Assessment of Automatic Strategies for Combining QRS Detections by Multiple Algorithms in Multiple Leads

Mariano Llamedo [1], Juan Pablo Martínez [2]

[1] GIBIO, Electronics Department, National Technological University, Medrano 951, C1179AAQ, Buenos Aires, Argentina

[2] BSICoS Group, Institute of Engineering Research, IIS Aragón, University of Zaragoza, Spain

E-mail: llamedom@frba.utn.edu.ar

Abstract. *Objective*: To develop and evaluate an algorithm for the selection of the best performing QRS detections from multiple algorithms and ECG leads. *Approach*: The detections produced by several publicly available single-lead QRS detectors are segmented in 20 seconds consecutive windows. Then a statistical model is trained to estimate a quality metric that is used to rank each 20-s segment of detections. The model describes each heartbeat in terms of 6 features calculated from the RR interval series, and one feature proportional to the number of heartbeats detected in other leads in a neighborhood of the current heartbeat. With the highest ranked segments, we defined several lead selection strategies (LSS) that were evaluated in a set of 1754 ECG recordings, from 14 ECG databases. The LSS proposed were compared with simple strategies such as selecting lead II or the first lead available in a recording. The performance was calculated in terms of the average sensitivity, positive predictive value and the F score. *Main results*: The best performing LSS, based on *wavedet* algorithm, achieved an F score of 98.7, with sensitivity $S = 99.2$ and positive predictive value $P = 98.3$. The F score for the simpler strategy using the same algorithm was 92.7. The LSS studied in this work have been made available in an open source toolbox to ease the reproducibility and result comparison. *Significance*: The results suggest that the use of LSS is convenient in order to select the best heartbeat locations among those provided by different detectors in different leads, obtaining better results than any of the algorithms individually.

Keywords: ECG, QRS detection, lead selection, multilead, multi-algorithm, pattern recognition.

1. Introduction

The improvement of automatic ECG analysis methods may produce an improvement in diagnosis of cardiovascular diseases, which are currently the first single cause of death globally (World Health Organization 2012). In particular, heartbeat detection (or QRS detection) is of great importance, since it is one of the first steps in any automated ECG analysis, and possible errors in it may limit seriously the performance of subsequent algorithms.

Automatic detection of heartbeats in the ECG has been extensively studied in the last decades. Several approaches were studied: based on digital filters (Pan & Tompkins 1985) or the first derivative of the ECG signal (Arzeno et al. 2006). In (Kohler et al. 2002), several detectors based on digital filters, wavelet transform and artificial neural networks were compared. Other methods based on Hilbert (Benitez et al. 2001), curve length (Paoletti & Marchesi 2006) and wavelet transform (Martínez et al. 2004) were evaluated. As a result, several algorithms were made publicly available such as (Pan & Tompkins 1985), the three Physionet's detectors: *sqrs*, *wqrs* and *gqrs*, (Zong et al. 2003, Goldberger et al. 2000), or the *wavedet* algorithm (Almeida et al. 2009) developed in our group, among many others. All of these algorithms and many others reported sensitivities (S) and positive predictive values (P) above 90% in public databases, as is shown in the extensive review of (Elgendi 2013). As discussed in (Elgendi 2013), most of the algorithms reviewed have been trained and evaluated in a single database, the MIT-BIH Arrhythmia database (mitdb) (Goldberger et al. 2000), causing a well-known optimistic bias in the performance estimation due to overfitting. When the algorithms are applied to ECG signals from other settings, other types of patients, or from other recording devices, the performance inevitably drops with respect to the reported results. In the last decade some of the QRS detectors were evaluated in a different database following a machine learning approach (Mehta & Lingayat 2008), (Mondelo et al. 2017) and (Ledezma & Altuve 2019) by including other public databases such as the INCART database (Goldberger et al. 2000). This important methodological change started to describe the generalization capability of the QRS detectors.

Although multichannel or multilead QRS detection is a well developed methodology (Almeida et al. 2009), very few approaches for lead selection have been studied, in contrast with the single-lead counterpart. Moreover, mixing detections from several single-lead algorithms to improve performance has been scarcely studied and the few articles found are very recent (Mondelo et al. 2017), (Ledezma & Altuve 2019). Other approaches that focus on quantifying the signal quality in order to decide if the signal is too noisy to be processed. A challenge on ECG signal quality estimation was organized in 2011, as a result several works can be found cited in (Clifford et al. 2012), and more recently in a review of that subject (Satija et al. 2018).

In a previous work, we showed that QRS detection performance is highly lead-dependent, and that an important performance improvement can be achieved by selecting the best performing lead (Llamedo & Martínez 2014). Moreover, we also presented a quality metric able to select the best lead in 70% of the recordings, and one of the three best leads in 93% of the cases (Llamedo et al. 2014), resulting in a promising strategy to perform lead selection. In this work, we extend this strategy by concatenating detection segments obtained from several QRS detectors and ECG leads. As a result, a new set of QRS detections is created from parts of the outputs of different detectors and leads, resulting in a multilead and multi-algorithm series. This study aims to analyze the detection improvement achieved by this methodology with respect to the baseline method consisting on always selecting lead II.

The rest of the article is organized as follows: in the next section we present the evaluation dataset used, the algorithms for QRS detection, a quality metric and the lead selection strategies to be evaluated. In the following section the results achieved are presented. Finally the discussion of the results, together with some final remarks conclude this article.

