Biomedical Signal Processing and Control 68 (2021) 102628

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Certainty in QRS detection with artificial neural networks☆

Jonas Chromik *, Lukas Pirl, Jossekin Beilharz, Bert Arnrich, Andreas Polze

Hasso Plattner Institute, University of Potsdam, Potsdam, Germany

ARTICLE INFO

Keywords:
QRS detection
Electrocardiography
Artificial neural networks
Machine learning
Signal-to-noise ratio

ABSTRACT

Detection of the QRS complex is a long-standing topic in the context of electrocardiography and many algorithms build upon the knowledge of the QRS positions. Although the first solutions to this problem were proposed in the 1970s and 1980s, there is still potential for improvements. Advancements in neural network technology made in recent years also lead to the emergence of enhanced QRS detectors based on artificial neural networks. In this work, we propose a method for assessing the certainty that is in each of the detected QRS complexes, i.e. how confident the QRS detector is that there is, in fact, a QRS complex in the position where it was detected. We further show how this metric can be utilised to distinguish correctly detected QRS complexes from false detections.

1. Introduction

The beating of the heart is one of the most basic and most important vital signs considered in medicine. Consequently, parameters describing the heartbeat are important tools for monitoring a patient's condition. One such parameter is the heart rate (HR) describing how fast the heart beats [1]. To compute parameters such as HR, we need to find out *when* the heart beats. This can be done by monitoring the heart's electrical activity, a method called electrocardiography. Electrocardiography produces an electrocardiogram (ECG), a biomedical signal that corresponds to said electrical activity. In this ECG, there are many different waves and spikes. The most striking deflection is the QRS complex which corresponds to the contraction of the heart's ventricles and hence to the heartbeat [2]. Thus, accurate detection of QRS complexes is important to compute vital parameters such as HR. Furthermore, downstream analyses heavily rely on the precise knowledge of the QRS complex's position. One example is alarm generation in patient monitors. These monitors check that vital parameters (e.g. HR) are in a predefined healthy range (e.g. between $60\,\mathrm{min}^{-1}$ and $120\,\mathrm{min}^{-1}$) and alarm otherwise. In addition, arrhythmia detection in patient monitors also relies on prior QRS detection [3]. Thus, errors in QRS detection propagate in multiple ways (see Fig. 1), leading to false alarms and subsequently alarm fatigue [4,5]. To prevent this, QRS detection needs to be as good as possible. Perfect accuracy is not always achievable due to different kinds of noise in the ECG signal, such as muscle artefacts or electrode motion [6]. In such cases were QRS detection is impaired, we want to at least know that it is impaired and hence results are not reliable. The aim of this paper is to introduce a *certainty* metric indicating how reliable the QRS detector's results are.

The rest of this work is structured as follows: In the remainder of Section 1 we give an overview on the state of the art in QRS detection and introduce the problem of uncertainty in QRS detection. Furthermore, we discuss potential applications of the certainty metric which is the main contribution of this work. Lastly, for Section 1, we give an overview of related work. In Section 2 we describe the materials this work is based upon, i.e. QRS detectors utilising artificial neural networks (ANNs) and ECG databases. In Section 3 we explain the inner workings of ANN-based QRS detectors as well as their shortcomings. Furthermore, we define the concept of certainty in QRS detection to address said shortcomings. In Section 4 we present the results of our evaluations on the proposed certainty metric. Also, we show how different factors such as noise, different beat types and training parameters influence certainty. In Section 5 we discuss the potential of certainty for increasing the QRS detectors performance as well as the reliability of downstream algorithms such as alarm generation in patient monitors. Finally, we describe limitations and future work.

1.1. State of the art

QRS detection is a long-standing field with its roots tracing back to the 1970s [7] and 1980s [8]. Initially, QRS detection relied on digital filters and moving window integration. Since then, a variety of

☆ This work was partly funded by the German Federal Ministry of Education and Research (BMBF) under grant number 16SV8559.
* Corresponding author.
E-mail address: jonas.chromik@hpi.de (J. Chromik).

https://doi.org/10.1016/j.bspc.2021.102628
Received 4 June 2020; Received in revised form 15 March 2021; Accepted 5 April 2021
Available online 24 April 2021

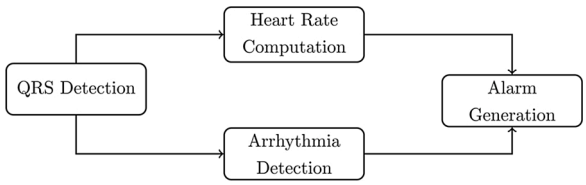

Fig. 1. Alarm generation in patient monitors relies on QRS detection in multiple ways.

approaches to QRS detection has been studied [9]. One of these approaches is QRS detection based on ANNs. This approach is the basis of this work and will be covered in greater detail in Section 2.1.

Nowadays QRS detectors exhibit good performance with sensitivity (Se) and positive predictive value (PPV) above 99% [9] on low-noise ECGs. In reality, however, there is noise leading to impaired performance in QRS detection [6]. Detecting and dealing with periods of impaired signal quality and detection performance manifests a problem in the state-of-the-art of QRS detection which we will discuss next.

1.2. Problem with the state of the art QRS detection

Most of the current approaches to QRS detection have in common that they yield the positions of the QRS complexes as their sole output.[1] While there are digital filters in place to reduce the influence of noise, there is no feedback given on data quality or the interpretability of the ECG. Consequently, ECGs with regularly shaped QRS complexes and low levels of noise are dealt with in the same way as ECGs with abnormal QRS complexes[2] and high levels of noise. Thus, errors in QRS detection pass unnoticed corrupting all analyses that build upon information on QRS positions. Consequently, vital parameters are erroneously computed and false alarms are triggered, leading to stress and even harm for patients and staff [4]. Hence, a mechanism for reporting certainty along with the QRS positions is required in QRS detectors.

1.3. Contribution of this work

In this work, we address the problem of uncertainty in QRS detection. We define a certainty metric for QRS detectors which are based on ANNs. Furthermore, we modify these detectors in order to have them report said certainty metric along with their detections. In addition, we investigate how the certainty metric can be used to distinguish between correctly detected QRS complexes and false detections.

The problem of low signal quality in ECG signal was already addressed in a variety of manners by related work which we discuss in Section 1.5. However, none of the predominant approaches quantifies how well a specific QRS detector can interpret the ECG signal which is the unique characteristic of this work.

Hence, the research questions answered in this work are:

RQ1 How to quantify how well a QRS detector can interpret a given ECG signal? (This is achieved by the certainty metric and through the example of ANN-based QRS detectors.)

RQ2 How can the certainty metric from RQ1 be utilised to distinguish between true and false detections?

RQ3 How do different factors such as signal-to-noise ratio (SNR) and beat types influence the certainty metric from RQ1?

[1] There are few exceptions where the QRS detector has a notion of uncertainty and reports this information, e.g. [10].

[2] Abnormal QRS complexes are caused by pathologies and exhibit a deviant shape. Therefore, they are harder to detect since QRS detectors are optimised for regularly shaped QRS complexes. This is either by design or accidentally by having normal beats as majority class in the training data.

1.4. Potential applications

The concept of certainty in QRS detection can be used for a variety of tasks. In the following, we want to give some examples.

Alarm generation in patient monitors. In intensive care units (ICUs), patient monitors are used for early detection of life-threatening situations in patients. ECG signals are used to detect conditions such as tachycardia, bradycardia, or arrhythmia. However, the quality of the ECG signal gets impaired for various reasons, such as patient movement, loose electrodes, changes in posture, or power-line interference [6]. Certainty can help to identify such impairments and selecting the ECG lead that is least affected. Thus, the quality of QRS detection is improved and hence we can achieve higher accuracy in derived metrics, such as heart rate. This, in turn, leads to fewer false alarms.

Holter record review. When reviewing long-term ECG recordings generated by wearable Holter monitors, huge amounts of data have to be checked manually by a healthcare professional. Certainty can help by providing an estimate for signal quality, giving guidance for the reviewer in the task of finding potential errors in QRS detection. Moreover, this kind of signal quality assessment is specific to the task of QRS detection, i.e. only indicating bad signal quality where QRS detection is impaired by this.

Active learning. In active learning, the machine learning model is allowed to request labels while not being in the training phase anymore (i.e. while doing predictions). The ultimate goal of this approach is to reduce the amount of training data needed. Whenever the model is uncertain about a given sample, a label is requested to improve the model for this very kind of sample [11]. In QRS detection, a QRS position could be requested from a medical professional whenever the ANN is uncertain (e.g. due to a previously unseen QRS shape), thus improving the model itself.

1.5. Related work

There are different approaches to dealing with noise in ECGs. In the following, we give an overview of the options one has when confronted with a noisy ECG. All approaches have in common, that they are separated from the actual QRS detection taking place either before or after. The unique characteristic of the certainty metric is, that it is built into the QRS detector and can hence give an immediate assessment on how much the noise impairs the task of QRS detection.

Denoising. Denoising techniques aim at separating signal and noise. This allows removal of noise while retaining only the clean signal. A survey on noise removal techniques can be found in [12]. Some kinds of noise such as power-line interference or baseline wander are easily removed by band-pass filters. There are, however, other kinds of noise such as muscle artefacts and electrode motion artefacts that are hard to separate from the signal [6]. This is because the frequency spectrum of these kinds of noise is very similar to the frequency spectrum of the signal (i.e. the heart's electrical activity). Therefore, separating signal from noise is not always feasible.

