult (mean F1 91.7% on the tuning records and 86.6% on the held-out records). PVC sensitivity was about 90% for both detectors on the held-out set, the only beat type for which neither detector reached 95%.

Limitations and future work

The study has several limitations. First, nine records and 120 s of each were used, which is too little to claim that the results hold across the database or across patients, and beat-type counts in the held-out set are small (28 RBBB, 44 LBBB). Second, the refinement settings were chosen on the five tuning records, using a zero-phase Butterworth prototype. The causal FIR implementation reported here selected a different timestamp for 114 of 727 detections (up to 18 samples in record 118), so the settings were not tuned on the exact filter that is evaluated. The clean-data and noise experiments were repeated on four held-out records, which reduces but does not remove this concern. Third, one simplified Pan-Tompkins-style detector was evaluated, with standard threshold constants and without search-back or T-wave discrimination. A simplified Hamilton-Tompkins-style detector was also built in Simulink, but it produced about two detections per beat on most records, apparently because of a defect in its threshold logic that was not resolved, and its results are not reported. The findings therefore describe one detector and do not show that the timing pattern holds across classical detectors. Fourth, the 119-sample delay was estimated empirically from the annotations of record 100 and applied to all records, so performance on record 100 is not fully independent of this calibration. An analytically derived delay or a separate calibration segment would give a stronger evaluation.

Fifth, each record received a different noise slice, so noise type is partly confounded with record. Sixth, the beat-level Wilcoxon test ignores within-record dependence, and with nine records (four held-out) the statistical uncertainty is substantial. Any confidence interval should resample records, not beats. Seventh, the timing analysis includes only beats matched by both detectors within ±150 ms. Finally, the positive mean offset of the refined detector (+4.3 ms) may partly reflect annotator conventions, and in record 207 detections inside the flutter episode are scored as false positives because flutter waves are not beat annotations. Evaluation protocols may treat such episodes differently [19], so the F1 of 86.3% may understate performance under other conventions. The record was not rescored with the episode excluded.

Future work should add further classical detectors (for example a verified Hamilton-Tompkins implementation and a wavelet-based detector), use full-length records and more of the database with a statistical test across records, make the search window depend on measured QRS width, add a signal-quality index for artifact-heavy segments, and measure the effect of the timing errors on HRV metrics directly.

A Simulink implementation of a simplified Pan-Tompkins-style R-peak detector was evaluated by beat morphology, added-noise condition and matching tolerance. Under the ±50 ms criterion, many baseline errors on right bundle branch block and paced beats reflected localization error rather than the absence of a nearby detection. The causal peak-refinement stage reduced the mean absolute timing error from 25.8 ms to 6.0 ms and increased the unweighted mean record-level F1 from 92.5% to 98.0% across nine 120 s recordings. Across the four prespecified held-out records, refined F1 was equal to or higher than baseline F1. The advantage was small under the ±75 ms criterion, indicating that the principal benefit was improved timestamp precision rather than the recovery of beats with no nearby detection. Refinement did not correct the spurious detections during the high-amplitude episode in record 207, the severe electrode-motion artifact at 0 dB, or the unmatched PVCs. Because the evaluation used nine short recordings, one simplified baseline detector and one fixed noise slice per record and noise type, the findings characterize this implementation and should not be interpreted as a ranking of R-peak detectors.

Acknowledgments

The ECG recordings and annotations used here were obtained from PhysioNet. The authors thank the creators and maintainers of the MIT-BIH Arrhythmia Database and the MIT-BIH Noise Stress Test Database.

Statements and Declarations

Funding No funding was received for conducting this study.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics approval This study used only publicly available, de-identified recordings from the MIT-BIH Arrhythmia Database and the MIT-BIH Noise Stress Test Database [14-17]. No new data were collected from human participants, and institutional ethics approval was not sought.

Data availability The ECG data are publicly available from PhysioNet (https://physionet.org/content/mitdb/1.0.0/ and https://physionet.org/content/nstdb/1.0.0/). The derived results are available from the corresponding author on reasonable request.

Code availability The MATLAB scripts and Simulink models are available from the corresponding author on reasonable request.

