Reduction of false arrhythmia alarms using signal selection and machine learning

2016 Physiol. Meas. 37 1204

(http://iopscience.iop.org/0967-3334/37/8/1204)

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

Physiol. Meas. **37** (2016) 1204–1216 doi:10.1088/0967-3334/37/8/1204

Reduction of false arrhythmia alarms using signal selection and machine learning

Linda M Eerikäinen[1], Joaquin Vanschoren[2], Michael J Rooijakkers[1], Rik Vullings[1] and Ronald M Aarts[1,3]

[1] Department of Electrical Engineering, Eindhoven University of Technology, 5612 AZ, Eindhoven, The Netherlands
[2] Department of Mathematics and Computer Science, Eindhoven University of Technology, 5612 AZ, Eindhoven, The Netherlands
[3] Department of Personal Health, Philips Research, Eindhoven, The Netherlands

E-mail: L.M.Eerikainen@tue.nl

Received 29 February 2016, revised 20 April 2016
Accepted for publication 5 May 2016
Published 25 July 2016

Abstract

In this paper, we propose an algorithm that classifies whether a generated cardiac arrhythmia alarm is true or false. The large number of false alarms in intensive care is a severe issue. The noise peaks caused by alarms can be high and in a noisy environment nurses can experience stress and fatigue. In addition, patient safety is compromised because reaction time of the caregivers to true alarms is reduced.

The data for the algorithm development consisted of records of electrocardiogram (ECG), arterial blood pressure, and photoplethysmogram signals in which an alarm for either asystole, extreme bradycardia, extreme tachycardia, ventricular fibrillation or flutter, or ventricular tachycardia occurs. First, heart beats are extracted from every signal. Next, the algorithm selects the most reliable signal pair from the available signals by comparing how well the detected beats match between different signals based on F_1-score and selecting the best match. From the selected signal pair, arrhythmia specific features, such as heart rate features and signal purity index are computed for the alarm classification. The classification is performed with five separate Random Forest models. In addition, information on the local noise level of the selected ECG lead is added to the classification. The algorithm was trained and evaluated with the PhysioNet/Computing in Cardiology Challenge 2015 data set. In the test set the overall true positive rates were 93 and 95% and true negative rates 80 and 83%, respectively for events with no information and events with information after the alarm. The overall challenge scores were 77.39 and 81.58.

Keywords: alarms, cardiac arrhythmia, intensive care, Random Forest

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

1. Introduction

The number of false alarms in intensive care units (ICUs) have been reported to be 40–86% (Lawless 1994, Tsien and Fackler 1997, Siebig *et al* 2010). The noise level in intensive care is high and alarms may reach a noise peak that exceeds 80 dB (Balogh *et al* 1993). Noise can cause stress and fatigue to the nursing staff (Konkani and Oakley 2012). In addition, patient safety is compromised because the excessive amount of alarms affects the reaction time of caregivers to respond (Graham and Cvach 2010).

The problem of false arrhythmia alarm reduction has been approached with various strategies. These approaches can be roughly divided into filtering methods, signal quality assessment, multi-parametric analysis, machine learning approaches, or combinations of the previous.

The filtering methods aim to suppress the variation and outliers in the signal that may cause the false alarms. Proposed methods include median filtering (Mäkivirta *et al* 1991), statistical signal filtering (Borowski *et al* 2011), and model-based filtering (Sayadi and Shamsollahi 2011).

A signal quality assessment of electrocardiogram (ECG) by combining several signal quality indices (SQIs) was proposed by Behar *et al* (2013) to suppress false cardiac arrhythmia alarms. Different SQIs were combined with machine learning methods. Clifford *et al* (2006) evaluated first the quality of arterial blood pressure (ABP) to either suppress an alarm directly or to use additional information from the ABP signal for the decision if the alarm should be suppressed. Aboukhalil *et al* (2008) evaluated both timing and signal abnormality information from ABP to suppress false ECG arrhythmia alarms. Instead of ABP, Deshmane (2009) used signal quality and onset information of photoplethysmograms (PPG).

The assessment of signal quality also has an important role in multi-parameter approaches. Li *et al* (2008) estimate heart rate (HR) by fusing ECG, ABP, and PPG, and using SQIs and a Kalman filter. HR features are then used for deciding whether an alarm should be suppressed. Optionally, they use SQIs and several other features from the signals with machine learning to suppress the false alarms.