2. Materials and Methods

2.1. ECG databases

The evaluation or test set is composed of 14 ECG databases, grouped into 5 categories: normal sinus rhythm (NSR), arrhythmia (AR), ST and T morphology changes (STT), stress-test (STR) and long-term (LT). Of the 14 databases used, 12 are publicly available online at (Goldberger et al. 2000) or (Couderc n.d.), while ahadb is distributed by ECRI institute (AHA 2010) and biosigna is distributed by Biosigna GmbH (Fischer et al. 2008). All the databases have expert-reviewed QRS complex annotations, used as gold-standard for performance evaluation. The most relevant details of the databases are summarized in Table 1. Overall, the evaluation set includes 1754 recordings with different SNR, arrhythmias, rhythms and morphology changes, representing an exhaustive set of evaluation for QRS detectors.

2.2. QRS detectors

The QRS detectors considered are listed in Table 2. They are representative of the state of the art, and most of them are publicly available (Goldberger et al. 2000, Demski & Soria 2016), including *wavedet* (Martínez et al. 2004), developed and used extensively by our group, and *aristotle* (Moody & Mark 1982) being the only non-public algorithm considered. All the algorithms in Table 2 were used with the pre-processing indicated in their reference publications.

The Pan and Tompkins algorithm (PT) is probably the most popular of the evaluated detectors. It is based on the use of digital filters to band-limit QRS complex energy, followed by a squaring nonlinearity and then by a low-pass filter to enhance the remaining QRS complex energy. The resulting signal is analyzed with several thresholds and a rule-based algorithm in order to perform the QRS detection (Pan & Tompkins 1985). In this work we used the open source implementation of PT available in toolboxes (Sameni 2006, Demski & Soria 2016). The algorithms from EP limited (EP1/2) are also based on the original PT algorithm, but with some enhancements described in (Hamilton & Tompkins 1986). The *wavedet* algorithm (WD) uses the wavelet transform with the derivative of a smoothing function as prototype wavelet, to enhance the energy from the QRS complexes in specific locations of the time-scale plane. The detection is performed by a set of threshold-based rules across several

Table 1. Databases included in the evaluation set

group	name	f_S (Hz)	# rec	leads	length	# beats	ref
NSR	nsrdb	128	18	2	1 day	1785791	Physionet
	fantasia	250	40	1	2 h	39411	Physionet
AR	ahadb	250	155	2	30 m	348514	AHA
	biosigna	500	50	12	1 h	290149	Biosigna
	mitdb	360	48	2	30 m	100718	Physionet
	svdb	128	78	2	30 m	184502	Physionet
	incartdb	257	75	12	30 m	175893	Physionet
STT	edb	250	90	2	2 h	790554	Physionet
	ltstdb	250	86	2-3	21-24 h	9201221	Physionet
STR	thew15	1000	909	12	15 m	1653250	THEW
	stdb	360	28	2	10-40 m	76175	Physionet
LT	ltdb	128	7	2	14-22 h	691505	Physionet
	nsrdb	128	18	2	1 day	1785791	Physionet
	ltstdb	250	86	2-3	21-24 h	9201221	Physionet
	ltafdb	128	84	2	1 day	9290757	Physionet
Total			1754			24628440	

Physionet: (Goldberger et al. 2000), AHA: (AHA 2010)

THEW: (Couderc n.d.), Biosigna: (Fischer et al. 2008)

Table 2. Algorithms description

name	evaluated in	multilead	ref
WD	mitdb, edb, CSE, qtdb	yes	(Martínez et al. 2004)
GQ, SQ, WQ	mitdb	no	(Goldberger et al. 2000)
PT	mitdb	no	(Pan & Tompkins 1985)
EP1/2	mitdb	no	(Hamilton & Tompkins 1986)
AR	mitdb	yes	(Moody & Mark 1982)

wavelet scales (Martínez et al. 2004). The *wqrs* detector (WQ) uses a similar threshold approach, but using a non-linear transform known as the length transform to enhance QRS complex energy (Zong et al. 2003). The main difference of the *gqrs* detector (GQ), unpublished but open source in (Goldberger et al. 2000), is the use of digital filters (lowpass followed by a matched filter) for QRS complex enhancing. The *sqrs* (SQ) and *aristotle* (AR) detectors were also used for comparative purposes, and described in (Goldberger et al. 2000, Moody & Mark 1982). Following the recommendation of the detectors included in the Physionet's WFDB toolbox, we downsampled ECG recordings when the sampling rate $f_S > 260$ Hz.

2.3. Estimation of the QRS detection quality

In this section, we describe the metric proposed to estimate the quality of QRS detection performed by a given detector at a given lead. Figure 1 can be used to follow the explanation presented in this section. The algorithm takes as input the QRS locations produced by the D available detectors at all the L available leads:

