IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement

Physiol. Meas. **37** (2016) E5–E23 doi:10.1088/0967-3334/37/8/E5

Editorial

False alarm reduction in critical care

Abstract

High false alarm rates in the ICU decrease quality of care by slowing staff response times while increasing patient delirium through noise pollution. The 2015 PhysioNet/Computing in Cardiology Challenge provides a set of 1250 multi-parameter ICU data segments associated with critical arrhythmia alarms, and challenges the general research community to address the issue of false alarm suppression using all available signals. Each data segment was 5 minutes long (for real time analysis), ending at the time of the alarm. For retrospective analysis, we provided a further 30 seconds of data after the alarm was triggered.

A total of 750 data segments were made available for training and 500 were held back for testing. Each alarm was reviewed by expert annotators, at least two of whom agreed that the alarm was either true or false. Challenge participants were invited to submit a complete, working algorithm to distinguish true from false alarms, and received a score based on their program's performance on the hidden test set. This score was based on the percentage of alarms correct, but with a penalty that weights the suppression of true alarms five times more heavily than acceptance of false alarms.

We provided three example entries based on well-known, open source signal processing algorithms, to serve as a basis for comparison and as a starting point for participants to develop their own code. A total of 38 teams submitted a total of 215 entries in this year's Challenge.

This editorial reviews the background issues for this challenge, the design of the challenge itself, the key achievements, and the follow-up research generated as a result of the Challenge, published in the concurrent special issue of *Physiological Measurement*. Additionally we make some recommendations for future changes in the field of patient monitoring as a result of the Challenge.

Keywords: critical care, false alarm reduction, arrhythmia

(Some figures may appear in colour only in the online journal)

1. Introduction

During the last decade, over a period of seven years, intensive care unit (ICU) admissions at US hospitals increased by 48.8% with a mean biennial increase of 14.2%. By comparison, overall emergency department (ED) visits increased by 5.8% biennially. In absolute terms, admissions jumped from 2.79 million in 2002–2003 to 4.14 million in 2008–2009, according to data from the National Hospital Ambulatory Care Survey (Mullins *et al* 2013). The three most common diagnoses for ICU admissions were unspecified chest pain, congestive heart failure, and pneumonia. Utilization rates of most tests and services delivered to patients admitted to the ICU from the ED increased, with the largest increase occurring in computed tomography (CT) and magnetic resonance imaging (MRI), which increased from 16.8% in 2002/2003 to 37.4% in 2008/2009, a 6.9% mean biennial increase. These findings suggested emergency physicians were sending more patients on to the ICU. The increase might be the result of an older, sicker population that needs more care (Mullins *et al* 2013).

ICU patients require a high level of acute care, with numerous bedside monitors which are continuously measuring both invasive and non-invasive variables. These monitors provide synchronous waveforms with both independent and complementary information. Huge ICU databases are therefore becoming available, and include parameters such as the electrocardiogram (ECG), the photoplethysmogram (PPG), the arterial blood pressure (ABP) waveform, and respiratory effort. In clinical practice these signals are processed individually to trigger an alarm when a derived parameter (such as heart rate) exceeds a pre-defined range. These alarms are frequently false alarms (FAs) and account for a large majority of all alarms generated in the ICU (Chambrin *et al* 1999).

The high rate of false alarms significantly burdens clinical staff, which can lead to decreased quality of care (Donchin and Seagull 2002, Imhoff and Kuhls 2006), impacting both the patient and the clinical staff through noise disturbances, desensitization to warnings, slowing of response times (Chambrin 2001) and missed true alarms (Allen and Murray 1996, Chambrin 2001, Hug *et al* 2011). ICU alarms produce sound intensities above 80 dB that can lead to sleep deprivation (Meyer *et al* 1994, Chambrin 2001, Parthasarathy and Tobin 2004), inferior sleep structure (Slevin *et al* 2000, Johnson 2001), stress for both patients and staff (Cropp *et al* 1994, Novaes *et al* 1997, Topf and Thompson 2001, Morrison *et al* 2003) and depressed immune systems (Berg 2001). There are also indications that the incidence of re-hospitalization is lower if disruptive noise levels are decreased during a patient's stay (Hagerman *et al* 2005). Furthermore, such disruptions have been shown to have an important effect on recovery and length of stay (Cropp *et al* 1994, Donchin and Seagull 2002). In particular, cortisol levels have been shown to be elevated (reflecting increased stress) (Topf and Thompson 2001, Morrison *et al* 2003), and sleep disruption has been shown to lead to longer stays in the ICU (Parthasarathy and Tobin 2004). ICU false alarm (FA) rates as high as 90% (Aboukhalil *et al* 2008) have been reported, with between 6% and 40% of ICU alarms having been shown to be true but clinically insignificant (requiring no immediate action) (Lawless 1994). In fact, only 2% to 9% of alarms have been found to be important for patient management (Tsien and Fackler 1997). In response to this, thresholds and filter settings for alarms are often manipulated on a case-by-case basis in response to an individual clinical user's preferences (to reduce annoyance), which may well be sub optimal in terms of the trade off between true and false alarms (Mullins *et al* 2013).

In the 2015 PhysioNet/Computing in Cardiology Challenge (Clifford *et al* 2015) (http://physionet.org/challenge/2015), we aimed to address the problem of high false alarm rates by encouraging the development of new algorithms to improve the specificity of ICU alarms. In this Challenge, we focused on five types of life-threatening arrhythmia events, which we defined as follows:

Asystole (ASY): No heartbeats for a period of 4 s or more.

Extreme bradycardia (EBR): Heart rate lower than 40 beats per minute; fewer than five beats occur within a period of 6 s.

Extreme tachycardia (ETC): Heart rate higher than 140 beats per minute; more than 17 beats occur within a period of 6.85 s.

Ventricular tachycardia (VTA): Five or more consecutive ventricular beats within a period of 2.4 s (a rate of 100 per minute)

Ventricular fibrillation or flutter (VFB): The heart exhibits a rapid fibrillatory, flutter, or oscillatory waveform for at least 4 s.

Participants in the Challenge were given samples of ICU patient waveforms that were identified by the bedside monitor as falling into one of the above categories, and were tasked with devising an algorithm to determine which of these alarms represented true arrhythmias, and which were caused by other factors (such as noise, patient movement, leads falling off, or mis-identification of ECG features on the part of the monitor.)

The Challenge was divided into two events. Event 1 was a simulation of the *real-time* alarm suppression problem: the algorithm needed to determine whether the alarm was true or false based solely on the information available before the alarm was first triggered. In Event 2, algorithms were also able to see 30 seconds worth of waveform data following the time of the alarm, and could use this information to *retrospectively* classify the alarm as true or false. The development of an algorithm that could reliably solve either of these problems would be a major step forward in patient care.