Author contributions Akib Ikbal: conceptualization, supervision, writing (review and editing). Sadekatul Shihab: methodology, software, investigation, formal analysis, writing (original draft). Both authors read and approved the final manuscript.

References

1. Pan J, Tompkins WJ. A real-time QRS detection algorithm. IEEE Trans Biomed Eng. 1985;BME-32:230–6. https://doi.org/10.1109/TBME.1985.325532
2. Hamilton PS, Tompkins WJ. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE Trans Biomed Eng. 1986;BME-33:1157–65. https://doi.org/10.1109/TBME.1986.325695
3. Köhler B-U, Hennig C, Orglmeister R. The principles of software QRS detection. IEEE Eng Med Biol Mag. 2002;21:42–57. https://doi.org/10.1109/51.993193
4. Martínez JP, Almeida R, Olmos S, Rocha AP, Laguna P. A wavelet-based ECG delineator: evaluation on standard databases. IEEE Trans Biomed Eng. 2004;51:570–81. https://doi.org/10.1109/TBME.2003.821031
5. Merah M, Taleb-Ahmed A, Beklaouz HL. R-peaks detection based on stationary wavelet transform. Comput Methods Programs Biomed. 2015;121:149–60. https://doi.org/10.1016/j.cmpb.2015.06.003
6. Kalidas V, Tamil L. Real-time QRS detector using stationary wavelet transform for automated ECG analysis. In: Proceedings of the 17th IEEE International Conference on Bioinformatics and Bioengineering (BIBE). 2017. pp. 457–61. https://doi.org/10.1109/BIBE.2017.00-12
7. Friesen GM, Jannett TC, Jadallah MA, Yates SL, Quint SR, Nagle HT. A comparison of the noise sensitivity of nine QRS detection algorithms. IEEE Trans Biomed Eng. 1990;37:85–98. https://doi.org/10.1109/10.43620
8. Ellis RJ, Zhu B, Koenig J, Thayer JF, Wang Y. A careful look at ECG sampling frequency and R-peak interpolation on short-term measures of heart rate variability. Physiol Meas. 2015;36:1827–52. https://doi.org/10.1088/0967-3334/36/9/1827
9. Portet F, Hernández AI, Carrault G. Evaluation of real-time QRS detection algorithms in variable contexts. Med Biol Eng Comput. 2005;43:379–85. https://doi.org/10.1007/BF02345816

10. Reklewski W, Miśkowicz M, Augustyniak P. QRS detector performance evaluation aware of temporal accuracy and presence of noise. Sensors. 2024;24:1698. https://doi.org/10.3390/s24051698
11. 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: Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2015. pp. 522–5.
12. Zahid MU, Kiranyaz S, Ince T, Devecioglu OC, Chowdhury MEH, Khandakar A, et al. Robust R-peak detection in low-quality Holter ECGs using 1D convolutional neural network. IEEE Trans Biomed Eng. 2022;69:119–28. https://doi.org/10.1109/TBME.2021.3088218
13. Yun D, Lee H-C, Jung C-W, Kwon S, Lee S-R, Kim K, et al. Robust R-peak detection in an electrocardiogram with stationary wavelet transformation and separable convolution. Sci Rep. 2022;12:19638. https://doi.org/10.1038/s41598-022-19495-9
14. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag. 2001;20:45–50. https://doi.org/10.1109/51.932724
15. Moody GB, Muldrow WE, Mark RG. A noise stress test for arrhythmia detectors. Comput Cardiol. 1984;11:381–4.
16. Moody G, Mark R. MIT-BIH noise stress test database (version 1.0.0). PhysioNet. 1999. https://physionet.org/content/nstdb/1.0.0/. Accessed 3 Nov 2026.
17. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101:e215–20. https://doi.org/10.1161/01.CIR.101.23.e215
18. Moody GB. MIT-BIH arrhythmia database directory (records). PhysioNet; revised 2010. https://physionet.org/content/mitdb/1.0.0/mitdbdir/records.htm. Accessed 3 Nov 2026.
19. ANSI/AAMI EC57:2012 (R2020). Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms. Arlington: Association for the Advancement of Medical Instrumentation; 2012.