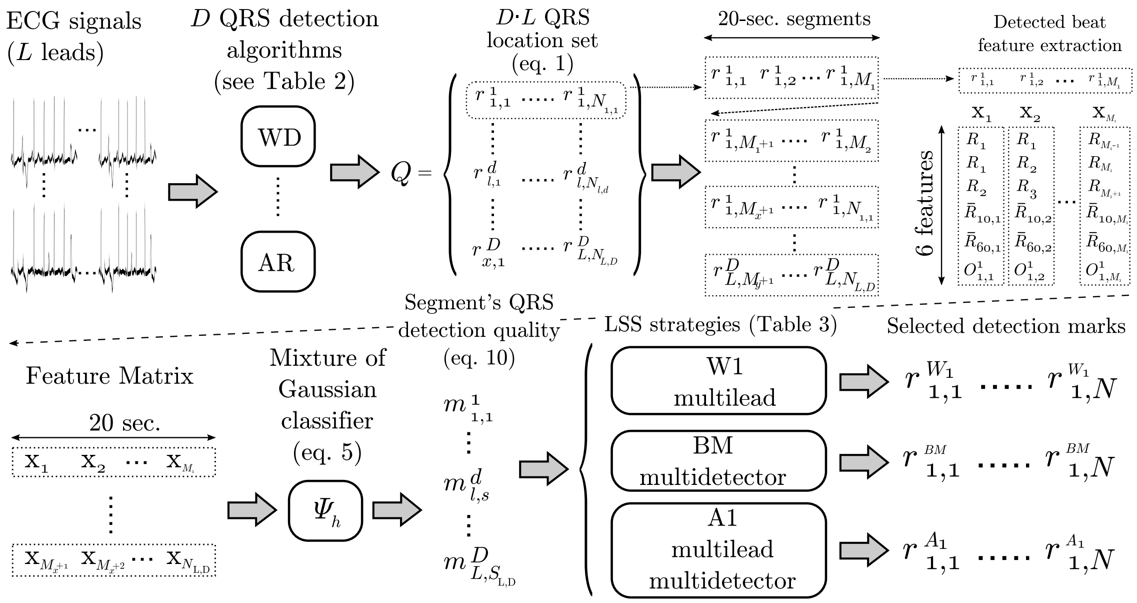

Figure 1. Block diagram of the presented algorithm.

$$Q = \left\{\mathbf{r}_1^1,\, \mathbf{r}_2^1, \cdots,\, \mathbf{r}_L^1, \mathbf{r}_1^2,\, \mathbf{r}_2^2, \cdots,\, \mathbf{r}_L^2, \cdots,\, \mathbf{r}_L^D\right\}, \tag{1}$$

where $\mathbf{r}_l^d$ is the set of detections obtained from detector $d = 1, \ldots, D$ at lead $l = 1, \ldots, L$. Where D is the number of detectors used, in this work we use $D = 1$ for single-algorithm approaches and $D > 1$ for multi-algorithm strategies, this will be explained in the following section. Note that the number of detections in each $\mathbf{r}_l^d$ element is usually different, and is denoted as $N_{l,d}$,

$$\mathbf{r}_l^d = \left\{r_{l,1}^d,\, r_{l,2}^d, , \cdots,\, r_{l,N_{l,d}}^d\right\}, \tag{2}$$

because each detector operates lead-by-lead, as illustrated in Figure 2 for a single recording and detector. All the detectors were used with the default parameter configuration for all the experiments performed in this work.

First, the recording is divided into 20-second segments. A 6-dimensional feature vector $\mathbf{x}_n$ is built for each detected beat $r_{l,n}^d$ (i.e., the *n-th* heartbeat detected in *l-th* lead for the *d-th* detector). Five of these features are related to the RR interval series for that detector and lead, specifically:

1) $R_{n-1} = r_{l,n-1}^d - r_{l,n-2}^d$, the previous RR interval,
2) $R_n = r_{l,n}^d - r_{l,n-1}^d$, the current RR interval and

3) $R_{n+1} = r^d_{l,n+1} - r^d_{l,n}$, the next RR interval,

as is shown in Figure 2. Note that the lead and detector dependence is intentionally omitted. The following two features are the RR baseline estimation with:

4) $\bar{R}_{10}$: the mean RR interval in the past 10 seconds,
5) $\bar{R}_{60}$: the mean RR interval in the past 60 seconds.

The remaining feature is related the the concordance of a given detector across leads

6) $O^d_{l,n}$: the number of heartbeats that co-occur in a 150 ms window of $r^d_{l,n}$ in all available leads.

For example, in a multilead ECG signal with L leads, the heartbeat $r^d_{l,n}$ has an associated value of $O^d_{l,n} = L - 1$ if this heartbeat was also detected in all the other leads, as shown in Figure 2. We consider a detection mark $r^d_{l,n}$ as co-occurrent with respect to $r^d_{x,y}$ if they are closer than 150 ms, following the EC38 standard. Note that erroneously detected beats are likely to have small values of $O^d_{l,n}$, as shown in Figure 2. In order to use the feature O_n with an arbitrary amount of leads, the range of values $[0 : L - 1]$ should be mapped to a fixed range [0:c]. This can be done through a sigmoid function, resulting in

$$O'_n = \frac{c}{1 + e^{-(a \cdot O_n + b)}}, \tag{3}$$

with parameters $a = \frac{18}{L-1}$, $b = -\frac{2a}{L-1}$ and $c = 1000$ that were previously set in (Llamedo et al. 2014). The interested reader is referred to the online implementation for further details: https://github.com/marianux/ecg-kit/blob/master/common/calc_co_ocurrences.m. With these 6 features we build a vector

$$\mathbf{x}_n = (R_{n-1},\, R_n,\, R_{n-1},\, \bar{R}_{10},\, \bar{R}_{60},\, O'_{n,l}). \tag{4}$$

computed for each $r^d_{l,n}$, as can be seen in Figure 1.

The algorithm follows with the computation of a quality metric $m^d_{l,s}$ for each 20-second segment s, as is shown in Figure 3.

In order to compute $m^d_{l,s}$, we used a three-class statistical model to estimate the probability that each heartbeat is a correctly detected beat or true positive (TP if distance between annotation and detection mark <150 ms), a false positive detection (FP), or a detected beat after false negative (FN), as proposed in (Llamedo et al. 2014). Thus each heartbeat will be adjudicated to the class (TP, FP or FN) with maximum *a posterior* probability.