2. Example algorithms

Key to rhythm detection is accurate heart rate estimation. Several ECG R-peak detection algorithms are freely available, several of which were used in the Challenge example entries.

eplimited (available at www.eplimited.com) (Hamilton and Tompkins 1986), which used digital filtering and a group of decision rules.

sqrs (available on PhysioNet (Goldberger *et al* 2000)) (Engelse and Zeelenberg 1979), which uses a single scan of the sampled data and combines digital filter preprocessing with a detector and feature extractor based on dynamically adjusted slope and timing criteria.

wqrs (available on PhysioNet) (Zong *et al* 2003b), which is based on the length transform.

gqrs (available on PhysioNet), which consists of a QRS matched filter with a custom built set of heuristics (such as search back).

coqrs (Nygårds and Sörnmo 1983, Clifford 2002, Oster *et al* 2013) based on the peak energy (no search back).

jqrs (Behar *et al* 2014a, 2014b) consists of a window-based peak energy detector but with replacement of the original band-pass filter with a QRS matched filter (Mexican hat) and an additional heuristic ensuring no detection were made during flat lines.

Detection of the onset of the pulses in the ABP and PPG signals can provide further information on rhythm and rate. An open-source algorithm, *wabp* (Zong *et al* 2003a), is available from PhysioNet. The algorithm consists of three components: (1) a low-pass filter which is to suppress high frequency noise that might affect the onset detection; (2) a windowed and weighted slope sum function (SSF) which is to enhance the up-slope of the pulse and to suppress the remainder of the pressure wave; (3) a decision rule which allows for detection of each SSF pulse onset.

We provided three example Challenge entries, based on these and other open-source algorithms, and implemented in various programming languages, to serve as a basis on which participants could develop their own code.

The simplest example entry (#1) used *wabp* and *gqrs*, along with the *gqfuse* tool (available on PhysioNet), to analyze all available signals and select the most stable sequence of RR intervals, in order to detect asystole, bradycardia, and tachycardia. To detect the onset of VFB, this entry analyzed the ECG and pulsatile signals separately (using *gqfuse* for each), and searched for a 10 s interval where the QRS rate and pulse rate were equal, followed by a 3 s interval in which the QRS rate increased by at least 25% and the pulse rate decreased by at least 75%. This entry did not attempt to detect VT.

A second example entry (#2) written in MATLAB used *wabp* to detect the beats and used *jSQI* (Sun *et al* 2006) and a template matching SQI (Li and Clifford 2012) to estimate the signal quality from ABP and PPG channels. For the ECG, an agreement level of two R-peak detectors (*ph gqrs* and *ph coqrs*) in a 10 s window, evaluated every second, known as *bSQI*, was used (Behar *et al* 2013).

Finally, the third sample entry (#3) was provided for Octave users, with functions from the WFDB toolbox for Octave/MATLAB Silva and Moody (2014). This sample entry ran three QRS detectors from the WFDB toolbox: *wqrs* on signal 1, *sqrs* on signals 1 and 2, and *gqrs* on signals 1 and 2. The results of the QRS detectors were then used to compute three tachograms. A decision was made on the veracity of the alarm based on the average pair-wise correlation between the tachograms 30 s prior to the alarm (a threshold was set arbitrarily based on the training data).

2.1. Signal quality

Signal quality indices (SQIs), which assess the signal quality or noise levels of the signals, can be extracted from the waveforms and used as weighting factors to allow for varying trust levels in the source data. The ECG signal quality has been extensively studied (Li *et al* 2008, 2014b, Clifford *et al* 2012, Behar *et al* 2013). For the benchmark algorithms, an agreement level of two R-peak detectors in a 10 s window, evaluated every second, known as *bSQI*, was used. Intuitively, the presence of noise and artifacts will lower the agreement level between two semi-independent detectors. The *bSQI* was recently successfully used on a database with pathological rhythms (Li *et al* 2008, Behar *et al* 2013). The ABP signal quality was evaluated using an open-source algorithm (Sun *et al* 2006) which flags a signal as bad quality if derived parameters from a blood pressure wave are not in reasonable physiological ranges. The PPG signal quality was also evaluated (Li and Clifford 2012) which matches a running PPG template with the pulsatile beat by dynamic time warping, simple matching, linear resampling matching and a clipping detection. When the signal quality was equal or greater than 0.9 and the corresponding HR or beat-to-beat interval derived from either the ABP or PPG did not surpass a predefined HR threshold (4 s for ASY, 40 bpm for EBR, 140 bpm for ETC, 100 bpm for VTA and 250 bpm for VFB), the alarm was suppressed as a false alarm.

It should be noted that no ECG signal quality metrics were used in our benchmark algorithms, although previous studies using the agreement of beat detectors for signal quality estimation have shown great promise in this area (Behar *et al* 2013). We also note that no ECG-based rhythm detection was used, although various open source algorithms were made available to the Challenge participants (Li *et al* 2014a).

2.2. Voting algorithms

We also implemented a voting approach to combine together varying numbers of algorithms. A simple unweighted voting of the *N* best performing final entries, ranked by their score on the training data (to prevent overfitting on the test scores, was implemented). *N* was varied from 1 to 37 with tied, absent or no vote was treated as ‘true’. In other words, a forward selection approach was used to select which algorithms should be combined.

3. Challenge data

Data for the Challenge consisted of waveform recordings from ICU patients in four hospitals in the USA and Europe, representing three major manufacturers of ICU monitoring equipment. For each arrhythmia alarm matching our selection criteria, we collected all available multiparameter waveforms (including at least five minutes of data before and after each alarm), as well as the alarm messages themselves, and any other status messages reported by the monitor. If possible, we also collected a list of the fiducial points and types of beats that were detected by the monitor; in some cases, the monitor did not provide this information. All of the signals were filtered in order to remove spectral characteristics that might identify the manufacturer or the country of origin. They were then resampled to 250 Hz and scaled to a 16-bit range. The specific names of the various alarm annotations were also normalized to anonymize the data.

3.1. Expert labeling

To build the 'gold standard' list of true and false alarms, a team of experts visually inspected the waveform record at the time of each alarm. Each annotator worked independently and was assigned a randomized list of patients to review. For each alarm, the annotator was initially shown 15 s of waveforms prior to the alarm and 5 s after it, but could resize and scroll the window in order to examine earlier and later portions of the record. If possible, the monitor-computed beat labels were also displayed.

After examining the alarm label and surrounding waveforms, the annotator was asked to press one of four buttons: *true*, *false*, *reject*, or *uncertain*. The *reject* label was used for records that were clearly fallacious (usually due to bugs in the monitor's data-exporting interface.) In order for an alarm to be included in the Challenge data set, it had to be independently reviewed by at least two annotators of whom a two-thirds majority had to agree that the alarm was either *true* or *false*.

3.2. Training and test data

From the set of 1564 alarms meeting all of the above criteria, we randomly picked 1250 to serve as training and test data for the Challenge (see table 1). The distribution of alarms was chosen to reflect the distribution of alarm types in the original data set (17% ASY, 11% EBR, 17% ETC, 47% VTA, 7% VFB) as well as to maintain the approximate true-to-false ratio for each alarm type. No single patient appeared in both the training and test sets, and no single manufacturer or hospital made up more than half of the records in either set.

Up to four signals were selected from each record: two ECG leads and up to two other signals including ABP, PPG, or respiration. The public training set consisted of 375 'short' records, containing only the five minutes leading up to the alarm, and 375 'long' records, containing a further 30 s after the alarm. The hidden test set consisted of 250 'short' records (used only for Event (1)) and 250 'long' records (used for both events). Each record was labeled with the alarm type, and in the case of the training set, whether the alarm was true or false. The records did not include the monitor-computed beat fiducial points or heart rate.