The PhysioNet/Computing in Cardiology Challenge 2015 (the Challenge) (Clifford *et al* 2015) provided an open data set with ECG, ABP, PPG, and respiratory data from intensive care inviting competitors to develop algorithms to reduce false arrhythmia alarms for five life-threatening arrhythmia types: asystole (ASY), extreme bradycardia (EBR), extreme tachycardia (ETC), ventricular tachycardia (VTA), and ventricular flutter or fibrillation (VFB). The data consists of 750 records for the training set and 500 records for the unrevealed test set. Both in training and test set half of the records are 5 min long and the other half contains an additional 30 s after the alarm. In every record the alarm occurs at 5 min from the beginning of the record. Every record contains 3–4 signals of which two are always ECG leads. The additional signals are either or both ABP and PPG signals, and in some cases a respiratory signal.

Many of the well-performing algorithms the Challenge had a signal quality or noise level assessment implemented in them (Couto *et al* 2015, Daluwatte *et al* 2015, Krasteva *et al* 2015, Zong 2015). Plesinger *et al* (2015) used testing of regular heart activity in multiple signals, and if no regular activity was detected, specific arrhythmia test was performed. Fallet *et al* (2015) had an approach based on robust HR estimation and signal purity index (SPI).

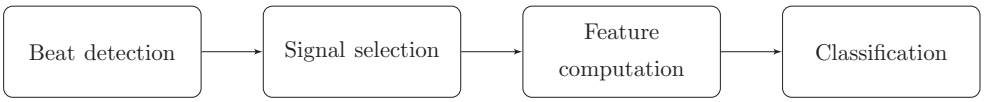

Figure 1. Flowchart of the algorithm.

In addition, machine learning approaches were presented by Hoog Antink and Leonhardt (2015) and Kalidas and Tamil (2015).

The algorithm presented in this paper compares the detected heart beats from ECG, ABP, and PPG signals, and selects which combination of two signals is the most reliable. The estimation of the reliability of this signal pair is based on the F_1-score (van Rijsbergen 1979) of the beats that are detected simultaneously in the two signals. The F_1-score is a measure combining sensitivity and precision and giving both an equal weight. After selecting the signal pair, arrhythmia specific features are computed from both signals. The features consist of HR features and SPI features. In addition, the algorithm uses the F_1-score as a quality feature for the classification. The classification between true and false alarms is performed with five separate Random Forest classifiers, one for each arrhythmia type. The performance is compared between two sets of features: having the F_1-score as the only quality feature and when adding local noise level around R-peaks as an additional feature.

2. Methods

The overall flowchart of our alarm reduction algorithm is presented in figure 1. In this section we will first describe the beat detection from ECG and pulsatile signals (section 2.1) followed by the signal selection (section 2.2). Next, the feature computation from the selected signal pair is described in section 2.3. Finally, classification and performance evaluation are presented in sections 2.4 and 2.5, respectively.

2.1. Beat detection

The first step of our algorithm is to detect heart beats from ECG and pulsatile signals. Before the beat detection, the signals were downsampled from 250 Hz to 125 Hz.

2.1.1. ECG beat detection. The beat detection from ECG is performed with a low-complexity R-peak detector (Rooijakkers *et al* 2012). First, in a preprocessing stage, the ECG is convoluted with a Mexican hat wavelet. The absolute value of the convoluted signal produces an output S from which the R-peaks are detected. The absolute value enables the use of a single threshold and suits for our case where the lead of the ECG is unknown.

The detection of R-peaks is executed in an iterative fashion in four stages: segment selection, threshold determination, peak detection, and signal-to-noise ratio (SNR) estimation. The segment for R-peak detection is selected in such a fashion that only one QRS complex is expected in the segment. The limits for the segment are based on the position of the previously detected R-peak and on the assumption that the heart rate is in the range of 32–210 beats per minute (bpm).

The threshold for every R-peak detection is determined by the previous threshold, T_{prev}, and a new threshold estimate, $\widehat{T}$, following a first order autoregressive process

$$T = \alpha \cdot \widehat{T} + (1 - \alpha) \cdot T_{\text{prev}}, \tag{1}$$

where α is a coefficient describing the dynamic behaviour of the threshold. The value of α was set to the default value 1/3. The new threshold estimate, $\widehat{T}$, is the product of the maximum amplitude of the preprocessed signal, $S_{\max}$, and the local noise level estimate, N_l, in the signal segment.

The N_l is indicative of the local SNR and is scaled to the range of [0, 1]. For the scaled noise level estimate, first an estimate of SNR is needed.