The different rhythms and morphologies present in the data corpus described in Table 1, make the probability distribution $f(\mathbf{x})$ not likely to be Gaussian. Therefore, we adopted a model based on the mixture of several Gaussians (van der Heijden et al. 2005) (MoG) in order to achieve a more adequate fit to the data distribution. In a previous work (Llamedo et al. 2014), the parameters of the statistical model were estimated (trained) with the output of the WD algorithm in a small subset of the first 20 (out of

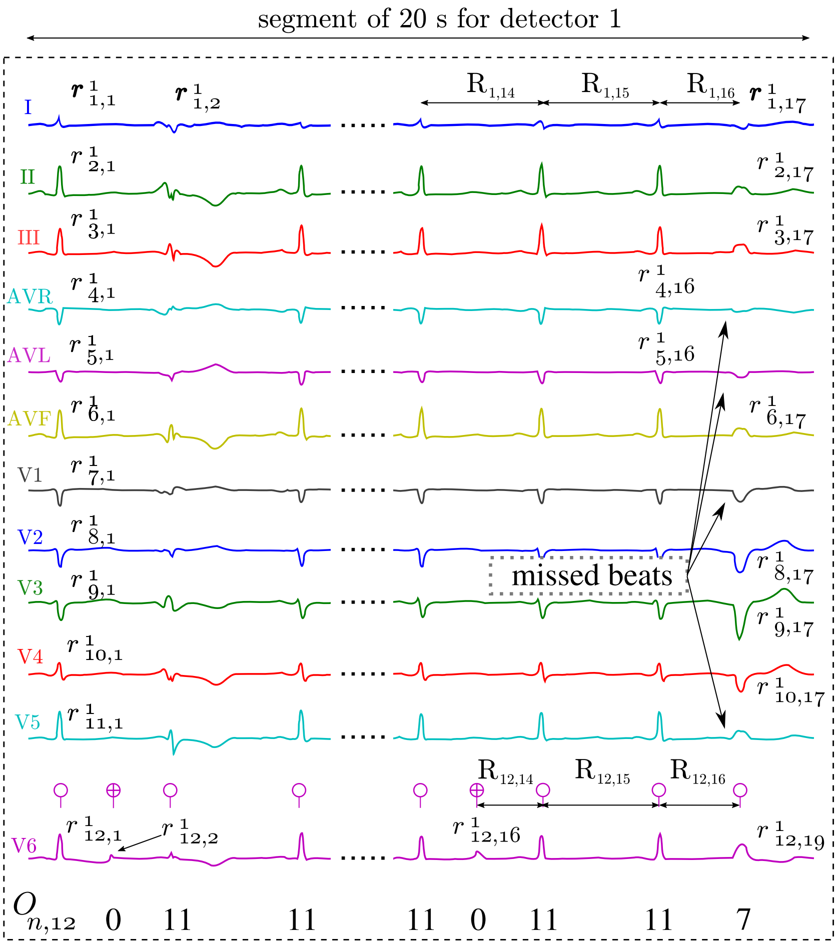

Figure 2. Example showing some feature values calculated for detector 1 in a 20-second segment. RR intervals from single lead QRS detections are shown in leads I and V6. The value of the co-occurrence feature $O_{n,l}$ is shown for lead V6 (l = 12). Note that values of 1 and 8 indicate the presence of a FP and a FN in lead V6.

909) recordings of thew15 database. Those parameters were not changed in this work. For the training of the model, the detection following a missed heartbeat was labelled as FN. When the algorithm is operating, a FN label means that the algorithm predicts that (at least) a beat has been missed between the previous detection and the current one. In the training phase of the classifier, the parameters of the density function

$$p(\mathbf{x}|\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_\mathrm{k}) = \sum_{k=1}^{K} \pi_k \cdot f(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_\mathrm{k}), \tag{5}$$

needs to be estimated where

$$f(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_\mathrm{k}) = \frac{1}{\sqrt{(2\pi)^6\,|\boldsymbol{\Sigma}_k|}}\, \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)} \tag{6}$$

which models the 6-dimensional feature vector $\mathbf{x}_n$ as the sum of K Gaussians with mixing coefficients π_k, in order to retain a more realistic structure of the data. For each class in h = {1, 2, 3} (TP, FP and FN respectively), a MoG parameter set