Signal quality indicators (SQIs). Apart from noise removal there are also approaches to noise assessment trying to determine the noise's intensity and even its type. Such metrics of overall signal quality are called signal quality indicators (SQIs). A 2014 review gives an overview of the predominant methods of signal quality assessment with special regard to their influence on heart rate and blood pressure derived metrics [13]. A 2016 paper assesses the interplay between SQIs and heartbeat (QRS) detection [14]. The limitation of the use of SQIs regarding this work is that SQIs provide am general signal-quality assessment. What we aim at in this work is a quality assessment for a specific task, i.e. whether the signal is usable for a specific QRS detector. This is not reliably achieved with SQIs.

Use of other physiological signals. With high levels of noise on the ECG, detection of QRS complexes using the ECG signal alone might not be feasible. In such a case, additional pulsatile physiological signals such as

photoplethysmography (PPG) and arterial blood pressure (ABP) can be used to facilitate QRS detection. This approach is called multimodal beat detection and was covered extensively by the 2014s PhysioNet/Computing in Cardiology Challenge of which [15] is the summary paper. This approach is, however, only feasible if other physiological signals are available which is not always the case.

ECG signal reconstruction. ECG signal reconstruction aims at reproducing the ECG signal from other physiological signals. This can be used when the ECG signal drops out (e.g. due to loose electrodes) or the signal-to-noise ratio becomes too low (i.e. too much noise). This topic was covered extensively by the 2010s PhysioNet/Computing in Cardiology Challenge of which [16] is the summary paper. Although the results of the challenge seem promising, the practical applicability for QRS detection in noisy ECGs is limited since additional physiological signals (other than the one to reconstruct) such as other ECG leads, ABP, PPG, or others are required. As in multimodal beat detection, signal availability is a limitation here. Even if other signals are available, one could use these signals for beat detection right away without the detour via signal reconstruction that might introduce additional errors.

Two-step QRS detection. We found one unpublished QRS detector, GQRS, with separate postprocessing utility.[3] This way, QRS detection becomes a two-step procedure. The detector is optimised for Se. The postprocessing tool, GQPOST, however, removes some of the detections according to a given threshold thereby increasing PPV at the expense of Se. This is similar to our approach since here, too, postprocessing is a specialised step that accounts for uncertainty in the prior detection process. However, since this is unpublished work, we do not know by which criteria QRS detections are rejected or retained.

Other applications of neural networks in electrocardiography. Apart from QRS complex detection and signal quality assessment for noisy ECGs, there are also other applications for neural network technology in the field of electrocardiography. Although this field is too extensive to fully cover it here, we want to at least mention some applications.

In long-term ECG recordings having only few leads, detection of atrial fibrillation (AFib) is an established topic. Recent approaches are using convolutional neural network (CNNs) [17] and long short-term memory neural network (LSTMs) [18]. This topic was also covered by the 2017s Physionet/CinC Challenge [19].

In short-term 12-lead ECGs, the detection of various cardiovascular diseases (CVDs) using neural networks is a recent topic of research interest. [20] and [21] provide two solutions to this problem. Furthermore, this is also the topic of the 2020s Physionet/CinC Challenge which additionally underlines the recency of this topic [22].

2. Materials

In this section, we introduce the materials that lay the foundations for our work. Materials are two-fold: First, we introduce three QRS detectors based on neural networks. We will modify these detectors in a way that they not only yield QRS positions but also a certainty score for each detected QRS position. Second, we introduce the ECG databases we use to evaluate our approach. Here we rely on a combination of clean ECGs and noise template (as created by [6]) which we merge in order to achieve a well-defined SNR.

2.1. QRS detection with artificial neural networks

In this section, we describe three ANN-based QRS detectors. Thereby we show an evolution from the beginning of ANN-based QRS detection in 1997 to very recent developments. Furthermore, we uncover how detection errors and uncertain detections manifest in this specific approach to QRS detection.

García-Berdonés detector. Detecting QRS complexes using artificial neural networks was first introduced by [23] in 1997 by means of a multilayer perceptron (MLP). We will refer to this detector as the *García-Berdonés* detector after its first author. Their approach involves three main steps. These steps also manifest the core concept of all subsequent ANN-based QRS detectors.

1. For each sample ("A") in the ECG signal, take a neighbourhood of n samples. This neighbourhood is called *window* and corresponds to a small snippet of the ECG signal (Fig. 2).
2. Use an MLP for binary classification of each window, whether it contains a QRS complex (class 1) or not (class 0). By processing all windows in this manner, we create a new signal which we call *trigger signal*. Ideally, the trigger signal exhibits a square pulse shape with pulse plateaus at the positions where the ECG signal exhibits QRS complexes (Fig. 6).
3. Find the midpoint of each pulse plateau. These midpoints correspond to the precise QRS positions in the ECG signal.

Šarlija detector. Advancements in neural network technology also lead to improved architectures for ANN-based QRS detectors. In 2017, [24] proposed a QRS detector based on a CNN instead of a MLP. We will refer to this detector as the *Šarlija* detector after its first author. Another novelty of the Šarlija detector is, that the pulse width in the trigger signal is uncoupled from the window size provided to the CNN. The CNN is trained to classify a window as "contains QRS complex" (class 1) if and only if the QRS complex is in a predefined detection area in the middle of the window (Fig. 3). Hereby, the CNN can be provided with a larger section of the ECG signal, hence providing more contextual information around the QRS complex, without inappropriately increasing the width of the pulse plateaus in the trigger signal.

Xiang detector. In 2018, [25] proposed a non-sequential CNN approach to QRS detection. We will refer to this detector as the *Xiang* detector after its first author. There are two signals generated from the provided ECG signal. The first signal is subjected to a difference operation and fed into a "part level CNN". The second signal is subjected to a window average and a difference operation and fed into an "object-level CNN". Afterwards, a MLP utilises the outputs of the two CNNs to make the final decision on whether a QRS complex was detected in the detection area or not. Thus, generating the samples for the trigger signal (Fig. 4).

Comparison and synthesis. All of the described approaches follow the same overall structure, shown in Fig. 5. *First*, there is a preprocessing step. This preprocessing step might apply some operations to the input ECG signal (e.g. filtering, difference operation, window averaging). These operations are specific to each detector. However, the

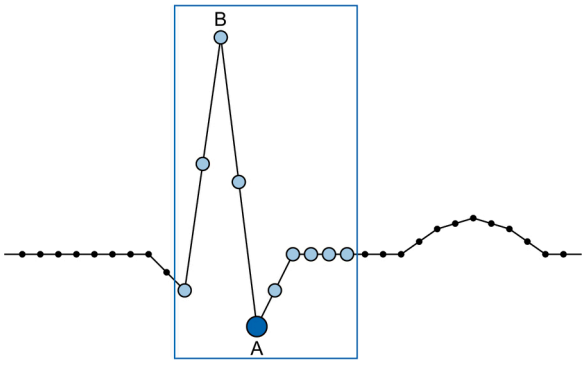

Fig. 2. All samples (blue dots) within a window (blue frame) are fed into the MLP for classification of the window as "does contain QRS complex" (class 1) or "does *not* contain a QRS complex" (class 0). Samples outside the window (black dots) are omitted, at least for this window. "A" marks the midpoint of the window and hence the centre of the neighbourhood. "B" marks the midpoint of the QRS complex as defined in the ground-truth data (see Section 2.2).

[3] https://www.physionet.org/physiotools/wag/gqrs-1.htm.

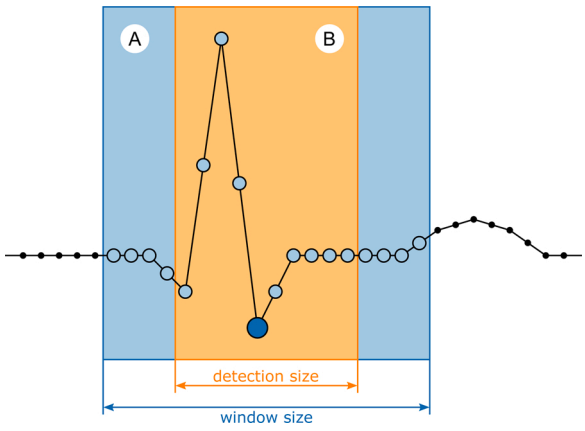

Fig. 3. All samples within area A (defined by window size) are used as input for the neural network. However, only if the QRS complex is within area B (defined by detection size), a detection shall be indicated.

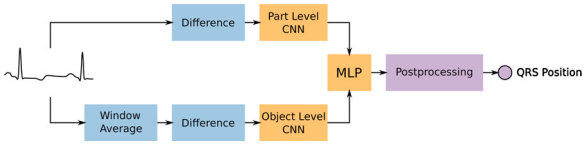

Fig. 4. Gross architecture of the Xiang detector. Blue boxes are part of the preprocessing step. Orange boxes are part of the ANN classifier. Purple indicates postprocessing with specific QRS positions only available after postprocessing. Before, there is only the trigger signal.