SNR is classically determined by

$$\mathrm{SNR} = 10 \cdot \log_{10}\left(\frac{P_\mathrm{s}}{P_\mathrm{n}}\right), \tag{2}$$

where P_s is the signal power and P_n is the noise power. For the purposes of R-peak detection algorithm, P_s is defined as the power around the detected R-peak $\hat{p}$, $S[\hat{p}]^2$, and P_n as the maximal power $N_{\max}^2$ in the segments between consecutive R-peaks. $N_{\max}$ is the maximum of S in the segments before and after the detected peak. P_s and P_n can now be replaced in equation (2) and the power of 2 can be taken from the logarithm as a multiple in front of the logarithm. In order to reduce the computational complexity of the function, the decadic logarithm is replaced by a binary logarithm and a low-complexity estimate of SNR is rewritten as

$$\widehat{\mathrm{SNR}} = \log_2(S[\hat{p}]) - \log_2(N_{\max}). \tag{3}$$

Except for a scaling factor, the low-complexity estimate of the function corresponds to its higher complexity equivalent, and $\widehat{\mathrm{SNR}}$ corresponds to signal quality. The $\widehat{\mathrm{SNR}}$ is further scaled as the local noise level estimate

$$N_l = \frac{6 - \widehat{\mathrm{SNR}}}{8}, \tag{4}$$

where a low value of N_l indicates the minimal noise level and $N_l \geqslant 6/8$ indicates an $\widehat{\mathrm{SNR}}$ below 0 dB.

After determining the threshold, the first preprocessed sample crossing the threshold is the peak position candidate. The peak position candidate is updated if in the vicinity a sample with a higher amplitude is found. If no candidate is found in the segment, two other iterations are performed with an extended segment and lowered threshold. If after three iterations no R-peak is found, the segment moves forward with 1 s.

2.1.2. Pulsatile signals beat detection. The pulse detection from the ABP and PPG was performed with an open-source ABP pulse onset detection algorithm, *wabp* (Zong *et al* 2003). The algorithm is available from PhysioNet (Clifford *et al* 2015).

The key concept of the algorithm is to transform the ABP waveform into a slope sum function (SSF) signal. The purpose of the SSF is to enhance the upslope of the waveform and suppress the remainder of the waveform. Before the SSF transformation, the signal is low-pass filtered to suppress high frequency noise that might affect the onset detection. The windowed and weighted SSF, z, at time i is defined as

$$z_i = \sum_{k=i-w}^{i} \Delta u_k, \quad \Delta u_k = \begin{cases} y_k - y_{k-1} & : \quad y_k - y_{k-1} > 0 \\ 0 & : \quad y_k - y_{k-1} \leqslant 0 \end{cases} \tag{5}$$

where w is the length of the window and y_k the low-pass filtered ABP signal (Zong *et al* 2003).

The final step of the algorithm is to establish a decision rule for the detection of each SSF pulse onset. This is done in two steps. First, adaptive thresholding is applied to detect the

SSF pulses that have appropriate amplitude. Second, a search is employed locally around the detection point to confirm the detection and to identify the likely onset of the pulse.

This algorithm is developed for ABP signals, but was used in this paper for PPG pulse onset detection as well. The scale of the PPG signal was adjusted to correspond to the scale of an ABP signal before the pulse detection.

2.2. Signal selection

False alarms are usually caused by an artifact or a disturbance in the signal that may resemble the physiological event the alarm is intended for, e.g. a flat signal can be misinterpreted as an asystole event because no heart beats are detected. Important in the false alarm reduction is to evaluate what is the quality of the signals or how accurate the features obtained from the signals are. If we can be certain that a feature, e.g. heart rate, is accurately measured, we can rely on the decision based on this feature.

Accurate signal quality measures during arrhythmia are difficult to develop and may not always work as desired. Behar *et al* (2013) combined several ECG SQIs for classifying ECG signal quality during arrhythmia. Their conclusion was that SQIs should be developed separately for every arrhythmia and sufficient data would be needed to develop classifiers for a quality classification.

In our previous work (Eerikäinen *et al* 2015), the signals selected for the feature computation were always one ECG lead and one pulsatile signal. However, there are cases where both ECG signals or the available pulsatile signal or signals may be corrupted. Figure 2 shows an example where both ECG leads are of bad quality, but both pulsatile signals have good quality instead.