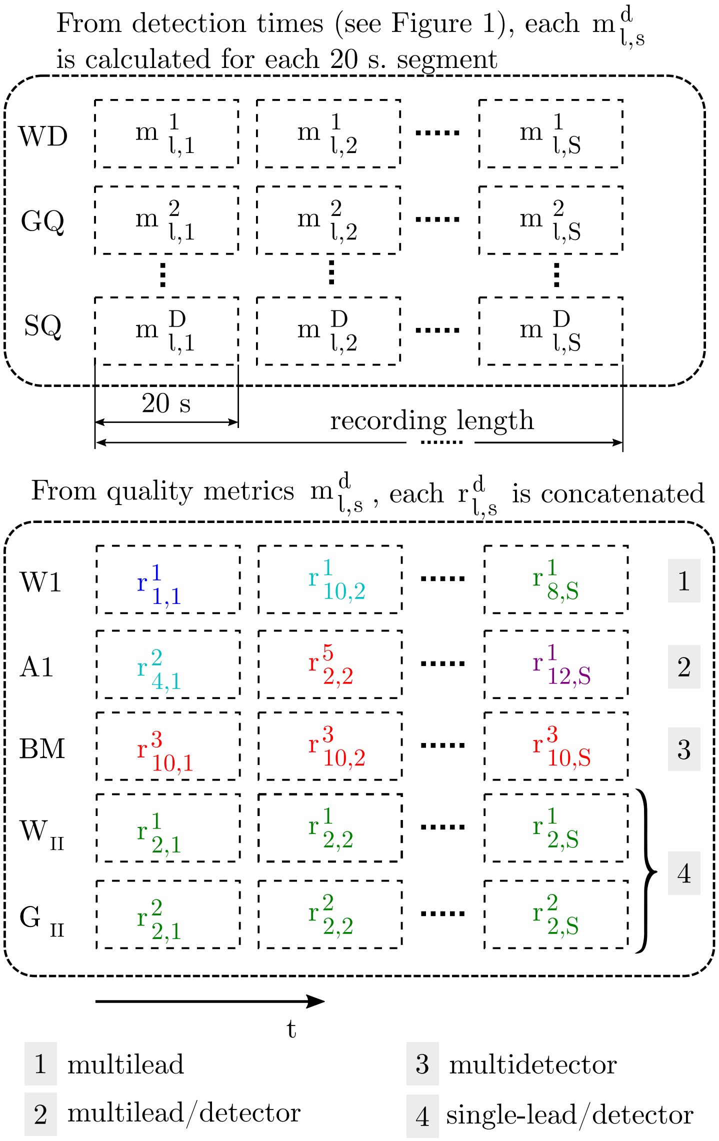

Figure 3. Scheme of the generation of the five LSS (bottom panel) from the quality metrics for different leads and detection algorithms (top). Each algorithm produces a set of detections $r^d_{l,n}$ (Fig. 2) from which a quality metric $m^d_{l,s}$ is calculated (Eq. 10) for each 20 s segment. For example, the LSS W_1 selects the 20-s segments from $r^1_{l,n}$ which achieves higher $m^1_{l,s}$. Note that W_1 is a multilead strategy, since always select from WD algorithm. In contrast, A_1 is a multilead and multi-algorithm LSS, while BM, a multi-algorithm LSS (colour indicates lead as in Fig. 2), selects a whole set of detections $r^d_{l,n}$ which achieves greater $\bar{m}^d_l$ (Eq. 11).

$\Psi_h = \{\pi_{k,h},\, \boldsymbol{\mu}_{k,h},\, \boldsymbol{\Sigma}_{k,h} | k = 1, \ldots, K\}$ is estimated by the maximum likelihood criterion, maximizing the log likelihood

$$L(\mathbf{x}_1, \ldots, \mathbf{x}_{N_h} \| \Psi_h) = \ln \prod_{n=1}^{N} p(\mathbf{x}_h | \Psi_n), \qquad (7)$$

for the N_h heartbeats of h class in the training set. The expectation maximization algorithm was used for this task (Bishop 2006).

Once the model parameters $\mathbf{\Psi}_h$ are estimated, the classifier assigns one label to the heartbeat $\mathbf{x}_n$ which results in the maximum *a posteriori* probability $p(\Psi_h|\mathbf{x}_n)$. Those labels are then used to estimate the quality of each segment s. The number of heartbeats adjudicated to classes TP, FP and FN during a segment s are denoted as N_{TP}, N_{FP}, N_{FN} respectively. Note that we omitted the detector, lead and segment dependence for simplicity. With the three amount of heartbeats per class, the estimated sensitivity can be calculated as

$$\hat{s}^d_{l,s} = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{8}$$

while the estimated positive predictive value is

$$\hat{p}^d_{l,s} = \frac{N_{TP}}{N_{TP} + N_{FP}}. \tag{9}$$

With these two estimates, the detection quality metric is calculated for each segment, lead and detector

$$m^d_{l,s} = \hat{s}^d_{l,s} \cdot w + \hat{p}^d_{l,s} \cdot (1 - w). \tag{10}$$

The weight parameter $w = 2/3$ was previously set for weighting twice $\hat{s}$ in order to favor detection of as many heartbeats as possible at the expense of more false positive detections. With the quality metric already calculated for each segment, the algorithm continues with the segment selection, as is shown in Figures 1 and 3.

2.4. Lead selection strategies

The proposed lead selection strategies (LSS) start with the detections $r^d_{l,s}$ produced by a set of QRS detection algorithms (see Table 2). Following the methodology described above, the quality metric $m^d_{l,s}$ (Eq. 10) can be computed for each lead l, detector d and signal segment s, as is shown in the top panel of Figure 3. The following LSS are then defined:

- W_1 selects, for each segment s, the $r^W_{l,s}$ corresponding to the lead l with best metric $m^W_{l,s}$ when applying the WD algorithm. All the selected sets of detections $r^W_{l,s}$ are then concatenated. It is therefore a multilead single-algorithm LSS.
- A_1 selects, for each segment s, the $r^d_{l,s}$ corresponding to the lead l and the algorithm d with best metric $m^d_{l,s}$ among *all* detectors in Table 2. All the selected sets of detections $r^d_{l,s}$ are then concatenated. This is therefore a multilead multi-algorithm LSS.
- $\mathrm{B_m}$ selects the $\mathbf{r}^d_{l,s}$ corresponding to the lead l and the algorithm d with best detection performance estimated as

  $$\bar{m}^d_l = \operatorname*{median}_{\mathrm{s}} \left\{ m^d_{l,s} \right\}. \tag{11}$$

  Note that it is a multi-algorithm multilead LSS, but the same combination of lead and detector is selected for all segments s. In contrast, A_1 may merge detections from several leads and detectors. This is indicated with the color code in the lower panel of Figure 3

To compare the performance of these three strategies, we also evaluated a simpler lead selection criterion which consists in selecting lead II, or the first available lead otherwise. This rule was evaluated for the best performing algorithms, WD and GQ (Llamedo et al. 2014), defining two LSS named $\mathrm{W_{II}}$ and $\mathrm{G_{II}}$ respectively.

The five LSS are summarized in Table 3. The interested reader is referred to the *ecg-kit* (Demski & Soria 2016), specifically to the scripts located in the *common* folder *wavedetMix.m*, *mixartif.m* and *calculateSeriesQuality.m*, for implementation details.