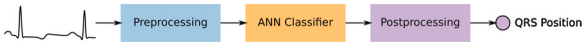

Fig. 5. Overall, abstract architecture of all ANN-based QRS detectors considered by this work (García-Berdonés, Šarlija, and Xiang). There is always a preprocessing step, an ANN used for classification, and a postprocessing step. The concrete implementations of these steps vary. The final output is always a (list of) QRS position(s).

preprocessing step always culminates in creating windows from the signal. In a *second* step, these windows are then used as input for an ANN classifier. The concrete architecture of the ANN is like the preprocessing step specific to each detector. The ANN generates from the windows another signal which we call trigger signal. Ideally, this trigger signal is a square pulse signal having a value of 0 in areas where no QRS complex is present and a plateau of value 1 where a QRS complex is present (Fig. 6). *Finally*, a postprocessing step is used to transform the trigger signal into QRS positions. The QRS positions are in the middle of the plateaus in the trigger signal.

Uncertainty and intermediate values. In practice, there is uncertainty:

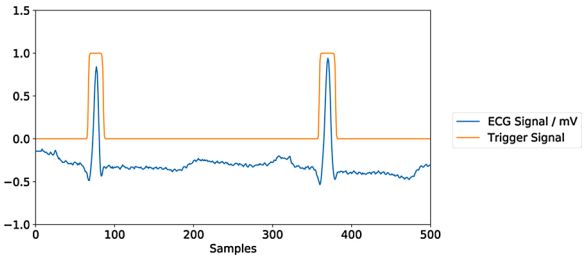

Fig. 6. ECG signal (in blue) and the corresponding trigger signal (in orange) in its supposed square pulse shape. Pulse plateaus can easily be distinguished from the rest of the trigger signal and hence detecting the QRS complexes is easily done via the trigger signal.

The output of the ANN (be it CNN or MLP) is not discrete (0 or 1) but a continuous value between 0 and 1, for example, 0.8. This is due to the nature of ANN classifiers. Such intermediate values occur when the ANN cannot definitely place the given window in one of the two classes (0 – no QRS complex; or 1 – QRS complex). Hence, intermediate values indicate that the ANN is uncertain whether a QRS complex is present or not. We use the term *flawed trigger signal* for trigger signals exhibiting such uncertainty and will utilise this behaviour later-on for computing the certainty metric. Therefore, we modify the postprocessing step, as shown in Fig. 7.

Furthermore, the output is not always consistent. For example, there is ripple, meaning a few samples where the trigger signal has a value close to zero in the middle of a plateau with values close to 1 [23] (Fig. 8). Hence, there are postprocessing steps involved to make sense from the trigger signal. This postprocessing step is where the certainty metric will be computed. Therefore, we will provide a detailed explanation of postprocessing in Section 3.

Implementation. We implemented all ANN-based QRS detectors discussed in this section (Section 2.1). As programming language for implementation, we chose Python[4] 3 (≥3.6) for all components. For implementing ANNs, we used the Keras[5] (≥2.2.0) [26] in conjunction with Tensorflow[6] (≥1.9.0) [27]. Furthermore, we used the native Python WFDB package[7] for reading the ECG files from the ECG databases described in Section 2.2. These databases were provided via Physionet [28].

2.2. ECG databases

For evaluation, we are using two types of ECG databases: First, the MIT-BIH Arrhythmia Database (`mitdb`) [29] as a source for clean ECGs with low levels of noise. And second, the MIT-BIH Noise Stress Test Database (`nstdb`) [6] which provides ECGs with added noise as well as the noise templates. Both databases are described in the following.

MIT-BIH Arrhythmia Database. The `mitdb`[8] contains 48 clean, two-channel ECGs. These ECGs contain cardiac arrhythmias and are fully annotated by two independent cardiologists. These annotations include information on heart rhythms, abnormal beat, but also specific QRS positions [29]. We use these expert annotations of QRS positions as ground truth for our works on QRS detection. Specifically, when generating windows from the ECG signal, we check with the annotation file whether there is a QRS complex in the time-frame covered by the window or in the detection area of the window, respectively (cf. Fig. 3). If so, the windows is labelled as class 1 (contains a QRS complex). Otherwise, it is labelled as class 0 (contains no QRS complex).

An alternative to the `mitdb` is the MIT-BIH Normal Sinus Rhythm Database (`nsrdb`).[9] This database also provides high-quality, low-noise ECGs. However, the nsrdb provides only ECGs from hearts exhibiting a sinus rhythm which is the physiological state of the heart. Since ECGs (especially long-term ECGs) usually contain some pathology of the heart, `mitdb` might be closer to the real use case of a QRS detector.

Fig. 7. The change this work proposes to the architecture shown in Fig. 5 (marked with a red frame). We propose generating not only QRS positions but also a certainty score for each position. Therefore, the postprocessing part needs to be changed.

[4] https://www.python.org/.
[5] https://keras.io/.
[6] https://www.tensorflow.org/.
[7] https://github.com/MIT-LCP/wfdb-python.
[8] https://physionet.org/content/mitdb/1.0.0/.
[9] https://www.physionet.org/content/nsrdb/1.0.0/.

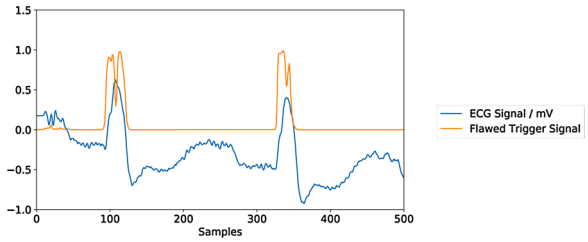

Fig. 8. ECG signal (in blue) and a corresponding trigger signal (in orange) that does not show the expected square pulse shape. In this case, plateaus are harder to detect and the detection process requires additional steps, i.e. discretization and ripple removal.

Hence we decided to use `mitdb` rather than nsrdb as a source for low-noise ECG signals.

MIT-BIH Noise Stress Test Database. The `nstdb`[10] contains ECGs with different levels of noise (SNR $\in \{-6\,dB,\ 0\,dB,\ 6\,dB,\ 12\,dB,\ 18\,dB,\ 24\,dB\}$) over some periods. These noisy ECGs where created using two clean recordings from the `mitdb` (#118 and #119) and noise templates. These noise templates were created by attaching ECG electrodes to the subject's limbs in a way that the heart's electrical activity is not visible in the recording. The subjects were then asked to be physically active. Afterwards, three kinds of noise were identified through visual inspection: baseline wander (`bw`), electrode motion (`em`), and muscle artefacts (`ma`). The authors of the database considered `em` as most troublesome since "it can mimic the appearance of ectopic beats and cannot be removed easily by simple filters, as can noise of other types" [6]. Thus, we too use `em` as a template for adding noise to clean ECGs.

3. Methods

In Section 2.1 we explained, how the trigger signal does not always exhibit the expected square pulse shape but rather tends to be flawed when the neural network does not definitely recognise the deflection in question as a QRS complex. In this section, we describe how intermediate signal amplitudes and ripple can be dealt with. On top of that, we define the concept of certainty in QRS detection as well as metrics for measuring certainty. These are the main contributions of this work. Beyond this, we demonstrate how certainty metrics can be utilised to distinguish between correctly detected QRS complexes and false detections. This, in turn, improves the accuracy of derived metrics (e.g. heart rate), avoiding false alarms and wrong diagnoses.

3.1. Finding QRS complexes with a flawed trigger signal

The first step in making sense from a flawed trigger signal is finding potential square pulses and hence candidates for QRS complexes. These candidates might be actual QRS complexes but could also be false detections. To distinguish actual QRS complexes from these candidates, we will refer to QRS complex *candidates* as *trigger points* throughout the rest of this work. Actual QRS complexes are referred to as QRS complexes.

The García-Berdonés detector [23] uses discretization and ripple removal as a means to make sense from the flawed trigger signal and to generate trigger points. Discretization is done via Eq. (1) whereas *ts* is the raw trigger signal as *ds* is the discretized signal. The relevant decision here is which value to choose for the discretization threshold (*th*). García-Berdonés et al. summarise this problem as follows: "Low (large) thresholds will decrease (increase) the number of missing detections but will increase (decrease) the number of false detections." [23]

$$\mathrm{ds}(x) = \begin{cases} 1, & \text{if } \mathrm{ts}(x) \geq \mathrm{th} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

[10] https://physionet.org/content/nstdb/1.0.0/.

After discretization, ripple has to be removed. Whereas ripple refers to a few consecutive samples of the discretized signal with value 0 within a plateau of samples with value 1 (Fig. 9). For this, we have to define a tolerance value (*to*) to specify how many consecutive 0-samples within a plateau of 1-samples count as ripple and when we consider this as two plateaus. For example, with $to = 2$, two consecutive samples with value 0 within 1-samples are considered to be ripple and the signal gets corrected to show one plateau (Eq. (2)). With $to = 1$ we interpret this pattern as two plateaus (without correction, Eq. (3)).

$$(1,1,1,0,0,1,1,1) \rightarrow (1,1,1,1,1,1,1,1) \quad \text{with} \quad to = 2 \tag{2}$$

$$(1,1,1,0,0,1,1,1) \rightarrow (1,1,1,0,0,1,1,1) \quad \text{with} \quad to = 1 \tag{3}$$

The problem we are facing in postprocessing is, that reasonable default values for the parameters *th* and *to* are hard to determine. García-Berdonés et al. propose various values between 0.001 and 0.999 for *th* [23], while Šarlija et al. and Xiang et al. do not address the problem at all [24,25].