In the current approach, the aim is to select for the feature computation the signal pair that has the most accurate beat detection on average. If a beat is detected in several signals, the beat is likely to be a real beat. Since we do not know beforehand in which of the signals the beats are detected most accurately, we cannot determine which of the signals and the detected beats in that signal are the reference. Therefore, measures such as sensitivity, also called true positive rate (TPR), and precision are not adequate for our purpose. They would produce two values per a signal pair because both signals would be needed to be considered as a reference. Sensitivity and precision are defined as

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

and

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

where TP are true positives, FP false positives, and FN false negatives.

We select to use a method based on F_1-score to select the signal pair that has the most beats detected in both signals simultaneously. The F_1-score is a harmonized mean between sensitivity and precision

$$F_1 = \frac{2 \cdot \text{sensitivity} \cdot \text{precision}}{\text{sensitivity} + \text{precision}}. \tag{8}$$

Inserting the equations (6) and (7) into (8) the F_1-score can be computed as

$$F_1 = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}. \tag{9}$$

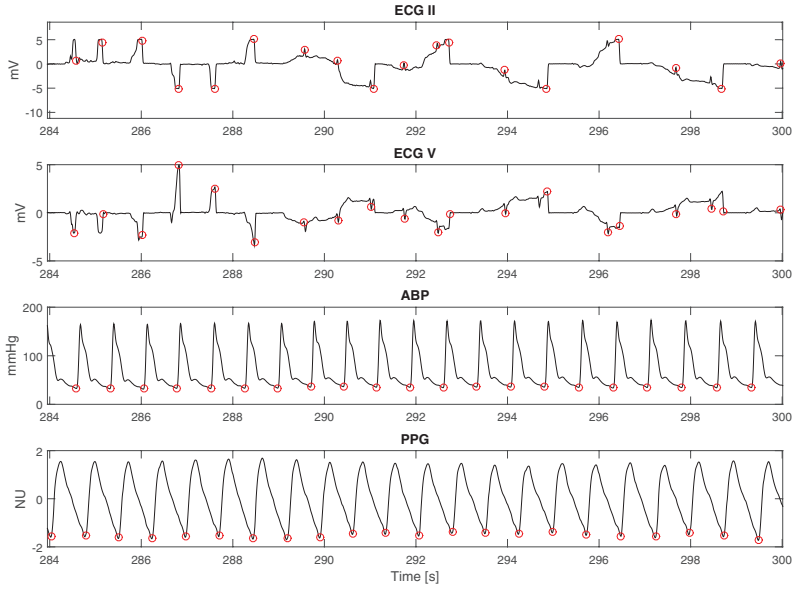

Figure 2. An example of bad quality ECG signals when pulsatile signals are good. The red circles indicate detected beats.

In the beat comparison, TPs are the beats detected both in the reference signal and in the signal that is compared. FPs are the beats detected in the compared signal, but not in the reference signal, and FNs the beats detected in the reference without detection in the compared signal. In a signal pair, when the roles of the signals are changed having now the compared signal as a reference, the number of FPs becomes the number of FNs and vice versa. Therefore, F_1-score gives one score for a signal pair that measures the agreement between the detected beats in the signals and there is no need to select either of the signals as a reference. When F_1-score is 1, all the beats in both signals match with each other, and when F_1-score is 0, none of the beats match. In figure 2 the F_1-score between ABP and PPG is 1. Previously, F_1-score has been used as a signal quality measure for ECG by comparing beats detected with different beat detectors from the same signal (Pimentel *et al* 2015). In our case, the compared beats are from different signal sources, i.e. ECG, ABP, and PPG.

Between the signals, two beats are considered as a match when they occur within 100 ms from each other. ECG, ABP, and PPG are measured from different locations of the body and are based on different measurement principles. The information about a heart beat is not visible simultaneously. The beat will be visible first in the ECG which is measured from the chest and the hemodynamic response caused by the beat is visible later in ABP and then in PPG. The delay between different signal types is compensated when finding matching beats between ECG and ABP, ECG and PPG, and ABP and PPG. The delay is computed as the mean delay from a period of 10 consecutive beats when the standard deviation of the 10 consecutive delays is less than 5% of the mean delay. If such a period is not present in the signal then the delay is not compensated.

F_1-score is computed for every signal pair, excluding respiratory signals, in a window before the alarm. The same window will be later used for the selected pair for feature computation which will be explained in section 2.3. Since the length of the arrhythmia event before triggering an alarm varies depending on the type of arrhythmia, the window length was optimized

for every arrhythmia type separately. The optimized window lengths for different arrhythmia types vary from 14 to 16 s.