Table 3. The LSS evaluated in this work

LSS	Description	Lead	Detector
$\mathrm{W_1}$	Concatenation of first ranked segments from W_X (wavedet mix)	multi	single
$\mathrm{A_1}$	Concatenation of first ranked segments from A_X (all detectors mix)	multi	multi
$\mathrm{B_m}$	*Any* algorithm single-lead detections with the best $m_{d,l}$ metric	single	multi
$\mathrm{W_{II}}$	*Wavedet* algorithm detections from lead II or first available	single	single
$\mathrm{G_{II}}$	*gqrs* algorithm detections from lead II or first available	single	single

2.5. Reference performance

In order to have a reference performance, an upper-bound was defined for each algorithm (UBP). It was computed as the performance achieved when selecting the lead with the best F_l^d score. The F score is defined as

$$F = \frac{2SP}{S+P}, \tag{12}$$

being the harmonic mean between sensitivity

$$S = \frac{M_{TP}}{M_{TP}+M_{FN}} \tag{13}$$

and the positive predictive value

$$P = \frac{M_{TP}}{M_{TP}+M_{FP}}. \tag{14}$$

Considering M_{TP} and M_{FP} as the number of correct and incorrect detections, and M_{FN} the number missed heartbeats, according to the EC38 standard. Note that the UBP is based on the knowledge of the gold standard annotations, so it can not be an implementable LSS.

Similarly, with the purpose of evaluating the performance of the metric m by selecting the best quality segments, we defined two UBP's for W_1 and A_1 respectively:

- W_X selects among $\{\mathbf{r}^{W1}, \mathbf{r}^{W2}, \mathbf{r}^{W3}\}$ the best performing detections in terms of F. Note that $\mathbf{r}^{W1}$ are the detections generated by LSS W1. The detections produced by W2 and W3 are defined in a similar fashion, but selecting the second and third ranked segments s, as explained above for W1.
- A_X selects among $\{\mathbf{r}^{A1}, \mathbf{r}^{A2}, \mathbf{r}^{A3}\}$ in a similar way as explained above for W_X.

Finally, we computed the performance of W_X and A_X in terms of S, P and F in the whole evaluation set serving as UBP for W_1 and A_1.

2.6. Experiment description

The experiment performed in this work pursues the evaluation of the five LSS, presented in Table 3, in the 14 databases presented in Table 1. The performance for each LSS was based on the S, P and the F score. Then, each metric is averaged over all databases. Since each group of databases is represented by different number of recordings or subjects, the most populated groups have greater influence in the final ranking (e.g. stress with 937 recordings). Then, two performances were calculated depending on the aggregation of the results: a) all database together, i.e. each of the 1754 recordings have the same weight on the final performance, and b) by averaging the results in the five groups of databases, i.e. each group of databases has the same weight on the global performance. Finally, in order to study the differences of the detections selected for each LSS with respect to the reference gold standard, the error was calculated and presented in terms of the standard deviation.

3. Results

The results obtained from the evaluation of all LSS are summarized in Table 4, grouped by database type and ranked by the median F score in the whole evaluation set. As a reference, the best upper-bound performance is also shown in Table 4. It can be observed that the LSS based on the metric m, i.e. W_1, A_1 and B_m, achieved the first ranking in all the evaluated database groups.

Table 5 presents the results for the whole evaluation set using both averaging methods considered. The performance is shown in terms of the median S, P and F in the whole evaluation set, using the later as the ranking criterion. The algorithm's best-lead performance as well as the best results in W_X and A_X are also shown for comparative purposes. Table 5 also shows the error with respect to the annotated heartbeat locations, in terms of the standard deviation of the error, in milliseconds.

For the final evaluation, W_1 was the best performing LSS while A_X and W_X were the best performing UBP, as shown in Table 5.

Table 4. Ranking of the five LSS evaluated in each database group based on the F score.

Rank	Sinus rhythm		Arrhythmia		Stress		ST-T changes		Long-term	
1	A_1	92.0	B_m	99.8	W_1	99.8	A_1	99.0	A_1	96.5
2	G_{II}	91.8	G_{II}	99.8	W_{II}	99.1	B_m	98.3	W_1	94.8
3	W_1	89.9	W_1	99.6	B_m	97.4	W_1	98.2	B_m	89.9
4	B_m	89.6	A_1	99.5	A_1	92.7	G_{II}	0	G_{II}	80.2
5	W_{II}	88.8	W_{II}	99.3	G_{II}	88.6	W_{II}	0	W_{II}	78.8
best*	GQ	92.6	GQ	99.9	W_X	99.8	GQ	99.4	A_X	97.2

* Upper-bound performance, see 2.5 for details.

Table 5. Final ranking for the algorithm's best-lead and five LSS in all databases for both weighting schemes.

Rank		per Recording				Rank		per Group			
		S	P	F	E			S	P	F	E
	W_X*	99.9	99.5	99.7	3		W_X*	99.3	98.3	98.8	3
1	W_1	99.9	99.5	99.7	3	1	W_1	99.2	98.3	98.7	3
	WD*	99.9	99.3	99.6	2		WD*	99.2	97.9	98.5	2
2	B_m	99.4	98.6	99.0	4		A_X*	96.9	99.4	98.1	7
	PT*	98.1	99.8	98.9	3		PT*	96.6	99.6	98.1	3
	A_X*	98.8	98.9	98.8	7	2	A_1	96.6	99	97.8	8
3	W_{II}	99.6	97.7	98.6	2		GQ*	96.8	98.9	97.8	6
4	A_1	97.9	98	97.9	8	3	B_m	96.7	98.4	97.5	4
	GQ*	98.5	95.7	97.1	6		EP1*	95.5	99.3	97.4	6
5	G_{II}	96.6	86	91.0	7		EP2*	96.4	94.2	97.2	6
	WQ*	99.3	84.3	91.2	7		WQ*	93.6	96	95.3	7
	AR*	3.1	84.6	6.0	7		AR*	95.4	99	94.8	7
	SQ*	0.7	87.7	1.4	5	4	W_{II}	88.1	95.1	92.7	2
	EP1*	0.4	96.2	0.8	6		SQ*	94.3	91.2	91.5	5
	EP2*	0.3	97.4	0.6	0	5	G_{II}	91.3	89.2	90.2	7

* Upper-bound performance. See 2.5 for details.