Through discretization and ripple removal a *rectified trigger signal* is created, as to be seen in Fig. 10. In this rectified trigger signal, plateaus are easily recognisable, and hence trigger point generation can be done by finding the midpoints of these plateaus. In the following, we want to show how the trigger signals (flawed and rectified) can be utilised to determine how much certainty there is in each trigger point.

3.2. Naive definition of the certainty metric

In an ideal case, where QRS detection is easy for the detector, the trigger signal will exhibit a square pulse shape. Consequently, the more the shape of the trigger signal differs from its supposed square pulse shape, the more uncertainty there is in the corresponding trigger point. We quantify this difference with two different metrics for certainty, *naive certainty* (C) and *adaptive certainty* (C'). Later in this work, we compare both metrics against each other. In the remainder of this section, we will define and demonstrate naive certainty. In Section 3.3, we will explain some shortcomings of naive certainty which we address with adaptive certainty.

Naive certainty (C) compares the area under the flawed trigger signal with the area under the rectified trigger signal. According to this, we define a certainty metric C_i as the certainty of trigger point i in Eq. (4).

$$C_i = \frac{\int_{b_i}^{e_i} t_f(x)\mathrm{dx}}{\int_{b_i}^{e_i} t_r(x)\mathrm{dx}} \tag{4}$$

Whereas b_i and e_i are the begin and end of plateau i, as shown in Fig. 11. t_f is the flawed trigger signal and t_r is the rectified trigger signal.

Fig. 12 shows a raw ECG signal, the corresponding trigger signal, and trigger points with colour-coded certainty. As to be seen, false-positive trigger points exhibit lower certainty (deep blue colour) that true positive trigger points having a greenish-yellow colour indicating high certainty. Hence, if we treat this approach as a generate-and-test-pattern, certainty may be utilised in the test step to distinguish between true and false detections. This is shown in Eq. (5). TP_a is the set of actual trigger points that supposedly correspond to QRS complexes. TP_c the set of

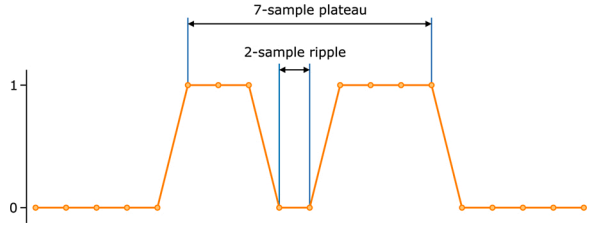

Fig. 9. Trigger signal with a 7-sample plateau having a ripple of 2 samples within.

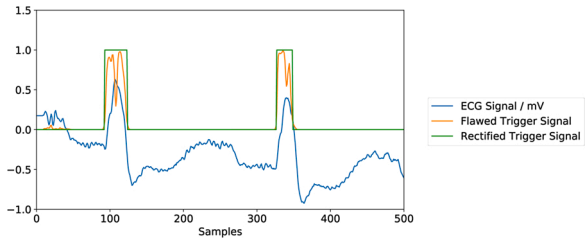

Fig. 10. ECG signal (in blue) together with a flawed trigger signal (in orange) that does not exhibit the expected square pulse shape. Furthermore, a rectified version of the flawed trigger signal is shown in green.

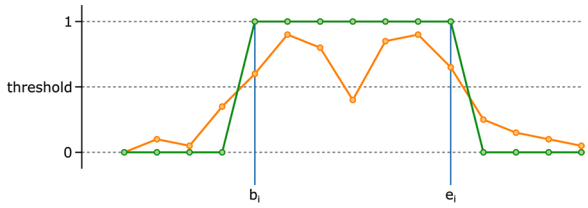

Fig. 11. Schematic drawing of flawed (orange) and rectified (green) trigger signal showing begin and end of a plateau, i.e. first and last sample above a predefined threshold.

Fig. 12. Three information are to be found in this figure. The bottom plot shows the raw ECG signal. The top plot shows the trigger signal generated from the ECG signal. The circles in the top plot are potential trigger points with colour-coded certainty.

trigger point candidates containing some trigger points with low certainty which probably do not correspond to QRS complexes. Finally, *ct* the certainty threshold used to distinguish between true and false detections. We show a method for finding suitable values for *ct* in Section 3.4.

$$\mathrm{TP}_a = \{\mathrm{tp} | \mathrm{tp} \in \mathrm{TP}_c \wedge \mathrm{certainty}(\mathrm{tp}) > \mathrm{ct}\} \tag{5}$$

3.3. Adaptive definition of the certainty metric

Naive certainty is defined as the ratio between the area under the flawed trigger signal and the area under the rectified trigger signal. However, this ratio tends to be skewed when the width of the rectified trigger signal varies. Examples of this can be seen in Fig. 25 and Fig. 26 and we will discuss this issue in greater detail in Section 5.3. For now, we want to provide an alternative definition of certainty which we call adaptive certainty since it adapts to the plateau width the ANN classifier should produce.

For the definition of adaptive certainty, we use the *expected plateau width* (w_e). The expected plateau width is the width of the square pulse the ANN classifier should produce under ideal circumstances. This width is equal to the width of the window that is used as input for the ANN – or the width of the detection area in the window, respectively. Eq. (6) shows the definition of adaptive certainty (C').

$$C_i' = \frac{\int_{b_i}^{e_i} t_f(x)\mathrm{dx}}{w_e} \tag{6}$$

In Sections 4.2 and 5.3, we will show how the two certainty metrics behave in comparison when using certainty to distinguish correct detections from false ones.

3.4. Certainty threshold determination

For distinguishing true and false trigger points using certainty, we have to find a proper certainty threshold. In the following, we describe a method for determining a suitable threshold. As a general notion, a lower certainty threshold will cause the QRS detector to yield more trigger points, including false ones, leading to lower PPV for the detector. On the contrary, higher certainty thresholds will yield fewer trigger points, potentially discarding true ones, leading to a lower Se of the detector.

If we plot the number of trigger points yielded against the certainty threshold, we expect to see a graph similar to the one shown in Fig. 13. For low certainty thresholds, the number of trigger points, including false ones, is high. Then, with increasing certainty threshold, low certainty (supposedly false) trigger points are discarded, leading to a plateau with a close to constant number of trigger points even if the certainty threshold increases further. At the far right side of the graph, the number of trigger points starts decreasing again, as even high certainty (supposedly true) trigger points are discarded. We expect to have an optimal certainty threshold in terms of F1 score (F1)[11] on the plateau to be found in the middle of the plot.

To evaluate whether QRS certainty behaves as expected and described, we need to test the approach proposed above on ECG data. In the following, we describe how evaluation is done.

3.5. Evaluation

When assessing certainty in QRS detection, situations have to be created, in which a QRS detector is to some degree uncertain about a detected QRS complex. With ANN-based QRS detectors, we can create such a situation by making the test data differ from the training data. We achieve this by using noisy test data, to make QRS detection harder.

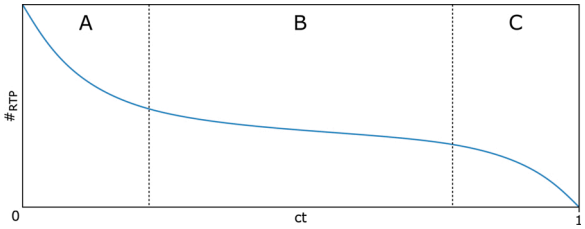

Fig. 13. Speculative figure showing the number of retained trigger points ($\#_{RTP}$) plotted subject to the certainty threshold (*ct*). Higher *ct* correspond to lower $\#_{RTP}$ since more trigger points are removed because their certainty is below the certainty threshold. This figure does not show real measurements but rather how the measurements (shown later in this work) are supposed to look like. In area A, low certainty (supposedly false) trigger points are included. In area B, we supposedly discard false trigger points while retaining true ones. In area C, the certainty threshold is sufficiently high to even discard trigger points with a rather high certainty, hence removing even supposedly true trigger points from the output.

[11] F1 is the harmonic mean of the PPV and Se, hence taking false positives and false negatives into account: $F1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{Se}}{\mathrm{PPV} + \mathrm{Se}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$.

Noisy test data are created using the ECG databases described in Section 2.2. We use clean ECGs from the `mitdb` and add noise provided by the `em` noise template in the `nstdb`. The exact method used for adding noise is described in Section 3.6.

On these noisy ECG data, we perform QRS detection as per our proposed method yielding trigger points[12] in conjunction with a certainty measure for each detection. On these trigger points, we apply the certainty threshold determination method described in Section 3.4. This means that we will plot the number of retained trigger points ($\#_{RTP}$) subject to the certainty threshold (*ct*) and verify that it, indeed, shows the shape described in Section 3.4 and Fig. 13. Furthermore, we will examine Se, PPV, and their harmonic mean F1 subject to *ct*. We will evaluate the following hypotheses:

- Low *ct* are associated with higher values for PPV and lower values for Se
- High *ct* are associated with higher values for Se and lower values for PPV
- Intermediate *ct* yield the highest F1

Furthermore, we will evaluate how increasing levels of noise influence certainty measures, hypothesising that certainty decreases with increasing levels of noise. In addition, we will compare naive certainty and adaptive certainty with respect to Se, PPV, and F1 using different discretization thresholds (*th*). With this evaluation, we want to find out whether varying plateau width in the rectified trigger signal are actually a problem. If this is the case, adaptive certainty should outperform naive certainty with respect to Se, PPV, and F1.