Based on the F_l-score, the most appropriate signal pair was selected for the feature computation. The signal combination of an ECG lead and a pulsatile signal or both pulsatile signals with the greatest F_l-score was selected for ASY, if the F_l-score was greater than zero, and for EBR and ETC if the F_l-score was greater than 0.5. Otherwise, the signal pair with the maximum F_l-score was selected. For VFB and VTA both ECG leads were always selected, and F_l-score was computed for the ECG pair.

2.3. Feature computation

During arrhythmia the heart rate is not in the normal range or there are irregularities in the heart beats. For every arrhythmia, one or more features were designed characterizing the arrhythmia based on the definition given in the Challenge description (Clifford *et al* 2015).

EBR and ETC are arrhythmias in which HR is either lower or higher than normal. The features for these two arrhythmias were based on their definitions: the minimum HR of five consecutive beats for EBR and maximum HR of 17 consecutive beats for ETC. During ASY there are no beats for at least four seconds, which can be characterized by the maximum interval between two consecutive beats.

In VFB the heart exhibits a rapid fibrillatory, flutter, or oscillatory waveform for at least four seconds (Clifford *et al* 2015). VTA, on the other hand, is characterized by a number of consecutive beats originating from the ventricles with an HR greater than 100 bpm. Therefore, a measure based solely on heart rate or inter-beat intervals is not sufficient for identifying the two arrhythmia.

Previously, good results for VFB and VTA classification have been reported with spectral purity index (Fallet *et al* 2015). The SPI was initially presented for electroencephalogram (EEG) analysis as a dimensionless parameter between 0 and 1 reflecting the signal bandwidth (Goncharova and Barlow 1990). The parameter has the maximum value 1 for a pure sine wave and diminishes as the bandwidth of the signal increases. The SPI of a signal is defined as the ratio between the squared, running second-order moment $\bar{\omega}_2$, and the running total power $\bar{\omega}_0$ and fourth-order moment $\bar{\omega}_4$ (Sörnmo and Laguna 2005),

$$\Gamma_{\mathrm{SPI}} = \frac{\bar{\omega}_2^2(n)}{\bar{\omega}_0(n)\bar{\omega}_4(n)}. \tag{10}$$

The spectral moments were implemented in the time domain according to Sörnmo and Laguna (2005). As done in the approach of Fallet *et al* (2015), before computing SPI, ECG signals were first downsampled to 35 Hz and smoothed using a 5-sample moving average filter. The window length for estimation of the spectral moments in the time domain was selected to be 4 s for VFB, since the length of the fibrillatory waveform should be at least 4 s. For VTA, the window length was 2 s. The SPI was then averaged in a window of 1 s, and the maximum and minimum of the averaged SPI in the window before the alarm were calculated as features.

The arrhythmia specific features used in the classification are listed in the table 1. The window length for computing the features varied between the arrhythmias. For ASY and VFB the window was 14 s, for EBR 15 s, and for ETC and VTA 16 s before the alarm. Moreover, the F_l-score used in the signal selection was added to the feature sets which are given to the classifier as an input. To evaluate whether the F_l-score alone is a sufficient quality feature for the classification, the feature sets for classification were created with and without adding also the median local noise level N_l of the ECG (see equation (4)).

Table 1. Features computed from the selected signal pair.

Arrhythmia	Features
ASY	Maximum inter-beat interval
EBR	Minimum heart rate of 5 consecutive beats
ETC	Maximum heart rate of 17 consecutive beats
VFB	Maximum SPI
VTA	Maximum and minimum SPI, maximum heart rate

Note. Two feature sets were created for every arrhythmia type with the above features: one with adding F_l-score and the other adding both the F_l-score and median N_l.

2.4. Classification

Our algorithm uses five different Random Forest classifiers, each trained for one type of arrhythmia. The choice for the classifier was made after first comparing a large number of classification algorithms. The Random Forests performed overall the best and were therefore selected as the classifier for our algorithm.

A Random Forest is a collection of a large number of tree-structured classifiers in which every tree in the classifier casts a unit vote for the most popular class. The trees in the Random Forest are grown by selecting randomly the inputs or combinations of inputs at each node of the tree to determine the split (Breiman 2001). Single-tree approaches have been presented previously with good results for integrating multiple signals for artifact detection in neonatal ICU (Tsien *et al* 2000) and patient specific alarming models (Zhang and Szolovits 2008). A binary classification tree was used also in the Challenge entry of Hoog Antink and Leonhardt (2015).