4. Scoring

Participants were asked to submit their entries in the form of a compressed archive that included everything needed to compile and run their program on a GNU/Linux system, together with

Table 1. Types of alarms and signals used in the challenge.

	Training ($N = 750$)		Test ($N = 500$)	
	False	True	False	True
ASY	100	20	90	12
EBR	45	45	38	26
ETC	8	131	5	68
VTA	253	90	176	45
VFB	52	6	34	6
PPG	227	178	158	83
ABP	59	63	58	39
Both	172	51	127	35
Total	458	292	343	157

Note. Each of the *N* records included two ECG channels.

the results that they expected their program to produce for the records in the public training set. When an entry was uploaded, the scoring system would first attempt to compile the program and run it over a randomly selected subset of the training set; if this did not produce the expected results, evaluation stopped and the error messages were sent back to the submitter.

Once the program was successfully compiled and validated, it was then invoked for each record in the test set.[1] If the program failed to produce output for a given record, it was treated as if it had classified that alarm as true.

For each category, the entry's score was computed based on the number of *true positives* (true alarms classified as true), *false positives* (false alarms classified as true), *true negatives*, and *false negatives*. The scoring function was designed to treat false negatives—genuinely life-threatening events that the program considered unimportant—especially harshly, and was defined as:

$$\text{score} = \frac{100 \cdot (TP + TN)}{TP + TN + FP + 5 \cdot FN}$$

5. Results of the Challenge

A total of 29 closed-source entries and 215 open-source entries were submitted in the Challenge in 2015. Table 2 provides a breakdown of the top scoring entries. A different contestant ranked highest in each separate alarm category, indicating that there was no best general algorithm. Interestingly, a simple majority vote of all the 38 competitors' final entries gave scores of 60.15 in the real-time event and 62.41 in the retrospective event. These moderate performances, well below the top 10 algorithms, indicating that simple voting schemes do no yield an improved performance in this context, since the performance tail is long. A voting algorithm using the $N = 13$ best performing final entries ranked by their score on the training data, provided the highest scores in both event 1 (84.26) and event 2 (87.04), although $N = 11$ was sufficient to beat the best performance in either event. Figure 1 illustrates the performance of a simple voting approach for both the retrospective and prospective parts of the Challenge. Note that performance only degrades above 13 algorithms.

[1] For the 250 'long' records, the program was invoked twice: once with the full record as input, and once with a truncated version.

Table 2. Final scores for the top 9 entrants in both events (real-time and retrospective) ranked by overall real-time score, the three example algorithms provided and a voting approach.

	Event 1 (Real-time)			Event 2 (Retrospective)		
Entrant	TPR (%)	TNR (%)	Score	TPR (%)	TNR (%)	Score
Pleššinger *et al* (2015)	92	88	**81.39**	95	88	84.96
Kalidas and Tamil (2015)	**94**	82	79.44	94	86	81.85
Krasteva *et al* (2015b)[a]	93	83	79.41[a]	93	84	79.56[a]
Couto *et al* (2015)	89	**91**	79.02	88	**92**	78.28
Fallet *et al* (2015)	**94**	77	76.11	**99**	80	**85.04**
Hoog Antink and Leonhardt (2015)	93	77	75.55	90	82	75.18
Eerikäinen *et al* (2015)	90	82	75.54	89	85	75.52
Ansari *et al* (2015)	89	84	74.48	89	87	76.57
Liu *et al* (2015)	89	79	71.68	93	78	75.91
Example algorithm 1	76	44	41.41	73	46	40.83
Example algorithm 2	86	38	45.07	84	38	44.37
Example algorithm 3	64	76	45.59	61	77	47.35
Voting algorithm ($N = 11$)	94	87	82.78	94	93	86.67
Voting algorithm ($N = 13$)	94	90	84.26	94	94	87.04

[a] Denotes an unofficial ('closed-source') entry. Underlined scores are the highest unofficial scores in the table.

Note. Best performances of competition entrants are in bold. TPR = fraction of true alarms correctly classified; TNR = fraction of false alarms correctly classified.

6. Review of articles in the special issue

A total of 13 articles were reviewed and revised in time to be accepted for this special issue. Most authors had originally entered the Challenge, and submitted updated versions of their algorithms, which should be made available by the authors through their open source licenses. The top reported results on the hidden test set for each alarm type were: ASY: 97.4% (Pleššinger *et al* 2016), EBR: 93.8% (Krasteva *et al* 2015a), ETC: 100.0% (Hoog Antink *et al* 2016), VFB: 88.7% (Rodrigues and Couto 2016), and VTA: 76.7% (Kalidas and Tamil 2016), yielding an average best of 91.2%.

Each algorithm published in this issue is reviewed below according to four standard stages of algorithm function:

1. Pre-processing and signal conditioning
2. Beat detection
3. Beat classification
4. Alarm classification

The purpose of this standardized summary is to glimpse at a myriad of advanced approaches used by the competitors in a format that allows the reader to quickly identify both the commonalities and the originality of all the approaches. Finally, the last two articles in this review (Daluwatte *et al* 2016, Tsimenidis and Murray 2016) did not attempt to reduce the number of false alarms, but rather provide some useful insights into the relationship between signal quality metrics and false alarm rates.

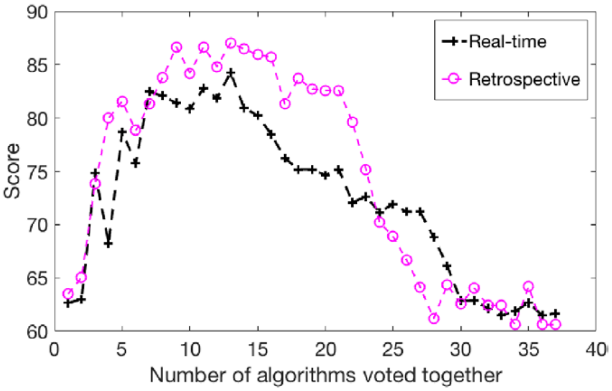

Figure 1. Performance of voting algorithms as a function of number of algorithms for both the real time and retrospective events. Algorithms were chosen by ranking them in descending order of score on the training data, and the test data score was reported (to prevent over-estimation of the score). Equal weights were given to all algorithms and a tied, absent or no vote was treated as 'true'.

6.1. Ansari et al (2016)

Ansari *et al* (2016) proposed an algorithm that uses several beat detectors within each channel, followed by beat classification, and heuristics to determine the veracity of the alarm. The proposed algorithm operates on 16 s of worth of data prior to the alarm. The algorithm achieved a performance accuracy on the final test data-set of ASY: 86.4%, EBR: 79%, ETC: 93.9%, VFB: 61% VTA: 67.6%, yielding a total average of: 76.2%.

6.1.1. Preprocessing. The preprocessing steps consisted of re-sampling the signals to 125 Hz. The ECG signals were band-pass filtered between 0.5–40 Hz, while the pulsatile signals were band-pass filtered between 0.5–10 Hz. Baseline and trend estimation and subtraction was accomplished with a 250 point median filter. The authors also removed pacemaker activity by thresholding on the peak amplitude.