Apart from noise, we investigate how certainty is influenced by deviant beat types in Section 4.3, how certainty levels vary with different detectors in Section 4.4, and the influence of training parameters on certainty in Section 4.5.

3.6. Artificially adding noise to a signal

We want to evaluate how noise in the ECG signal affects QRS certainty and subsequently the performance of the QRS detector. We use the `mitdb` as source for low-noise (clean) ECG signals. The `nstdb` is used as a source for noise templates. To add noise to the clean signal, we perform the following steps:

1. Determine the signal power of the clean signal and the noise template.
2. Scale the samples of the noise template up or down to achieve a specific SNR.
3. Add samples of noise template to clean signal.

The SNR is defined in terms of signal power, as shown in Eq. (7). The power of an electrical signal s can be defined in terms of its root mean square amplitude (Eq. (8)). Using this, we can define SNR in terms of signal and noise samples (Eq. (9)). Assuming clean signal and noise template have the same length in terms of number of samples ($m = n$), we can simplify the equation by removing $\frac{1}{m}$ and $\frac{1}{n}$ (Eq. (10)).

$$\mathrm{SNR} = \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} \tag{7}$$

$$P_{\mathrm{s}} = \frac{1}{n}\sum_{i=0}^{n} \mathrm{s}(i)^2 \tag{8}$$

$$\mathrm{SNR} = \frac{\frac{1}{m}\sum_{i=0}^{m} \mathrm{signal}(i)^2}{\frac{1}{n}\sum_{i=0}^{n} \mathrm{noise}(i)^2} \tag{9}$$

$$\mathrm{SNR} = \frac{\sum_{i=0}^{n} \mathrm{signal}(i)^2}{\sum_{i=0}^{n} \mathrm{noise}(i)^2} \tag{10}$$

Since we do not want to describe the current SNR which we would get when adding the samples of signal and noise without scaling the noise, we define a target SNR (SNR_t) in Eq. (11) where k is the factor used for scaling each of the noise's samples. Eqs. (12)–(15) show, how to find k.

$$\mathrm{SNR}_t = \frac{\sum_{i=0}^{n} \mathrm{signal}(i)^2}{\sum_{i=0}^{n} (k \cdot \mathrm{noise}(i))^2} \tag{11}$$

$$\mathrm{SNR}_t = \frac{\sum_{i=0}^{n} \mathrm{signal}(i)^2}{k^2 \cdot \sum_{i=0}^{n} \mathrm{noise}(i)^2} \tag{12}$$

$$k^2 = \frac{\sum_{i=0}^{n} \mathrm{signal}(i)^2}{\mathrm{SNR}_t \cdot \sum_{i=0}^{n} \mathrm{noise}(i)^2} \tag{13}$$

$$k^2 = \frac{\mathrm{SNR}}{\mathrm{SNR}_t} \tag{14}$$

$$k = \sqrt{\frac{\mathrm{SNR}}{\mathrm{SNR}_t}} \tag{15}$$

Now we can scale up all samples of the noise template by multiplying them with k. Finally, we can add the samples of the scaled noise template to the clean signal to produce a noisy ECG signal with constant SNR. This noisy ECG signal can then, in turn, be used for evaluating how the certainty metric behaves when signal quality is impaired.

4. Results

In the following, we evaluate the certainty metrics proposed in this work. We show how certainty can be used to distinguish correctly detected QRS complexes from false detections. To perform this distinction, we have to find a suitable certainty threshold. This threshold is under the influence of external factors, such as noise. Higher levels of noise cause larger artefacts that show stronger resemblance to QRS complexes than smaller artefacts would do. Hence, there is not a single and always adequate certainty threshold, rather we have to determine the certainty threshold for each record separately.

4.1. Certainty threshold determination

For finding a certainty threshold suitable for the signal's level of noise, we proposed considering the number of detected trigger points ($\#_{RPT}$) subject to the certainty threshold (*ct*). Fig. 14 shows a plot of this relationship. As assumed, $\#_{RPT}$ decreases rapidly with increasing *ct* for small *ct*, then $\#_{RPT}$ reaches an area of low slope until, for large *ct*, $\#_{RPT}$ starts decreasing again with increasing *ct*. This holds even for higher levels of noise, as Fig. 15 shows. More noise causes $\#_{RPT}$ to be higher for small *ct*, due to a higher number of low-certainty trigger points caused

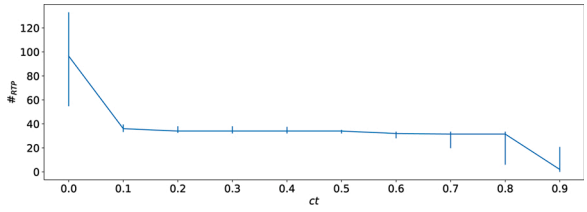

Fig. 14. The number of detected trigger points ($\#_{RPT}$) subject to the certainty threshold (*ct*) for records with SNR = 6*dB*. The line shows the median, the error bars show the 1st and 3rd quartile. As suspected in Section 3, there is an area of low slope in the middle of the plot. certainty thresholds forming this area of low slope are produce best QRS detection performance, as Fig. 16 shows.

[12] Candidates for QRS positions.

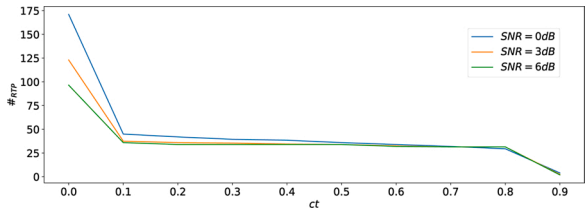

Fig. 15. Median $\#_{RPT}$ subject to ct for records with different signal-to-noise ratios. For all SNRs there is an area of low slope for intermediate ct values. For more extreme ct values ($ct < 0.1$ or $ct > 0.8$), the $\#_{RPT}$ decreases considerably with increasing ct.

by noise. Using the proposed method, values for ct between 0.1 and 0.8 seem to be reasonable.

To evaluate whether ct values in the area of low slope are actually corresponding to the best detector performance, we assess Se, PPV, and F1 subject to ct. Fig. 16 shows the results of this evaluation for the $\#_{RPT}/ct$ plot shown in Fig. 14. Fig. 17 shows the results of this evaluation for the $\#_{RPT}/ct$ plot shown in Fig. 15. These results are discussed in Section 5.1. The effect that increased levels of noise have on the certainty metric is discussed in Section 5.2.

4.2. Comparison of certainty metrics

In Section 3 we introduced two alternative definitions for certainty. Naive certainty (C) is defined in terms of the rectified trigger signal and adaptive certainty (C') is defined in terms of the expected plateau width. In the following, we compare C and C' under two different conditions. For the first comparison, we used a low discretization threshold for the trigger signal ($th = 0.05$). For the second comparison, we used a high discretization threshold for the trigger signal ($th = 0.5$).

Fig. 18 shows the $\#_{RPT}/ct$ plots of C and C' for $th = 0.05$. We can see that the plots only differ for high certainty thresholds ($ct > 0.8$) where adaptive certainty retains more detected trigger points (higher $\#_{RPT}$). Fig. 19 shows that with C' we achieve higher Se (and consequently higher F1) than with C for high certainty thresholds ($ct > 0.8$). Higher Se means more true positives in relation to false negatives. Consequently, C' reduces the number of false negatives compared to C.

For $th = 0.5$ we can see in Fig. 20 that the $\#_{RPT}/ct$ plot of C' shows fewer detected trigger points than C for equal certainty thresholds in the area of low slope ($ct \in [0.1, 0.9]$). Fig. 21 shows that C' shows higher PPVs for equal certainty thresholds than C. Consequently, F1 is also higher for C'. Higher PPVs mean that there are more true positives in relation to false positive. Consequently, C' reduces the number of false positives compared to C.

These results and the reasons for this behaviour are discussed in Section 5.3.

4.3. Impact of varying beat types on certainty

We trained a García-Berdonés detector that was modified to report certainty along with its detections on the `mitdb` recordings 100 to 104. These ECG recordings contain only negligible amounts of non-normal beats.[13] Afterwards, we used this pre-trained detector to detect QRS complexes on recording 106 that exhibits 520 premature ventricular contractions (PVCs) (non-normal beats) in its 2027 QRS complexes. Fig. 22 shows how certainty in PVCs is lower than certainty in normal beats. False positive detections, however, exhibit even lower certainty than PVCs.

To check whether the certainty values of (A) normal beats and PVCs and (B) PVCs and false detections are statistically significantly different, we performed a Mann-Whitney U test[14] with a significance level $\alpha = 0.05$. For (A) normal beats and PVCs we computed $p = 0.0381$ which indicates statistically significant difference. For (B) PVCs and false detections we computed $p = 4.93 \times 10^{-91}$ which also indicates statistically significant difference.