The classification accuracy improves when instead of having a single tree more trees form an ensemble (Breiman 2001). Random Forests have been previously presented in an alarm classification setting as an analogy to a statistical hypothesis test for 'situation is alarm relevant' versus 'situation is not alarm relevant' (Sieben and Gather 2007). Several physiological measures were given as an input to the Random Forest and the rate of false alarms was reduced by 45–30% on average. In our approach, we use fewer and more event targeted features as inputs for the Random Forest.

In the final algorithm, the records with F_l-score zero are assigned directly as false alarms and are not classified with the Random Forest. Therefore, before training the models, feature vectors for records having F_l-score zero were removed from the training set. In total 24 records of false alarms and one true alarm were removed. In addition, the ETC record 't409l', labeled as a false alarm was removed from the training set. In the record, both pulsatile signals had clear recognizable beats that matched completely with each other and the HR was 157–158 bpm for at least 17 consecutive beats, therefore this alarm was assumed to be true.

The classifiers were tested both with 100 and 500 trees. Asystole was the only arrhythmia type for which the performance in the training set improved when increasing the number of trees to 500, and therefore the asystole classifier was selected to consist of 500 trees. The remaining four classifiers consist of 100 trees, since the performance did not improve by adding more trees.

For nearly all the arrhythmia types, the distribution between the two classes, i.e. true and false alarms, was skewed. To balance the class distribution, the classifiers were trained using a cost matrix C,

$$C = \begin{pmatrix} 0 & \frac{A_{\text{true}}}{A_{\text{false}}} \\ 5 & 0 \end{pmatrix}, \qquad \text{if } A_{\text{true}} \geqslant A_{\text{false}}$$

and

$$C = \begin{pmatrix} 0 & 1 \\ 5 \cdot \dfrac{A_{\text{false}}}{A_{\text{true}}} & 0 \end{pmatrix}, \qquad \text{if } A_{\text{true}} < A_{\text{false}}$$

where A_{true} is the number of true alarms and A_{false} the number of false alarms in the training set. $C(1, 2)$ is the penalty given for a misclassified false alarm and $C(2, 1)$ the penalty given for a misclassified true alarm. Misclassification of true alarms is a more severe error than misclassification of false alarms. The multiple five was adopted from the automated score computed by the Challenge test system, which is defined in the section 2.5. In the data set for VFB, there were only 6 true alarms compared to 52 false alarms. Hence, for the classifier for VFB, including the multiple of five would have increased the weight too much and was therefore omitted.

2.5. Performance evaluation

The performance of the algorithm was evaluated with three different measures: true positive rate (TPR) or sensitivity defined in equation (6), true negative rate (TNR), and a Challenge score which is a weighted accuracy. The TNR and the score are computed as

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{11}$$

and

$$\text{score} = \frac{100 \cdot (\text{TP} + \text{TN})}{\text{TP} + \text{TN} + \text{FP} + 5 \cdot \text{FN}}. \tag{12}$$

In the training set, the estimates for the performance measures were produced with k-fold cross-validation. The number of sets k was set to 10 when there were more than 10 samples in the smaller class. Otherwise, k was set to the size of the smaller class to ensure that there was at least one sample from both of the classes. The k sets were generated in a way that the class distribution in every set represents the class distribution of the training set. For the unrevealed test set, the performance measures were computed by the scoring system.

The algorithm was implemented and evaluated in Matlab 2014b (The MathWorks Inc., Natick, MA).

3. Results

The results for the algorithm are listed in table 2. On the left are the results without using information about the median N_l around the R-peaks in ECG and on the right when information about the N_l is added.

In the training set, adding the local noise level improved the results for two arrhythmia types: ASY and VTA. With the best performing feature combinations, scores of 85 or higher were achieved for all the arrhythmias except VTA. The TPRs were 83–98% and the TNRs 63–94%.

The best results in the test set were achieved when the local noise level was added to the features, except for EBR where no change occurred. The overall TPRs were 93 and 95% depending on whether the test set was real-time or when additional retrospective data after the alarm was included. The overall TNRs were 80 and 83% and the scores 77.39 and 81.58 for real-time and retrospective data, respectively.

Table 2. Results without / with local noise level.