E is the standard deviation of error in ms. See 2.5 for details.

4. Discussion

In this work we extended a previously presented strategy for selecting the best QRS detections from a multilead recording (Llamedo et al. 2014) to an LSS capable of selecting detections from multiple leads and detectors. The working hypothesis is that if it was possible to select the best performing algorithm and lead, we may also be able to create new series of heartbeat detections by concatenating the best performing segments of several series, and eventually outperform the best performing single-lead marks.

The first novelty with respect to previous works in QRS detection is the multilead and multi-algorithm extension with A_1 LSS, as shown in Figure 3. The second novelty is the extensive evaluation set used to assess the performance of the proposed LSS. The performance was compared with respect to the UBP calculated for each algorithm in Table 2. As shown in Table 1, we used a total of 1754 recordings, grouped into 5 recording types, whose cumulative length is equivalent to a single-lead recording of more than 10 years of duration. To the best of our knowledge, this represents the most extensive dataset used to evaluate the QRS detectors performance at the moment of writing this manuscript. This extensive evaluation can be understood as an analysis of the generalization ability of each detector, given that none of them was previously evaluated nor tuned to these databases.

In addition to the performance of the LSS, we also computed the UBP for each algorithm, and W_X and A_X. Those performance values were used only as a reference for comparison, since they can not be implemented.

As shown in Table 4, the LSS based on the proposed quality estimation metric m (Eq. 10 and 11) were the better ranked for all the analyzed groups. For all recording types, the performance of $\mathrm{W_1}$, $\mathrm{A_1}$ and $\mathrm{B_m}$ was better or equal than $\mathrm{G_{II}}$ and $\mathrm{W_{II}}$, which are trivial LSS based on *gqrs* (GQ) and *wavedet* (WD) algorithms. The best performing LSS was in all cases very close to the best UBP, with a difference in the F score ranging from 0.0% in stress test databases to 0.7% in long-term databases. In the same Table it can also be observed that A_X and W_X were the best-performing UBP for the stress and long-term groups, suggesting that the multi-detector/lead composition can perform better than the best single-lead in those settings. This was also confirmed in Table 5, where A_X and W_X were the best-performing UBP for both aggregation schemes. This results also suggest the usefulness of the quality estimation metric m, which is used to select the best detections that conform A_X and W_X.

When we calculated performances of the LSS over the whole evaluation set, it can be observed in Table 5 that $\mathrm{W_1}$ obtained the best ranking for both aggregation methods, resulting in the most promising LSS. Considering the *per Group* average performance, $\mathrm{W_1}$ performed equally well as the best UBP A_X, with an F score of 98.1, while $\mathrm{A_1}$ was slightly below, at 97.5. The fact that $\mathrm{W_1}$ outperforms $\mathrm{A_1}$ suggests that the metric m fails to detect the best possible detections, since the locations used by $\mathrm{W_1}$ are also included in $\mathrm{A_1}$. However, the metric allowed $\mathrm{W_1}$ to outperform all LSS and most UBP's on selecting a given lead for detector WD. It is noticeable that the best single-lead performances of the rest of the algorithms in Table 2 achieved a lower F score than $\mathrm{W_1}$. This result suggests that the metric m is adequate for the ranking and posterior selection of promising candidates among several heartbeat detections. Similar conclusions can be obtained from results obtained by aggregating all recordings with the same weight.

Note that the metric m relies in the proper modeling of $f(\mathbf{x})$. In this work, the parameters were estimated in a quite small dataset (20 recordings of thew15). In future works it will be analyzed whether adding more data to the training set could improvement performance of m. With respect to the classes considered in the model, note that the missed beats (FN) are detected mainly thanks to the abrupt deceleration in the RR interval tachogram. This highlights the importance of the features related to the mean heart rhythm ($\bar{R}_{10}$, $\bar{R}_{60}$): an abrupt increase in R_n is likely to occur in the presence of FN, while a decrease may be related to FP.

Another important aspect of a detector's performance is how bad it performs when running in a set of signals recorded in completely different settings. Some of the evaluated detectors absolutely fail to detect heartbeats in several types of recordings. This can be seen in the best-lead performance achieved by detectors EP1/2, SQ, WQ and PT. These detectors failed to detect heartbeats in at least 5% (89) of the recordings. The case of the PT algorithm is particularly interesting, as its median performance achieved is as high as $F_{50} = 98.9$ %, which contrasts with the 5 *th.* percentile $F_5 = 0$ %. When analyzing the performance by group of databases, the PT algorithm, as it was expected, fails more often in stress recordings which are the most frequent in our evaluation dataset. On the other hand, algorithms such as WD and GQ present quite balanced performances, achieving higher rankings in Table 5.

One possible drawback of concatenating detections from several leads and/or algorithms could be an increase in the variance of the time error with respect to the true location of the heartbeats. This aspect was also studied for all the experiments performed in this work. We found that WD algorithm had an error with standard deviation (sd) of 2 ms for its best lead, while the W_1 obtained 3 ms, and A_1 achieved 8 ms. On the other hand, the algorithms GQ and WQ presented an sd of 6 ms and 7 ms respectively. As these results suggest, the increased sd due to the segment concatenation is negligible in those LSS. Finally, the strategy B_m achieved an sd of 4 ms, as can be noted, it is lower than in A_1 since this strategy does not merge QRS detections from different leads and algorithms. The use of strategies and detectors with small variance is relevant for applications where stability in QRS locations is needed, as in heart rate variability analysis. In any case, any series of QRS detections can be refined in a post-processing step by using cross-correlation methods.