6.1.2. Beat detection. Ansari *et al* (2016) implemented 7 different QRS detectors for each ECG signal, and 3 peak detectors for each of the pressure signals. The fiducial points for all peaks were re-aligned by picking the maximum within 50 ms of the detected beat for ECG signals, and the maximum within 50 ms before or 1 s after the detected beat for the ABP or PPG signals. The outputs of all the 20 beat detectors were then fused by adding their binary outputs (with at least 1 beat under AS, at least 2 for other alarms).

6.1.3. Beat classification. ECG beat classification was performed for the VFB and VTA alarms only. The beat classifier was a decision tree that utilized features derived from the Stockwell Transform on a 200 ms window.

6.1.4. Alarm classification. A decision tree classifier was trained with five fold cross validation in order to determine the veracity of a beat. The final decision regarding the alarm veracity was made based on a set of heuristics.

Table 3. Performances of the competitors in both events (real-time and retrospective) ranked by overall real-time score during the follow-up phase (Spring 2016).

	Real-time			Retrospective		
Authors	TPR (%)	TNR (%)	Score	TPR (%)	TNR (%)	Score
Plešsinger *et al* (2016)	93	**87**	**81.62 (81.39)**	95	**88**	84.96
Krasteva *et al* (2016)	92	87	80.07 (79.41[a])	93	88	81.75 (79.56[a])
Kalidas and Tamil (2016)	94	82	79.44	94	86	80.29 (81.85)
Hoog Antink *et al* (2016)	**95**	78	78.20 (75.55)	93	76	74.45 (75.18)
Eerikäinen *et al* (2016)	93	80	77.39 (75.54)	95	83	81.58 (75.52)
Fallet *et al* (2016)	**95**	76	77.07 (76.11)	**99**	80	**85.04**
Ansari *et al* (2016)	89	85	76.23 (74.48)	88	84	73.40 (76.57)
Rodrigues and Couto (2016)	92	78	74.28 (79.02)	92	78	74.46 (78.28)
Liu *et al* (2016)	89	79	71.68	93	78	75.91
Sadr *et al* (2016)	**95**	65	69.92	98	66	74.03
Tsimenidis and Murray (2016)	92	66	67.88	92	69	68.71
Daluwatte *et al* (2016)			—			—
Zong *et al* (2016)			—			—

[a] Denotes an unofficial ('closed-source') entry.

Note. Best performances are in bold. TPR = fraction of true alarms correctly classified; TNR = fraction of false alarms correctly classified. Numbers in the parentheses are the score of competition entrants if they are different with those in the follow-up phase.

6.2. Eerikäinen et al (2016)

Eerikäinen *et al* (2016) produced an algorithm that achieved a performance accuracy on the final test data-set of ASY: 89.2%, EBR: 71.5%, ETC: 99.1%, VFB: 81.8% VTA: 68.1%, yielding a total average of: 77.3% and a retrospective average of 81.5%.

6.2.1. Preprocessing. All signals were down-sampled to 125 Hz and the processing window length was optimized for each arrhythmia type (varying from 14 to 16 s prior to the alarm). Noise levels were estimated based on the power estimated from the regions in-between beats.

6.2.2. Beat detection. Beat detection on the ECG waveforms were performed using a QRS detector based on wavelets and auto-regressive modeling of the R-peak (Rooijakkers *et al* 2012). The pulsatile peaks were detected via the open source detector *wabp*.

6.2.3. Alarm classification. A random forest classifier was trained for each of the five different types of alarms. The technique focused on comparing pairs of beats. Two beats were considered a match if they were within 100 ms of each other. Delays across channels were compensated for if the standard deviation of 10 consecutive beats was less than 5% of the mean delay. For the VTA and VFB alarms, only the F1 statistic between ECG leads was used, in addition to spectral purity indexes. An alarm with an F1 equal to zero was identified to be false.

6.3. Fallet et al (2016)

Fallet *et al* (2016) proposed an algorithm that detects beats in the ECG sand the pulsatile signals, provided their signal quality is good. The authors also use a spectral purity metric to aid on the classification of VTA and VFB alarms. The algorithm achieved a performance accuracy on the final test data-set was ASY: 84.2%, EBR: 82.4%, ETC: 86.9%, VFB: 87.1% VTA: 72.7%, yielding a total average of: 77.07% and an average on the retroactive category of 85.0%.

6.3.1. Preprocessing. The preprocessing stage for this algorithm consisted of 50 Hz power line noise removal. For the calculations of spectral purity indexes, the signal was down-sampled to 35 Hz and a 5-sample moving average filter was applied. The signal quality for the pulsatile waveforms was estimated through the *ppgSQI* and *jSQI* methods (Clifford *et al* 2015).

6.3.2. Beat detection. The QRS component of the ECG signal was detected through a morphological analysis approach with an adaptive approach from Yazdani and Vesin (2014). Beat detection on the pulsatile signals was performed using the algorithm proposed by Arberet *et al* (2013). The heart rate time series was then derived through a multi-channel oscillator based adaptive frequency tracking algorithm.

6.3.3. Beat classification. The spectral purity index (Goncharova and Barlow 1990, Sörnmo and Laguna 2005) was used a feature to distinguish between normal, ventricular tachycardia, ventricular flutter/fibrillatory arrhythmia (the index was expected to be higher for abnormal rhythms).

6.3.4. Alarm classification. A set of heuristics rules was developed for the final alarm classification. In the case of the ASYS alarm, the algorithm applied majority voting based on the heart rate series from individual ECG and pulsatile channels. The pulsatile channels were only used if the quality was above a certain threshold. A linear discriminant analysis classifier was used for the retrospective event to corroborate the ECG output, but again, only if the pulsatile signal quality was sufficiently high. If the pulsatile quality was low, a set of heuristic thresholds was applied to the minimum heart rate from the last five consecutive beats using 16 s before and 5 s after the alarm. The extreme tachycardia alarm only used pulsatile waveforms: if the quality was good, the alarm was checked against the pulsatile rate, else it defaulted to true. Ventricular flutter/fibrillatory alarms were checked through the maximum average spectral purity index calculation over a 3 s window, and no pulsatile information was used. Finally, ventricular tachycardia alarms used a set of heuristic rules encompassing pulsatile waveform heart-rate series, as well as current versus previous values of the ECG spectral purity indexes.

6.4. Hoog Antink et al (2016)

The algorithm proposed by Hoog Antink *et al* (2016) used 16 s of data prior to the alarm event. The algorithm achieved a performance accuracy on the final test data-set of ASY: 76.7%, EBR: 74.2%, ETC: 100%, VFB: 72.8% VTA: 71.5%, yielding a total average of: 78.2% and retrospective average of 74.4%.

6.4.1. Preprocessing. The pre-processing steps for this algorithm included re-sampling of the signals to 100 Hz, band-pass filtering with a pass-band region of 1–30 Hz. The signals were also normalized to zero mean and unit variance using statistics calculated on 5 min of data prior to the alarm.

6.4.2. Beat detection. Beat detection was achieved through the Bayesian fusion of several inter-beat interval estimators that rely on self-similarity: lag adaptive short-time autocorrelation, average magnitude difference function, and maximum amplitude pairs (Brüser *et al* 2013). A quality metric based on the reliability of the fused estimates was derived from the peak height to area of the fused similarity curve.