4.4. Certainty differences between detectors

We compare certainties across the three detectors discussed in this work, i.e. García-Berdonés detector, Šarlija detector, and Xiang detector. Therefor, we used the same procedure as in Section 4.3. We trained the detectors on `mitdb` recordings 100 to 104 with negligible amounts of non-normal beats and tested on recording 106 which contains approximately 25% non-normal beats. Using this approach, we evaluate both the detector's ability to recognise previously seen beat types and its ability to generalise to deviant shapes. In practice, it would make more sense to train the detector on a variety of beat types. However, we decided to use this approach in order to make detection harder and hence certainty differences more striking. Fig. 23 shows certainty differences across different detectors. Interestingly, the Xiang detector shows higher certainty values while both the García-Berdonés detector and the Šarlija detector exhibit lower certainty values on a similar level. Apparently, the two-part approach using an object-level CNN and a part-level CNN – which is the novelty in the Xiang detector – makes a bigger difference in terms of certainty than the use of a CNN instead over an MLP – which is the novelty of the Šarlija detector over the García-Berdonés detector.

4.5. Impact of training parameters

With ANN-based QRS detection, there is a large number of parameters that need to be set. Hyperparameter optimisation concerning parameters like learning rate, number of training epochs, etc. is a general problem in machine learning, especially with neural networks. The problem is already covered by other works [30] and there exist tools for optimising hyperparameters in different machine learning toolkits such as scikit-learn's hyper-parameter optimisers,[15] the Keras Tuner[16] and Tensorflow's HParams Dashboard.[17] Hence, we will not cover this topic here. A special parameter to ANN-based QRS detectors is the window size which is why we will investigate this parameter more closely.

As a general notion, larger windows provide more information to the ANN but are computationally more expensive and hence makes the detector slower. In contrast, smaller windows provide less information and make detection faster. In the García-Berdonés detector, the window size must be smaller than RR interval because otherwise the square pulses in the trigger signal would fuse and QRS detection would become impossible. With the Šarlija detector and the Xiang detector this problem is averted by distinguishing between window size and detection size (cf. Section 2.1).

An analysis on the García-Berdonés detector shows, that certainty increases with larger window sizes but quickly converges so that even larger window sizes provide no additional benefit at some point. This is to be seen in Fig. 24

5. Discussion

In Section 3 we showed that an ANN-based QRS detector reacts to

[13] https://archive.physionet.org/physiobank/database/html/mitdbdir/tables.htm.

[14] From the boxplot (Fig. 22) we cannot assume normal distribution which rules out the Student's t-test. Also, the samples are unpaired which rules of the Wilcoxon signed-rank test.

[15] https://scikit-learn.org/stable/modules/classes.html.

[16] https://www.tensorflow.org/tutorials/keras/keras_tuner.

[17] https://www.tensorflow.org/tensorboard/hyperparameter_tuning_with_hparams.

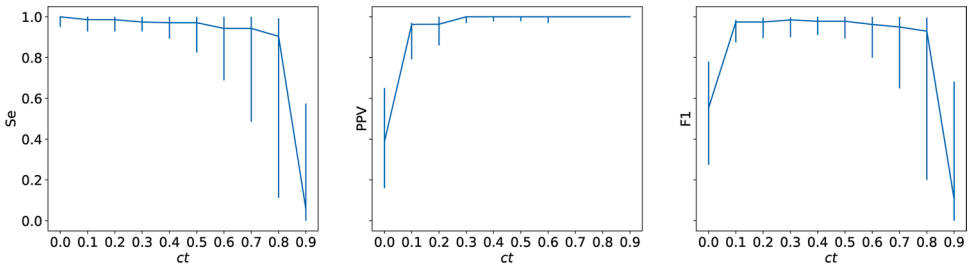

Fig. 16. Performance of an ANN-based QRS detector for different values for *ct*. SNR = 6 db. The line shows median, the error bars show the 25th and 75th percentiles. certainty thresholds between 0.1 and 0.8 yield are associated with high performance. This range of *ct* values is also the area of low slope in Fig. 14.

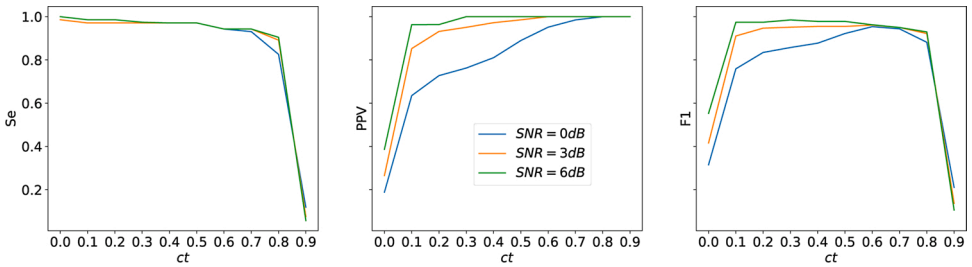

Fig. 17. Median performance of an ANN-based QRS detector for different values for *ct* and SNR. Although there are differences in performance, the overall shape of the plots is similar for all SNRs. Se decreases with increasing *ct* and has a cutoff at $ct \approx 0.8$ due to increasing numbers of false negatives. The PPV increases with increasing *ct* due to more false positives getting discarded. F1 is the harmonic mean of Se and PPV and has hence an inverted U shape.

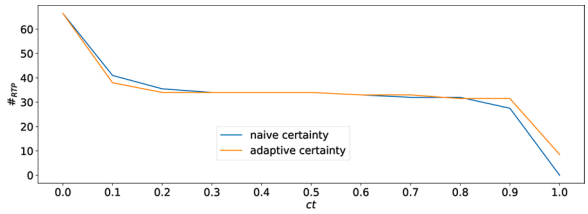

Fig. 18. $\#_{RPT}/ct$ plot comparison for adaptive certainty and naive certainty using a low discretization threshold of $th = 0.05$. Adaptive certainty shows higher numbers of detected trigger points ($\#_{RPT}$) for high certainty thresholds ($ct > 0.8$).

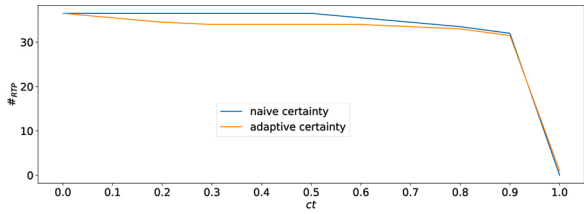

Fig. 20. $\#_{RPT}/ct$ plot comparison for adaptive certainty and naive certainty using a high discretization threshold of $th = 0.5$. Adaptive certainty shows lower numbers of detected trigger points ($\#_{RPT}$) for the area of low slope ($ct \in [0.1, 0.9]$).

hard-to-interpret ECG signals with distinct changes in the trigger signal. These changes are caused either by deviant beat types or by noise. To quantify these changes, we provided two definitions for a certainty metric, thus answering RQ1 (cf. Section 1.3).

Furthermore, we introduced a method for finding a suitable certainty threshold to distinguish between true and false QRS complex detections (Section 3.4), thus answering RQ2. The performance results of this method are presented in Section 4.1 and further discussed in Section 5.1.

Regarding RQ3, we examined how different factors such as increasing levels of noise (Sections 4.1 and 4.2), deviant beat types (Section 4.3), different detector architectures (Section 4.4), and training parameters (Section 4.5) influence certainty.

Especially the effects of increasing levels of noise are discussed in greater detail in Section 5.2, since this also has implications on the certainty threshold for rejecting false positives. Finally, in Section 5.3 we compare the two alternative definitions of certainty (naive certainty and adaptive certainty) we provided.

5.1. Certainty threshold evaluation

The graphs shown in Figs. 14 and 15 behave both as predicted in Section 3.4 (cf. Fig. 13). To evaluate whether *ct* values yielded by this approach actually correspond to the highest QRS detection performance, we plotted Se, PPV, and F1 subject to *ct* in Figs. 16 and 17 . These figures show that Se, PPV, and F1 respond to changes in *ct* as expected and described previously. There is an area of high values and a low slope

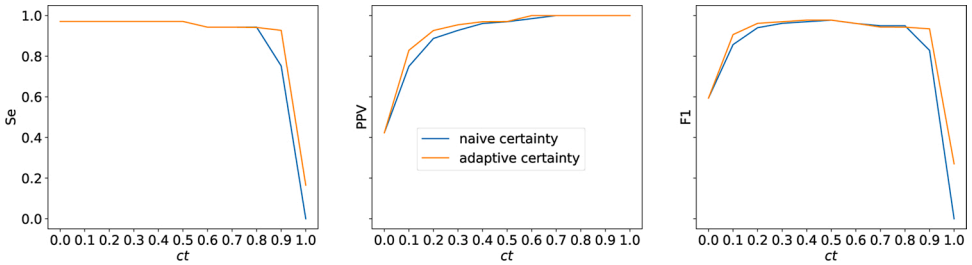

Fig. 19. Comparison of performance characteristics for adaptive certainty and naive certainty. Adaptive certainty shows higher Se for high certainty thresholds ($ct > 0.8$). Since F1 is the harmonic mean of Se and PPV and both certainties show the same performance in terms of PPV, F1 is also higher for adaptive certainty and $ct > 0.8$. This means that adaptive certainty produces fewer false negatives than naive certainty for high certainty thresholds.