The experimental part of this work was implemented using the *ecg-kit* toolbox for Matlab (http://marianux.github.io/ecg-kit/)(Demski & Soria 2016) and public databases (Goldberger et al. 2000, Couderc n.d.) in order to ensure the reproducibility of the results presented. Strategies B_m, W_1 and A_1 are also implemented in the *ecg-kit* for comparison with future detectors (See *wavedetMix*, *mixartif* and *calculateSeriesQuality*).

5. Conclusion

The results presented in this work suggest that lead selection strategies W_1, A_1 and B_m outperform the simple strategy of always selecting lead II or the first lead available (G_{II}, W_{II}). Moreover, strategy W_1, which selects the lead with best QRS detections using a wavelet-based QRS detector, outperforms the maximum theoretical performance of all considered algorithms, and it was only slightly below the UBP defined by A_X and W_X.

In conclusion, the quality metric m used to select and concatenate the most convenient segments of QRS detections, resulting in LSS $\mathrm{W_1}$, $\mathrm{A_1}$ and $\mathrm{B_m}$, outperform all the best performing single-lead from all algorithms evaluated in this work. This suggest that the presented method is convenient to improve the performance of automatic QRS complexes detection in a broad set of settings, as those evaluated in this work.

Acknowledgments

This work was supported by projects ASUTNBA0003860, ICUTNBA0004864 and ICUTNBA0004887 from UTN FRBA, Argentina; by project TEC2016-75458-R from Spanish Ministry of Economy and Competitiveness (MINECO), Spain, and and by Gobierno de Aragón (Reference Group BSICoS T39-17R) cofunded by FEDER 2014-2020 "Building Europe from Aragon". The CIBER of Bioengineering, Biomaterials and Nanomedicine is an initiative of ISCIII.

References

AHA 2010 ‘American heart association ECG database’.
URL: *https://www.ecri.org*

Almeida R, Martínez J P, Rocha A P & Laguna P 2009 *IEEE Transactions on Biomedical Engineering* **56**(8), 1996–2005.

Arzeno N M, Poon C S & Deng Z D 2006 *in* ‘2006 International Conference of the IEEE Engineering in Medicine and Biology Society’ IEEE pp. 1788–1791.

Benitez D, Gaydecki P, Zaidi A & Fitzpatrick A 2001 *Computers in biology and medicine* **31**(5), 399–406.

Bishop C 2006 *Pattern Recognition and Machine Learning* Information Science and Statistics Springer.
URL: *http://books.google.it/books?id=kTNoQgAACAAJ*

Clifford G D, Behar J, Li Q & Rezek I 2012 *Physiological Measurement* **33**(9), 1419–1433.
URL: *https://doi.org/10.1088/0967-3334/33/9/1419*

Couderc J P n.d. ‘The telemetric and holter ECG warehouse initiative (thew)’.
URL: *thew-project.org*

Demski A J & Soria M L 2016 *Journal of Open Research Software* **4**(1), 2–5.

Elgendi M 2013 *PLoS ONE* **8**(9), e73557.
URL: *http://dx.doi.org/10.1371/journal.pone.0073557*

Fischer R, Sinner M, Petrovic R, Tarita E, Kääb S & Zywietz T K 2008 *in* ‘Computers in Cardiology’ Vol. 35 pp. 453–456.

Goldberger A L, Amaral L A N, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C K & Stanley H E 2000 *Circulation* **101**(23), e215–e220.

Hamilton P S & Tompkins W J 1986 *Biomedical Engineering, IEEE Transactions on* **BME-33**(12), 1157–1165.

Kohler B U, Hennig C & Orglmeister R 2002 *IEEE Engineering in Medicine and Biology Magazine* **21**(1), 42–57.

Ledezma C A & Altuve M 2019 *Medical & Biological Engineering & Computing* **57**(8), 1673–1681.
URL: *https://doi.org/10.1007/s11517-019-01990-3*

Llamedo M & Martínez J P 2014 *in* ‘Computers in Cardiology, 2014’ pp. 357–360.

Llamedo M, Martínez J P & Laguna P 2014 *in* ‘Computers in Cardiology, 2014’ pp. 721–724.

Martínez J P, Almeida R, Olmos S, Rocha A & Laguna P 2004 *IEEE Transactions on Biomedical Engineering* **51**, 570–581.

Mehta S S & Lingayat N S 2008 *Computers in biology and medicine* **38**(1), 138–145.

Mondelo V, Lado M, Méndez A, Vila X & Rodríguez-Liñares L 2017 *Journal on Advances in Theoretical and Applied Informatics* **3**(1), 5–9.
URL: *https://revista.univem.edu.br/jadi/article/view/2436*

Moody G B & Mark R G 1982 *in* 'Computers in cardiology' Vol. 9 pp. 39–44.
URL: *http://ecg.mit.edu/george/publications/ecg-cinc-1982.pdf*

Pan J & Tompkins W J 1985 *Biomedical Engineering, IEEE Transactions on* **0**(3), 230–236.

Paoletti M & Marchesi C 2006 *Computer Methods and programs in biomedicine* **82**(1), 20–30.

Sameni R 2006 *Open Source ECG Toolbox (OSET).*
URL: *http://ecg.sharif.ir/*

Satija U, Ramkumar B & Manikandan M S 2018 *IEEE Reviews in Biomedical Engineering* **11**, 36–52.

van der Heijden F, Duin R, de Ridder D & Tax D 2005 *Classification, Parameter Estimation and State Estimation* John Wiley & Sons.

World Health Organization 2012 'Cardiovascular diseases'.
URL: *http://www.who.int/cardiovascular_ diseases/en/*

Zong W, Moody G & Jiang D 2003 *in* 'Computers in Cardiology, 2003' pp. 737–740.