6.4.3. Alarm classification. The classifiers chosen for the alarm validation included binary classification trees, regularized linear discriminant analysis, a support vector machine, and a random forest. The authors utilized a combination of both alarm specific and global classifiers (i.e, classifiers trained to detect a general false alarm). Their final choices were linear discriminant analysis for EBR, VFB, and VTA, a binary classifier for ETC, and a random forest model for ASY. A superset of 88 features was developed from: 24 beat-to-beat interval statistics and correlogram analysis of interval time series. From this superset, subsets were selected according to alarm types.

6.5. Kalidas and Tamil (2016)

The algorithm proposed by Kalidas and Tamil (2016) used 10 s of data prior to the alarm. The algorithm achieved a performance accuracy on the final test data-set of ASY: 80.7%, EBR: 71.7%, ETC: 99.1%, VFB: 74.1% VTA: 76.7%, yielding a total average of: 79.4% and retrospective average of 80.2%.

6.5.1. Preprocessing. Baseline wander was estimated with a low-pass filter with a 1 Hz cutoff and then subtracted from original signal. Flat line artifact was detected by testing for identical sample values in 2 s windows. ‘Zig–zag’ artifacts were detected by testing for alternating positive and negative slopes in consecutive samples over 2 s periods.

6.5.2. Beat detection. The (Pan and Tompkins 1985) algorithm was used to detected QRS complexes in the ECG signal. Pulsatile peaks were detected through first order differentiation.

6.5.3. Alarm classification. No pulsatile signal information was used for VFB and VTA arrhythmia alarms. For each alarm type, an individual support vector machine and set of heuristics was developed. The features used into these classifiers included the ECG-derived heart rate, and PPG-derived heart rate if morphology was considered valid (excluding the VFB and VTA alarms). The VFB and VTA alarms also included an additional set of features related to the power spectra of the ECG waveforms.

6.6. Krasteva et al (2016)

The algorithm proposed by Krasteva *et al* (2016) used 3–7.5 s windows prior to the alarm event, with the specific duration tuned for the each alarm type. The algorithm achieved a performance accuracy on the final test data-set of ASY: 88.0%, EBR: 93.8%, ETC: 90.7%, VFB: 72.7% VTA: 72.6%, yielding a total average of: 80.0%.

6.6.1. Preprocessing. The ECG channels were fused to form two data streams: a magnitude (second norm) and a velocity (second norm of the first order derivative). The ECG signal quality was estimated using 3 frequency bands on 4 s interval windows: high frequency was used to estimate spikes from artifacts and pacemakers, medium frequency range was used to estimate the signal level and power line interference (with intra-beat temporal statistics used to estimate power line noise level), and the low frequency band was used to estimate baseline wander. Pulsatile signals were low-pass filtered with a 1 Hz cut-off. The pulsatile signal quality was estimated with a periodicity index, and mean peak-to-peak amplitude values.

6.6.2. Beat detection.

A nonlinear filtering approach, with adaptively updated upper and lower thresholds, was used for QRS detection. The beat detector had a conventional refractory period of 150 ms.

6.6.3. Beat classification.

A beat classifier was developed for supra-ventricular and ventricular ectopic beats. A decision tree model was also used, based on features that included: information from template correlation matching, beat morphology features, and RR statistics (Krasteva *et al* 2014, 2015a).

6.6.4. Alarm classification.

The alarm classification algorithm used a set of heuristic rules based on heart rate, dominant frequency for ventricular rate, phase space area from both the ECG magnitude and velocity, and pulsatile quality metrics.

6.7. Liu et al (2016)

Liu *et al* (2016) proposed an algorithm which processed 60 s of data prior to the alarm event. The algorithm achieved a performance accuracy on the final test data-set of ASY: 88.7%, EBR: 77.7%, ETC: 89.9%, VFB: 67.7% VTA: 61.0%, yielding a total average of: 71.6% and retrospective average of 75.9%.

6.7.1. Preprocessing.

The ECG and pulsatile signals were band-pass filtered with the pass-band frequency region of 5–40 Hz for the ECGs and a pass-band frequency region of 5–35 Hz for the pulsatile waveforms.

6.7.2. Beat detection.

The authors developed an ECG R wave detection algorithm that used the average maximum amplitude from 6 non-overlapping segments. Pulsatile beats were detected via *wabp* . The final detected beats were validated based on intra- and inter-channel verification of the detected beats along with a set of rules involving the number of detected beats, R amplitude, and distance metrics between the heart rate time series.

6.7.3. Beat classification.

A set of heuristics were applied to classify beats. The features included: morphology analysis based on correlation against template, the ratio between changed beats and total beats in segment, QRS width, and maximum heart rate.

6.7.4. Alarm classification.

A set of decision rules was applied to channels that passed a data quality check (if the result of the test failed, the alarm was set to false). The features used for the second classification step included number of valid feature points, heart rate, and maximum heart rate at current analysis window.

6.8. Plešsinger et al (2016)

Plešsinger *et al* (2016) developed an algorithm that used information across multiple channels and sought to detect regions contaminated by artifacts. The algorithm achieved a performance accuracy on the final test data-set was ASY: 97.4%, EBR: 83.5%, ETC: 87.8%, VFB: 80.3% VTA: 75.0%, yielding a total average of: 81.6% and a retrospective average of 84.9%.

6.8.1. Preprocessing. The preprocessing step for this algorithm started with the detection of artifacts based on the temporal statistics of the signal under analysis. Noise and pacemaker activity were estimated based on spectral content of the 50–70 Hz frequency band. The pulsatile signals were low passed filtered with cut-off frequency at either 5 or 20 Hz. The following time windows prior to the alarm event were used to process the alarm data: ASY = 14 s, EBR = 16 s, ETC = 14 s, VFB = 13 s, VTA = 10 s.

6.8.2. Beat detection. The ECG QRS detection was based on an analysis of Fourier and Hilbert Transform derived envelopes, with a 110 ms refractory period between detection. Pulsatile based beat detection was evaluated on estimated temporal slope values.

6.8.3. Beat classification. Beat classification was performed using spectral features and descriptive residue statistics over 120 ms and 500 ms windows.

6.8.4. Alarm classification. The alarm classification stage for the ASY, VTA, and VFB alarms included using the count of invalid features obtained during the preprocessing stage described above. Additional features included statistics for the RR series obtained from multiple channels. The sum of the invalid areas had to be zero in order for the algorithm to accept the RR series as a regular rhythm for the specific channel. Finally, a set of heuristic rules was applied based on the derived RR series and the invalid region statistics.

6.9. Rodrigues and Couto (2016)

Rodrigues and Couto (2016) proposed an algorithm that uses two open-source beat detectors on the ECG waveforms as well as *wabp* on the pulsatile signals. The authors also performed beat classification based on the phase of the R wave in the ECG signals. The algorithm achieved a performance accuracy on the final test data-set of ASY: 83.6%, EBR: 71.4%, ETC: 99.1%, VFB: 88.7% VTA: 61.4%, yielding a total average of: 74.2% and a retrospective average of 74.4%.

6.9.1. Preprocessing. All signals were re-sampled to 125 Hz, and the ECG waveforms were processed for pacemaker detection and removal. Baseline noise was removed by first estimating it with a 125 sample median filter, followed by subtraction from the original signal. Flat signal regions were identified by thresholding on low variance over 2 s windows.