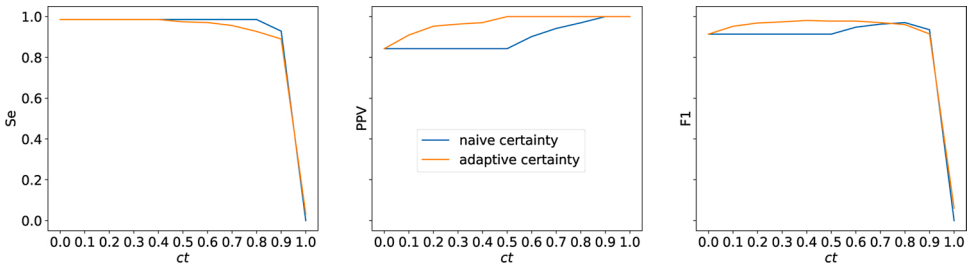

Fig. 21. Comparison of performance characteristics for adaptive certainty and naive certainty. Adaptive certainty shows higher PPVs. Since the F1 is the harmonic mean of Se (Se) and PPV and both certainties show the same performance in terms of Se, the F1 is also higher for adaptive certainty. This means that adaptive certainty produces fewer false positives than naive certainty.

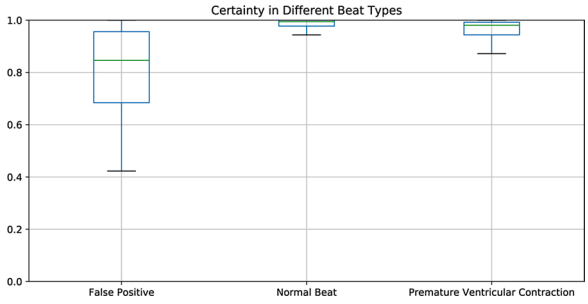

Fig. 22. Values of the certainty metric in different beat types. False detections exhibit the lowest certainty, followed by non-normal beat types such as PVCs. As expected, the highest certainty is to be found with normal beats since this is the kind of beat the detector was trained on.

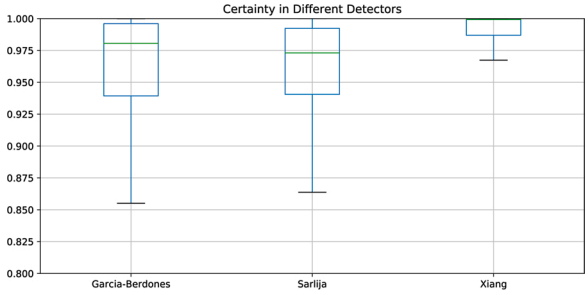

Fig. 23. Values of the certainty metric as produced by different detectors. The Xiang detector exhibits higher certainty values than both other detectors. Note that the y axis is scaled to $[0.8; 1.0]$.

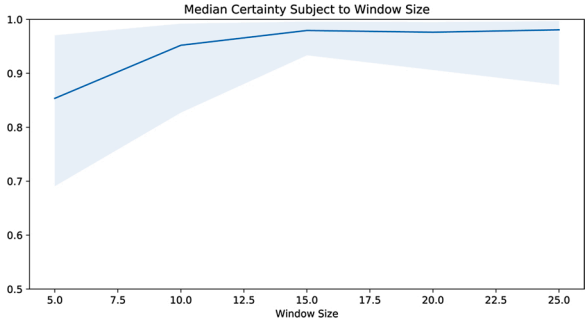

Fig. 24. Median certainty subject to window size as produced by a García-Berdonés detector. The light blue are represents the IQR. We can observe that certainty increases with increasing window size up to a point of convergence which is at 15 in this case.

in the F1[18] between $ct = 0.1$ and $ct = 0.8$. Thus, searching for areas of low slope in the $\#_{RPT}/ct$ plot is suitable for finding proper certainty thresholds.

5.2. Effects of noise

When increasing the level of noise we can make two observations, to be seen in Fig. 17:

1. The numbers of true positives and false negatives do not change severely with increasing levels of noise. Thus, Se curves are approximately the same for different levels of noise.
2. The number of false positives increases with increasing levels of noise. Therefore the curve of the PPV shows lower values with higher levels of noise.

We can conclude that with increasing levels of noise, the influence of false positives on the overall performance of the QRS detector increases. Hence, we have to choose higher values for ct to discard these false positives. This is supported by Fig. 17 showing that for higher levels of noise (lower SNR), a F1 value close to maximum is reached with higher ct only (i.e. $ct \approx 0.7$). In the corresponding $\#_{RPT}/ct$ plot (Fig. 15) this is also to be seen: Curves for higher levels of noise (e.g. blue curve, SNR = 0 dB) have higher slopes for lower ct values ($ct \in [0.1, 0.4]$) than curves for lower levels of noise (e.g. green curve, SNR = 6 dB).

5.3. Why does adaptive certainty outperform naive certainty?

Generating a rectified trigger signal from a flawed trigger signal is done by an algorithm expecting a discretization threshold as an additional parameter. For certainty assessment, low discretization thresholds are used to avoid prematurely rejecting plateaus. However, a too low threshold can create too wide plateaus for actual QRS complexes, leading to lower certainty when using C. This is shown in Fig. 25. Consequently, the plateau is rejected due to too low certainty, creating a false negative. Using C', the plateau width is predefined and fixed and thus the problem does not occur.

With higher discretization thresholds, we face the opposite situation. Plateaus caused by noise are narrow and hence only covering parts where the flawed trigger signal has high values. This is shown in Fig. 26. Consequently, the plateau's certainty is higher, and therefore the plateau is retained even though it is caused by noise, creating a false positive.

[18] F1 is the harmonic mean of Se and PPV. Thus, F1 can be considered a good metric for the overall performance in QRS detection, since it combines the measure for the occurrence of false positives (PPV) and the measure for the occurrence of false negatives (Se).

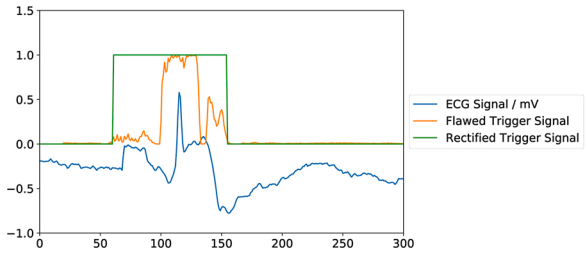

Fig. 25. Trigger signal rectification with a discretization threshold of $th = 0.05$. Low discretization thresholds are useful for making sure all plateaus in the flawed trigger signal are discovered. However, in this case the low discretization threshold in combination with ripple removal creates a too wide plateau (to be seen in the rectified trigger signal). Consequently, the plateau's certainty is lower than expected.

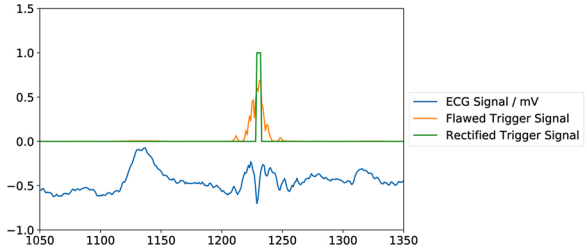

Fig. 26. Trigger signal rectification with a discretization threshold of $th = 0.5$. Higher discretization thresholds avoid the problem of too wide plateaus for actual QRS complexes. However, high discretization thresholds also make plateaus caused by noise narrower. Consequently, the plateau's certainty is higher than expected.

5.4. Conclusion

In this work, we defined the concept of certainty in ANN-based QRS detection. We examined one exemplary use case for certainty within the QRS detector itself: Distinguishing correctly detected QRS complexes from false detections. This helps in making sense from the trigger signal without having to rely heavily on arbitrarily set thresholds. However, improving the performance of the QRS detector itself is not the only possible use case. Detecting QRS complexes is not a self-purpose but rather a single step in a pipeline. From the knowledge of the QRS complex positions, we can derive metrics like heart rate which in turn can be used for triggering alarms in patient monitor or making diagnoses. With certainty assessment in the QRS detector, we cannot only derive the heart rate but also give an estimate on how reliable this information is, thus retaining more information for higher-level decisions (such as raising an alarm or not).

It has to be noticed that certainty is not only influenced by the shape of the QRS complex but also by the neural network and how it was trained. The more training and test data differ, the lower the certainty is. Hence, low certainty does not necessarily imply bad signal quality but rather that the neural network in use can hardly make sense from the signal.

5.5. Limitations and future work

We proposed the certainty metric as a SQI that is specific to the task of QRS detection with a specific detector. However, this approach still has some limitations that demand future work.

Different detectors. So far, we defined the certainty metric only in terms of ANN-based QRS detectors. This limits the practical applicability since ANN-based QRS detection is only one of many approaches [9]. Also, due to the window-based approach, QRS detection with ANNs introduces a time lag which further limits its applicability for real-time use cases such as patient monitoring. Other approaches to QRS detection need to be investigated in order to define certainty for a larger variety of detectors. To this end, future work needs to investigate how uncertainty manifests in different detectors and how it can be quantified.

Different signals. While this paper defines certainty in terms of QRS detection, i.e. heartbeat detection in ECG signals, other physiological signals can be used for beat detection as well. For example, there are ABP and PPG signals, where pulse waves can be detected in. Future work needs to investigate whether the concept of certainty can also be applied to beat detection in other physiological signals.