	Training set			Test set		
Arrhythmia	TPR	TNR	Score	TPR	TNR	Score
ASY	90 / 90	89 / 90	85.72 / **86.63**	83 / 89	97 / 98	88.62 / **92.02**
EBR	95 / 93	85 / 83	**85.00** / 80.46	92 / 92	72 / 72	71.56 / 71.56
ETC	98 / 98	88 / 88	**93.35** / 91.33	100 / 100	0 / 80	95.50 / **99.10**
VFB	83 / 67	94 / 94	**88.58** / 83.11	89 / 78	73 / 96	70.79 / **81.82**
VTA	86 / 89	66 / 63	62.57 / **63.14**	84 / 88	75 / 71	67.56 / **68.14**
Real-time	—	—	—	92 / 93	78 / 80	75.00 / **77.39**
Retrospective	—	—	—	93 / 95	84 / 83	79.20 / **81.58**

Note. The higher score between the two feature combinations is written in bold.

The best TPR was for ETC and was 100%, i.e. no true alarms were missed. The TNR for ETC was 80%. In the set there were 5 false alarms (Clifford *et al* 2015), which means that all except one false alarm were suppressed. The best TNR (98%) was for ASY. Based on the overall score, the performance in alarm classification was the best for ETC and then for ASY. The worst performance was for VTA both in the training and test set.

4. Discussion and conclusion

In this paper, an alarm classification algorithm was presented based on a signal comparison and selection with F_1-score, computation of arrhythmia relevant features, and a classification with Random Forest classifiers. The signal selection based on F_1-score does not use any signal specific quality information. The F_1-score gives a value that represents how well beats detected in the signals are in agreement with each other. Therefore, F_1-score provides a means to quantify the reliability of features based on detected beats in two signals. This is in contrast to other methods that rely on signal quality estimation of single signals. Different beat detectors may perform differently in the presence of different types of noise and an SQI does not necessarily include the information on how reliable the beat detection is.

The alarm classification when having only F_1-score as a quality measure gave relatively good results. The overall TPR was 92% and overall score 75.00. When both F_1-score and the median local noise level were used as quality indicators, the overall score increased to 77.39 and overall TPR of the algorithm was 93% in real-time events. This TPR is nearly as good as the best TPR achieved (94%) with the same test set in the Challenge (September 2015) (Clifford *et al* 2015). Misclassifying a true alarm as a false alarm is more severe than not suppressing a false alarm. From a clinical point of view the TPRs should be further increased for alarm suppression algorithms. The score computed for overall comparison of the algorithms weighted misclassified true alarms five times more severe than misclassified false alarms. A higher weight for false negative classifications in the evaluation score could be also considered.

Interestingly, for bradycardia the results remained the same independent of the addition of the local noise level. Moreover, the results are poorer in the test set. Looking at the results in the Challenge, Krasteva *et al* (2015) had the best reported score (93.81) for bradycardia alarm classification in the test set. This score was achieved by including features from both ECG and from pulsatile signals in the classification. Their results were better in the test set than in the training set. They also report a relatively large decrease in performance in the test set compared to the training set when only ECG based features are used. This could indicate that ECG based features are not sufficient for accurate classification of bradycardia alarms and that

there might be information in the pulsatile signals that is better represented in the test set than in the training set. Adding a feature from pulsatile signals could improve the classification of bradycardia alarms.

For three arrhythmia types, the results improved when the performance was evaluated on the hidden test set. Usually, the results in the training set are better because classifiers are optimized for the training data. This suggests that the training set might not represent the test set completely and the data of the test set might be less noisy. Moreover, the amount of data particularly for VFB and ETC was very small in one of the classes, containing only 8 ETC false alarms and 6 VFB true alarms (Clifford *et al* 2015). Increasing the amount of data could help in producing more robust solutions.

A system with a hidden test data for evaluating the algorithm performances enables a consistent and an objective way to compare different solutions. Using a closed system, however, has also its limitations. Further analysis of the results remains limited because the cases where the algorithm fails in the test set cannot be seen and the differences in data between training and test set cannot be analyzed. In addition, the complete information how the results are produced is not available. Some of the scores given by the system do not seem to comply with the number of false and true alarms reported in Clifford *et al* (2015) if the scores are simply calculated according to the equation (8). The cause for this discrepancy remains unknown.

The classification model was selected as the Random Forest for all the arrhythmias. In the solution of Hoog Antink and Leonhardt (2015) different machine learning techniques were used depending on the arrhythmia. In their approach, very different strategies depending on arrhythmia provided an optimal solution. The performance of our algorithm could be improved by optimizing the selection of the classifier for every type of arrhythmia separately.