6.9.2. Beat detection. ECG QRS detection was performed using *gqrs* and *osea* software packages (Hamilton 2002). The beats on the pulsatile signals were detected with the *wabp* software. The authors developed their own specific beat detectors for ventricular fibrillation beats by fitting a parabola on 125 ms windows. Following the method of Li *et al* (2008), a quality index was developed based on the fraction of matched beats from *gqrs* and the *osea* software packages (Hamilton 2002) on the ECG channels.

For pulsatile signals , the quality was estimated using the morphology of consecutive beats estimated from correlation and dynamic time warp analysis, per (Li and Clifford 2012). The detected beats were fused based on quality indexes and a tolerance window of 150 ms. Pulsatile beats were compensated with a delay estimated from initial detections.

6.9.3. Beat classification. Beat classification was based on a set of heuristics modified from the *osea* software package (Hamilton 2002). These set of rules included statistics derived from inter-beat interval and QRS duration. Rodrigues and Couto (2016) also developed a four-category feature, termed 'polarity' that characterized the different types of phases of the R wave into: positive, negative, positive-negative, negative-positive (the last two representing biphasic R waves).

6.9.4. Alarm classification. Alarm classification was calculated from a set of decision rules based on signal quality, but with priority weight given to ECG signals.

6.10. Sadr et al (2016)

Sadr *et al* (2016) proposed an algorithm that uses features and processing specific arrhythmias being tested. The algorithm achieved a performance accuracy on the final test data-set was ASY: 82.4%, EBR: 71.13%, ETC: 99.1%, VFB: 65.5% VTA: 68.0%, yielding a total average of: 69.9% and a total average on the retrospective event of 74.0%.

6.10.1. Preprocessing. Baseline removal was performed by first estimating the baseline component through median filtering and then subtracting this baseline component from the original signal.

6.10.2. Beat detection. A Hilbert Transform based QRS detector based used for estimating the ECG beats (Benitez *et al* 2001). The *wabp* algorithm was used to detect the peaks on the ABP and PPG waveforms, and a quantile algorithm was also used to locate peaks on the PPG waveform.

6.10.3. Alarm classification. The alarm verification was performed on a 16 s window of data prior to the alarm. For all of the alarms with the exception of VTA, the alarm data streams had to pass four signal quality checks in order to be deemed a true alarm, otherwise they were tagged as being false. Pulsatile signal information was not used for the ETC and VTA alarms. The classification also consisted of decision trees based on several extracted features customized to each alarm type, including: threshold crossing intervals, auto-correlation function values, complexity measures, and QRS template parameters.

6.11. Zong et al (2016)

Zong *et al* (2016) is unique in that it proposed an algorithm based on pulsatile waveform features. The algorithm was developed and tested using the MIMIC II database (Saeed *et al* 2011) rather than the Challenge data, and was not open sourced.

6.11.1. Preprocessing. Pulsatile signals were low-pass filtered with cutoff set to 16 Hz, and a signal quality estimate was obtained using the technique described in Zong *et al* (2004).

6.11.2. Beat detection. Beat detection was performed with the pulsatile signals using *wabp* and with a forced detection after a period of 2 s from the last detected pulse.

6.11.3. Beat classification. The pulsatile beats were classified based on the abnormality index from Sun *et al* (2006)

6.11.4. Alarm classification. The alarm classification was achieved using features from pulsatile signals that included: pulse-to-pulse interval, amplitude, maximum slope, signal quality, and rhythm. The classifier was developed based on set of heuristic rules specific to each alarm type.

6.12. Daluwatte et al (2016)

The focus of this article was on developing a better understanding between signal quality and false alarms. The authors developed arrhythmia specific quality indexes, and investigated if existing quality indexes can distinguish between true alarms versus noise. Two humans annotated each ECG signal 10s prior to the alarm as either of high or low quality. Disagreements were not included in the analysis. The authors used 18 signal quality indexes from existing literature, and selected the top three algorithms from ROC analysis on the manually annotated data-set. The ECG beats were detected using the U3 transform (Paoletti and Marchesi 2006).

6.13. Tsimenidis and Murray (2016)

The article by Tsimenidis and Murray (2016) investigated the relationship between ECG quality and false alarms. The authors investigated the signal quality of ECG leads 8 s prior to the alarm event. They broke the analysis down into three frequency bands: low frequency from 0.1–1 Hz, mid frequency from 10–20 Hz, and high frequency from 20–40 Hz. The ECG's major power spectrum component was expected to be located mostly from 1–10 Hz. The authors report that the power on all the three frequencies was significantly greater for a false alarm versus a true alarm.

7. Conclusions

In summary, the PhysioNet/Computing in Cardiology Challenge 2015 provided several key additions to the field of false alarm suppression in critical care. First we note that for the top performing entrants, it was the VTA alarm that proved the hardest to classify accurately. This is partly because, at low heart rates, the signal becomes 'more normal looking' in the other signals. This was previously noted in Aboukhalil *et al* (2008). To-date, there has been little to address this issue, although the current Challenge has made a move towards this. Second, we note that retrospective scores were generally higher than 'real-time' scores, with the highest performing retrospective approach only suppressing 1% of the true alarms, while 80% of the false alarms were suppressed. Although debatable, this may be acceptable as a clinical algorithm if a 30 s window were acceptable. We suggest that this may spur a re-consideration of the Association for the Advancement of Medical Instrumentation (AAMI) guidelines for maximum alarm latency. Third, we note that voting algorithms together can produce superior results to even the best algorithm. Such an approach can also lead to a more robust implementation, although it may be significantly more computationally intensive. It is also important to note that too many naive voters (more than 13 in the case of the 2015 Challenge) can reduce the accuracy of the label or answer. In Zhu *et al* (2014) and Zhu *et al* (2015) a voting system for algorithm (and human) annotations of physiological data was described, which incorporates both the physiology and the individual annotator's accuracy as a function of objective features (such as signal quality) to produce a weighted voting scheme to guarantee that all voters add extra information. We suggest that such approaches will become ever more important as computational power becomes increasingly less expensive. We also note that this means that all competitors in the challenge added something to the final answer!

Finally we note some limitations of the competition. A larger database is needed with more patients, longer recordings, more leads and abnormalities (such as arrhythmias). We intend to work with industry and researchers alike to enhance the Challenge database in all these areas and would be grateful for continued contributions of data and source code, which we will post together with all the open source algorithms and annotated data from the 2015

PhysioNet/Computing in Cardiology Challenge. The latter can be found on PhysioNet's website at http://physionet.org/challenge/2015.

Acknowledgments

This work was funded in part by the National Institutes of Health, grant R01-GM104987 and by the National Institute of General Medical Sciences, under NIH cooperative agreement U01-EB-008577 and NIH grant R01-EB-001659. We also would like to thank Arvind Ananthan, Jaka Cikač, Barbara Drew, Quan Ding, Scott Eaton, Xiao Hu, Franc Jager, Cadathur 'Raj' Rajagopalan and Anže Rezelj, as well as GE Healthcare, MathWorks, Mindray North America and Philips Healthcare for their valuable assistance. Finally, we thank all the competitors themselves, without whom there would be no competition.