Usefulness in practice. In this work, we only investigated how the certainty metric can be used within the field of QRS detection itself. For example, to improve QRS detection by choosing a suitable certainty threshold. It remains to be investigated, how the certainty reported by the QRS detection can be utilised in later-on, for example for alarm generation in medical monitors. Further work has to be done in determining the potential of certainty for practical purposes, such as distinguishing real medical emergencies from false alarms due to bad signal quality.

Credit author statement

Jonas Chromik: Conceptualisation, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualisation.
Jossekin Beilharz: Conceptualisation, Resources, Supervision.
Lukas Pirl: Resources, Writing – Review & Editing, Supervision.
Bert Arnrich: Resources, Writing – Review & Editing, Supervision.
Andreas Polze: Supervision.

Declaration of Competing Interest

The authors report no declarations of interest.

References

[1] V.C. Scanlon, T. Sanders, Essentials of Anatomy and Physiology, 5th Edition, F.A. Davis Co, Philadelphia, 2007 oCLC: ocm68694088.
[2] G.S. Wagner, D.G. Strauss, Marriott's Practical Electrocardiography, 12th Edition, Lippincott Williams & Wilkins, Philadelphia, PA, 2013.
[3] G.D. Clifford, I. Silva, B. Moody, Q. Li, D. Kella, A. Shahin, T. Kooistra, D. Perry, R. G. Mark, The PhysioNet/computing in cardiology challenge 2015: reducing false arrhythmia alarms in the ICU, in: 2015 Computing in Cardiology Conference (CinC), IEEE, Nice, France, 2015, pp. 273–276, https://doi.org/10.1109/CIC.2015.7408639. http://ieeexplore.ieee.org/document/7408639/.
[4] M. Cvach, Monitor alarm fatigue: an integrative review, Biomed. Instrum. Technol. 46 (4) (2012) 268–277.
[5] B.J. Drew, P. Harris, J.K. Zègre-Hemsey, T. Mammone, D. Schindler, R. Salas-Boni, Y. Bai, A. Tinoco, Q. Ding, X. Hu, Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients, PLOS ONE 9 (10) (2014) e110274, https://doi.org/10.1371/journal.pone.0110274.
[6] G.B. Moody, W. Muldrow, R.G. Mark, A noise stress test for arrhythmia detectors, Comput. Cardiol. 11 (3) (1984) 381–384.
[7] W. Engelse, C. Zeelenberg, A single scan algorithm for qrs-detection and feature extraction, Comput. Cardiol. 6 (1979) (1979) 37–42.
[8] J. Pan, W.J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. BME-32 (3) (1985) 230–236, https://doi.org/10.1109/TBME.1985.325532.
[9] B. Kohler, C. Hennig, R. Orglmeister, The principles of software QRS detection, IEEE Eng. Med. Biol. Mag. 21 (1) (2002) 42–57, https://doi.org/10.1109/51.993193.
[10] J. Boston, G. Akyol, Using an uncertainty measure in a fuzzy QRS detector. Proceedings of the First Joint BMES/EMBS Conference. 1999 IEEE Engineering in Medicine and Biology 21st Annual Conference and the 1999 Annual Fall Meeting of the Biomedical Engineering Society (Cat. No.99CH37015), vol. 2, IEEE, Atlanta, GA, USA, 1999, p. 916, https://doi.org/10.1109/IEMBS.1999.804070.
[11] B. Settles, Active Learning Literature Survey, University of Wisconsin, Madison, Wisconsin, USA, 2010. Tech. Rep.
[12] C. Haritha, M. Ganesan, E.P. Sumesh, A survey on modern trends in ECG noise removal techniques. 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), IEEE, Nagercoil, India, 2016, pp. 1–7, https://doi.org/10.1109/ICCPCT.2016.7530192.
[13] N. Gambarotta, F. Aletti, G. Baselli, M. Ferrario, A review of methods for the signal quality assessment to improve reliability of heart rate and blood pressures derived parameters, Med. Biol. Eng. Comput. 54 (7) (2016) 1025–1035, https://doi.org/

10.1007/s11517-016-1453-5, place: Heidelberg Publisher: Springer Heidelberg WOS: 000379014500002.
[14] C. Daluwatte, L. Johannesen, L. Galeotti, J. Vicente, D.G. Strauss, C.G. Scully, Assessing effect of beat detector on detection dependent signal quality indices, 2016 Computing in Cardiology Conference (CinC) (2016) 921–924, iSSN: 2325-887X.
[15] G. Moody, B. Moody, I. Silva, Robust detection of heart beats in multimodal data: the PhysioNet/computing in cardiology challenge 2014, in: Computing in Cardiology, IEEE, IEEE, Cambridge, MA, USA, 7, 2014, p. 4.
[16] G.B. Moody, The PhysioNet/computing in cardiology challenge 2010: mind the gap, in: Computing in Cardiology, IEEE, Cambridge, MA, USA, 26, 2010, p. 4.
[17] B.M. Mathunjwa, Y.-T. Lin, C.-H. Lin, M.F. Abbod, J.-S. Shieh, Ecg arrhythmia classification by using a recurrence plot and convolutional neural network, Biomed. Signal Process. Control 64 (2021) 102262.
[18] G. Petmezas, K. Haris, L. Stefanopoulos, V. Kilintzis, A. Tzavelis, J.A. Rogers, A. K. Katsaggelos, N. Maglaveras, Automated atrial fibrillation detection using a hybrid cnn-lstm network on imbalanced ecg datasets, Biomed. Signal Process. Control 63 (2021) 102194.
[19] G.D. Clifford, C. Liu, B. Moody, H.L. Li-wei, I. Silva, Q. Li, A. Johnson, R.G. Mark, Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017, 2017 Computing in Cardiology (CinC) (2017) 1–4.
[20] Y. Deng, Z. Gao, S. Xu, P. Ren, Y. Wen, Y. Mao, Z. Li, St-net: synthetic ecg tracings for diagnosing various cardiovascular diseases, Biomed. Signal Process. Control 61 (2020) 101997.
[21] F. Ertuğrul, E. Acar, E. Aldemir, A. Öztekin, Automatic diagnosis of cardiovascular disorders by sub images of the ecg signal using multi-feature extraction methods and randomized neural network, Biomed. Signal Process. Control 64 (2021) 102260.
[22] E.A.P. Alday, A. Gu, A.J. Shah, C. Robichaux, A.-K.I. Wong, C. Liu, F. Liu, A.B. Rad, A. Elola, S. Seyedi, et al., Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020, Physiol. Meas. 41 (12) (2020) 124003.
[23] C. García-Berdonés, J. Narváez, U. Fernández, F. Sandoval, A new QRS detector based on neural network, in: J. Mira, R. Moreno-Díaz, J. Cabestany (Eds.), Biological and Artificial Computation: From Neuroscience to Technology, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 1997, pp. 1260–1269.
[24] M. Šarlija, F. Jurišić, S. Popović, A convolutional neural network based approach to QRS detection, Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis (2017) 121–125, https://doi.org/10.1109/ISPA.2017.8073581.
[25] Y. Xiang, Z. Lin, J. Meng, Automatic QRS complex detection using two-level convolutional neural network, BioMed. Eng. OnLine 17 (January) (2018), https://doi.org/10.1186/s12938-018-0441-4.
[26] F. Chollet, et al., Keras, 2015. https://keras.io.
[27] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale Machine Learning on Heterogeneous Systems, Software Available From tensorflow.org, 2015. https://www.tensorflow.org/.
[28] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J. E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (June (23)) (2000) e215–e220.
[29] G.B. Moody, R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. Mag.: Q. Mag. Eng. Med. Biol. Soc. 20 (May–June (3)) (2001) 45–50.
[30] M. Claesen, B. De Moor, Hyperparameter Search in Machine Learning, 2015 (arXiv preprint), arXiv:1502.02127.

Glossary

`bw`: baseline wander
`em`: electrode motion
`ma`: muscle artefacts
`mitdb`: MIT-BIH Arrhythmia Database
`nsrdb`: MIT-BIH Normal Sinus Rhythm Database
`nstdb`: MIT-BIH Noise Stress Test Database
ABP: arterial blood pressure
adaptive certainty: c
AFib: atrial fibrillation
ANN: artificial neural network
certainty: m
CNN: convolutional neural network
CVD: cardiovascular disease
ECG: electrocardiogram
expected plateau width: The width of the square pulse the ANN classifier should produce under ideal circumstances. This width is equal to the width of the window that is used as input for the ANN. Denoted with w_e
F1: F1 score
flawed trigger signal: A trigger signal that does not exhibit a square pulse shape. Intermediate values (between 0 and 1) and ripple are to be found. Denoted with t_f
HR: heart rate
IQR: interquartile range
LSTM: long short-term memory neural network
MLP: multilayer perceptron
naive certainty: Certainty metric defined via the flawed trigger signal and the rectified trigger signal. Denoted with C
PPG: photoplethysmography
PPV: positive predictive value
PVC: premature ventricular contraction
rectified trigger signal: Square pulse signal generated from the flawed trigger signal by means of discretization and ripple removal. Denoted with t_r
Se: sensitivity
SNR: signal-to-noise ratio
SQI: signal quality indicator
trigger point: a
trigger signal: i
window: s