The amount of false alarms was 61% and 69% in the training and test set, respectively. The distribution of false and true alarms varies between arrhythmia types, but for four out of five types the number of false alarms was greater than the number of true alarms. The data is collected from monitors from three different manufactures and from four different hospitals in USA and Europe. Clifford *et al* (2015) Hence, the data does not represent one particular manufacturer or hospital, but a more general situation. There is a need for alarm reduction algorithms, and our algorithm based on signal selection from multiple signals and alarm classification by machine learning provides promising results. Further improvements can be made by possibly adding features and selecting the classifiers for every arrhythmia separately.

Acknowledgments

The authors would like to thank L Dekker (PhD MD, cardiologist at the Catharina Hospital Eindhoven) and C van Pul (PhD MSc, clinical physicist at the MMC Hospital Veldhoven) for sharing their expertise.

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. Inform.* **41** 442–51

Balogh D, Kittinger E, Benzer A and Hackl J M 1993 Noise in the ICU *Intensive Care Med.* **19** 343–6

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

Borowski M, Siebig S, Wrede C and Imhoff M 2011 Reducing false alarms of intensive care online-monitoring systems: an evaluation of two signal extraction algorithms *Comput. Math. Methods Med.* **2011** 1–11

Breiman L 2001 Random Forests *Mach. Learn.* **45** 5–32

Clifford G, Aboukhalil A, Sun J, Zong W, Janz B, Moody G and Mark R 2006 Using the blood pressure waveform to reduce critical false ECG alarms *Computers in Cardiology* pp 829–32

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 *Computing in Cardiology* pp 273–6

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

Daluwatte C, Johannesen L, Vicente J, Scully C G, Galeotti L and Strauss D G 2015 Heartbeat fusion algorithm to reduce false alarms for arrhythmias *Computing in Cardiology* pp 745–8

Deshmane A V 2009 False arrhythmia alarm suppression using ecg, abp, and photoplethysmogram *Master's Thesis* Massachusetts Institute of Technology

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 *Computing in Cardiology* pp 293–6

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

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

Graham K C and Cvach M 2010 Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms *Am. J. Crit. Care* **19** 28–34

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

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

Konkani A and Oakley B 2012 Noise in hospital intensive care units—a critical review of a critical topic *J. Crit. Care* **27** 522e1–9

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

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

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 *Physiol. Meas.* **29** 15–32

Mäkivirta A, Koski E, Kari A and Sukuvaara T 1991 The median filter as a preprocessor for a patient monitor limit alarm system in intensive care *Comput. Methods Prog. Biomed.* **34** 139–44

Pimentel M A F, Santos M D, Springer D B and Clifford G D 2015 Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices *Physiol. Meas.* **36** 1717–27

Plesinger 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 *Computing in Cardiology* pp 281–4

Rooijakkers M J, Rabotti C, Oei S G and Mischi M 2012 Low-complexity R-peak detection for ambulatory fetal monitoring *Physiol. Meas.* **33** 1135–50

Sayadi O and Shamsollahi M B 2011 Life-threatening arrhythmia verification in ICU patients using the joint cardiovascular dynamical model and a bayesian filter *IEEE Trans. Biomed. Eng.* **58** 2748–57

Sieben W and Gather U 2007 Classifying alarms in intensive care—analogy to hypothesis testing *Artificial Intelligence in Medicine* pp 130–8

Siebig S, Kuhls S, Imhoff M, Gather U, Schölmerich J and Wrede C E 2010 Intensive care unit alarms—how many do we need? *Crit. Care Med.* **38** 451–6

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

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

Tsien C L, Kohane I S and McIntosh N 2000 Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit *Artif. Intell. Med.* **19** 189–202

van Rijsbergen C J 1979 *Information Retrieval* 2nd edn (London: Butterworths)
Zhang Y and Szolovits P 2008 Patient-specific learning in real time for adaptive monitoring in critical care *J. Biomed. Inform.* **41** 452–60
Zong W, Heldt T, Moody G and Mark R 2003 An open-source algorithm to detect onset of arterial blood pressure pulses *Computers in Cardiology* pp 259–62
Zong W 2015 Reduction of false critical ECG alarms using waveform features of arterial blood pressure and / or photoplethysmogram signals *Computing in Cardiology* pp 289–92