References

Aboukhalil A, Nielsen L, Saeed M, Mark R G and Clifford G D 2008 Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform *J. Biomed. Inf.* **41** 442–51

Allen J and Murray A 1996 Assessing ECG signal quality on a coronary care unit *Phys. Meas.* **17** 249

Ansari S, Belle A, Ghanbari H, Salamango M and Najarian K 2016 Suppression of false arrhythmia alarms in the ICU: a machine learning approach *Phys. Meas.* **37** 1186–203

Ansari S, Belle A and Najarian K 2015 Multi-modal integrated approach towards reducing false arrhythmia alarms during continuous patient monitoring: the PhysiOnet challenge 2015 *IEEE Computing in Cardiology Conf.* pp 1181–4

Arberet S, Lemay M, Renevey P, Sola J, Grossenbacher O, Andries D, Sartori C and Bertschi M 2013 Photoplethysmography-based ambulatory heartbeat monitoring embedded into a dedicated bracelet *IEEE Computing in Cardiology Conf.* pp 935–8

Behar J, Johnson A, Clifford G D and Oster J 2014a A comparison of single channel fetal ECG extraction methods *Ann. Biomed. Eng.* **42** 1340–53

Behar J, Oster J and Clifford G D 2014b Combining and benchmarking methods of foetal ECG extraction without maternal or scalp electrode data *Phys. Meas.* **35** 1569

Behar J, Oster J, Li Q and Clifford G D 2013 ECG signal quality during arrhythmia and its application to false alarm reduction *IEEE Trans. Biomed. Eng.* **60** 1660–6

Benitez D, Gaydecki P, Zaidi A and Fitzpatrick A 2001 The use of the hilbert transform in ECG signal analysis *Comput. Biol. Med.* **31** 399–406

Berg S 2001 Impact of reduced reverberation time on sound-induced arousals during sleep *Sleep* **24** 289–92

Brüser C, Winter S and Leonhardt S 2013 Robust inter-beat interval estimation in cardiac vibration signals *Physiol. Meas.* **34** 123

Chambrin M-C 2001 Alarms in the intensive care unit: how can the number of false alarms be reduced? *Crit. Care* **5** 184–8

Chambrin M-C, Ravaux P, Calvelo-Aros D, Jaborska A, Chopin C and Boniface B 1999 Multicentric study of monitoring alarms in the adult intensive care unit (ICU): a descriptive analysis *Intensive Care Med.* **25** 1360–6

Clifford G D 2002 Signal processing methods for heart rate variability *PhD Thesis* University of Oxford

Clifford G-D, Behar J, Li Q and Rezek I 2012 Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms *Physiol. Meas.* **33** 1419–33

Clifford G D, Silva I, Moody B, Li Q, Kella D, Shahin A, Kooistra T, Perry D and Mark R G 2015 The PhysioNet/Computing in Cardiology Challenge 2015: reducing false arrhythmia alarms in the ICU *IEEE Computing in Cardiology Conf.* pp 273–6

Couto P, Ramalho R and Rodrigues R 2015 Suppression of false arrhythmia alarms using ECG and pulsatile waveforms *IEEE Computing in Cardiology Conf.* pp 749–52

Cropp A-J, Woods L-A, Raney D and Bredle D-L 1994 Name that tone. The proliferation of alarms in the intensive care unit *Chest* **105** 1217–20

Daluwatte C, Johannesen L, Galeotti L, Vicente J, Strauss D G and Scully C G 2016 Assessing ECG signal quality indices to discriminate ECGs with artefacts from pathologically different arrhythmic ECGs *Phys. Meas.* **37** 1370–82

Donchin Y and Seagull F-J 2002 The hostile environment of the intensive care unit *Curr. Opin. Crit. Care* **8** 316–20

Eerikäinen L M, Vanschoren J, Rooijakkers M J, Vullings R and Aarts R M 2015 Decreasing the false alarm rate of arrhythmias in intensive care using a machine learning approach *IEEE Computing in Cardiology Conf.* pp 293–6

Eerikäinen L M, Vanschoren J, Rooijakkers M J, Vullings R and Aarts R M 2016 Reduction of false arrhythmia alarms using signal selection and machine learning *Phys. Meas.* **37** 1204–16

Engelse W-A-H and Zeelenberg C 1979 A single scan algorithm for QRS-detection and feature extraction *Comput. Cardiol.* **6** 37–42

Fallet S, Yazdani S and Vesin J-M 2015 A multimodal approach to reduce false arrhythmia alarms in the intensive care unit *IEEE Computing in Cardiology Conf.* pp 277–80

Fallet S, Yazdani S and Vesin J-M 2016 False arrhythmia alarms reduction in the intensive care unit: a multimodal approach *Phys. Meas.* **37** 1217–32

Goldberger A L, Amaral L A, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C-K and Stanley H E 2000 Physiobank, physiotoolkit, and PhysioNet components of a new research resource for complex physiologic signals *Circulation* **101** e215–20

Goncharova I I and Barlow J S 1990 Changes in eeg mean frequency and spectral purity during spontaneous alpha blocking *Electroencephalogr. Clin. Neurophysiol.* **76** 197–204

Hagerman I, Rasmanis G, Blomkvist V, Ulrich R, Eriksen C-A and Theorell T 2005 Influence of intensive coronary care acoustics on the quality of care and physiological state of patients *Int. J. Cardiol.* **98** 267–70

Hamilton P 2002 Open source ECG analysis *IEEE Computers in Cardiology* pp 101–4

Hamilton P-S and Tompkins W-J 1986 A real-time QRS detection algorithm *IEEE Trans. Biomed. Eng.* **33** 1157–65

Hoog Antink C B, Leonhardt S and Walter M 2016 Reducing false alarms in the ICU by quantifying self-similarity of multimodal biosignals *Phys. Meas.* **37** 1233–52

Hoog Antink C and Leonhardt S 2015 Reducing false arrhythmia alarms using robust interval estimation and machine learning *IEEE Computing in Cardiology Conf.* pp 285–8

Hug C W, Clifford G D and Reisner A T 2011 Clinician blood pressure documentation of stable intensive care patients: an intelligent archiving agent has a higher association with future hypotension *Crit. Care Med.* **39** 1006

Imhoff M and Kuhls S 2006 Alarm algorithms in critical care monitoring *Anesth. Analg.* **102** 1525–37

Johnson A-N 2001 Neonatal response to control of noise inside the incubator *Pediatric Nursing* **27** 600–5

Kalidas V and Tamil L 2015 Enhancing accuracy of arrhythmia classification by combining logical and machine learning techniques *IEEE Computing in Cardiology Conf.* pp 733–6

Kalidas V and Tamil L 2016 Cardiac arrhythmia classification using multi-modal signal analysis *Phys. Meas.* **37** 1253–72

Krasteva V, Jekova I, Leber R, Schmid R and Abächerli R 2015a Superiority of classification tree versus cluster, fuzzy and discriminant models in a heartbeat classification system *PloS One* **10** e0140123

Krasteva V, Jekova I, Leber R, Schmid R and Abächerli R 2015b Validation of arrhythmia detection library on bedside monitor data for triggering alarms in intensive care *IEEE Computing in Cardiology Conf.* pp 737–40

Krasteva V, Jekova I, Leber R, Schmid R and Abächerli R 2016 Real-time arrhythmia detection with supplementary ECG quality and pulse wave monitoring for reduction of false alarms in ICUs *Phys. Meas.* **37** 1273–97

Krasteva V, Leber R, Jekova I, Schmid R and Abacherli R 2014 Classification of supraventricular and ventricular beats by qrs template matching and decision tree *IEEE Computing in Cardiology Conf.* pp 349–52

Lawless S T 1994 Crying wolf: false alarms in a pediatric intensive care unit *Crit. Care Med.* **22** 981–5

Li Q and Clifford G D 2012 Dynamic time warping and machine learning for signal quality assessment of pulsatile signals *Phys. Meas.* **33** 1491–501

Li Q, Mark R G and Clifford G D 2008 Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter *Phys. Meas.* **29** 15–32

Li Q, Rajagopalan C and Clifford G 2014a Ventricular fibrillation and tachycardia classification using a machine learning approach *IEEE Trans. Biomed. Eng.* **61** 1607–13

Li Q, Rajagopalan C and Clifford G-D 2014b A machine learning approach to multi-level ECG signal quality classification *Comput. Methods Programs Biomed.* **117** 435–47

Liu C, Zhao L and Tang H 2015 Reduction of false alarms in intensive care unit using multi-feature fusion method *IEEE Computing in Cardiology Conf.* pp 741–4

Liu C, Zhao L, Tang H, Li Q, Wei S and Li J 2016 Life-threatening false alarms rejection in ICU: using rule-based and multi-channel information fusion method *Phys. Meas.* **37** 1298–312

Meyer T-J, Eveloff S-E, Bauer M-S, Schwartz W-A, Hill N-S and Millman R-P 1994 Adverse environmental conditions in the respiratory and medical ICU settings *Chest* **105** 1211–6

Morrison W-E, Haas E-C, Shaffner D-H, Garrett E-S and Fackler J-C 2003 Noise, stress, and annoyance in a pediatric intensive care unit *Crit. Care Med.* **31** 113–9

Mullins P M, Goyal M and Pines J M 2013 National growth in intensive care unit admissions from emergency departments in the United States from 2002 to 2009 *Acad. Emergency Med.* **20** 479–86

Novaes M-A, Aronovich A, Ferraz M-B and Knobel E 1997 Stressors in ICU: patients' evaluation *Intensive Care Med.* **23** 1282–5

Nygårds M-E and Sörnmo L 1983 Delineation of the QRS complex using the envelope of the ECG *Med. Biol. Eng. Comput.* **21** 538–47

Oster J, Behar J, Colloca R, Li Q, Li Q and Clifford G D 2013 Open source java-based ECG analysis software and android app for atrial fibrillation screening *Comput. Cardiol* pp 731–4

Pan J and Tompkins W J 1985 A real-time qrs detection algorithm *IEEE Trans. Biomed. Eng.* **32** 230–6

Paoletti M and Marchesi C 2006 Discovering dangerous patterns in long-term ambulatory ECG recordings using a fast QRS detection algorithm and explorative data analysis *Comput. Methods Programs Biomed.* **82** 20–30

Parthasarathy S and Tobin M-J 2004 Sleep in the intensive care unit *Intensive Care Med.* **30** 197–206

Plešsinger F, Klimes P, Halamek J and Jurak P 2015 False alarms in intensive care unit monitors: detection of life-threatening arrhythmias using elementary algebra, descriptive statistics and fuzzy logic *IEEE Computing in Cardiology Conf.* pp 281–4

Plešsinger F, Klimes P, Halamek J and Jurak P 2016 Taming of the monitors: reducing false alarms in intensive care units *Phys. Meas.* **37** 1313–25

Rodrigues R and Couto P 2016 Detection of false arrhythmia alarms with emphasis on ventricular tachycardia *Phys. Meas.* **37** 1326–39

Rooijakkers M J, Rabotti C, Oei S G and Mischi M 2012 Low-complexity r-peak detection for ambulatory fetal monitoring *Physiol. Meas.* **33** 1135

Sadr N, Huvanandana J, Nguyen D T, Kalra C, McEwan A and de Chazal P 2016 Reducing false arrhythmia alarms in the ICU using multimodal signals and robust QRS detection *Phys. Meas.* **37** 1340–54

Saeed M, Villarroel M, Reisner A T, Clifford G D, Lehman L-W, Moody G B, Heldt T, Kyaw T H, Moody B and Mark R G 2011 Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database *Crit. Care Med.* **39** 952

Silva I and Moody G-B 2014 An open-source toolbox for analysing and processing PhysioNet databases in MATLAB and Octave *J. Open Res. Softw.* **2** e27

Slevin M, Farrington N, Duffy G, Daly L and Murphy J-F 2000 Altering the NICU and measuring infants' responses *Acta Paediatrica* **89** 577–81

Sörnmo L and Laguna P 2005 *Bioelectrical Signal Processing in Cardiac and Neurological Applications* (New York: Academic)

Sun J, Reisner A and Mark R 2006 A signal abnormality index for arterial blood pressure waveforms *Comput. Cardiol.* pp 13–6

Topf M and Thompson S 2001 Interactive relationships between hospital patients' noise induced stress and other stress with sleep *Heart Lung* **30** 237–43

Tsien C L and Fackler J C 1997 Poor prognosis for existing monitors in the intensive care unit *Crit. Care Med.* **25** 614–9

Tsimenidis C and Murray A 2016 False alarms during patient monitoring in clinical intensive care units are highly related to poor quality of the monitored electrocardiogram signals *Phys. Meas.* **37** 1383–91

Yazdani S and Vesin J-M 2014 Adaptive mathematical morphology for qrs fiducial points detection in the ECG *IEEE Computing in Cardiology Conf.* pp 725–8

Zhu T, Dunkley N, Behar J, Clifton D A and Clifford G D 2015 Fusing continuous-valued medical labels using a Bayesian model *Ann. Biomed. Eng.* **43** 2892–902

Zhu T, Johnson A E W, Behar J and Clifford G D 2014 Crowd-sourced annotation of ECG signals using contextual information *Ann. Biomed. Eng.* **42** 871–84

Zong W, Heldt T, Moody G and Mark R 2003a An open-source algorithm to detect onset of arterial blood pressure pulses *Comput. Cardiol.* pp 259–62

Zong W, Moody G-B and Jiang D 2003b A robust open-source algorithm to detect onset and duration of QRS complexes *Comput. Cardiol.* **30** 737–40

Zong W, Moody G-B and Mark R-G 2004 Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure *Med. Biol. Eng. Comput.* **42** 698–706

Zong W, Nielsen L, Gross B, Brea J and Frassica J 2016 A practical algorithm to reduce false critical ECG alarms using arterial blood pressure and/or photoplethysmogram waveforms *Phys. Meas.* **37** 1355–69

Gari D Clifford[1,2]**, Ikaro Silva**[3]**, Benjamin Moody**[3]**, Qiao Li**[1]**, Danesh Kella**[1]**, Abdullah Chahin**[4]**, Tristan Kooistra**[4]**, Diane Perry**[4] **and Roger G Mark**[3]

[1] Department of Biomedical Informatics, Emory University, Atlanta GA, USA
[2] Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta GA, USA
[3] Institute for Medical Engineering & Science, Massachusetts Institute of Technology, USA
[4] Beth Israel Medical Center, Harvard University, Boston MA, USA

E-mail: gari@gatech.edu