IOP Publishing Physiological Measurement

Physiol. Meas. **33** (2012) 1503–1516 doi:10.1088/0967-3334/33/9/1503

Robust classification of neonatal apnoea-related desaturations

Violeta Monasterio[1,2,3], **Fred Burgess**[3] **and Gari D Clifford**[3]

[1] CIBER de Bioingenieria, Biomateriales y Nanomedicina (CIBER-BBN), Zaragoza, Spain
[2] Aragon Institute of Engineering Research, Universidad de Zaragoza, Zaragoza, Spain
[3] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK

E-mail: violeta.monasterio@unizar.es

Received 6 February 2012, accepted for publication 9 May 2012
Published 17 August 2012
Online at stacks.iop.org/PM/33/1503

Abstract

Respiratory signals monitored in the neonatal intensive care units are usually ignored due to the high prevalence of noise and false alarms (FA). Apneic events are generally therefore indicated by a pulse oximeter alarm reacting to the subsequent desaturation. However, the high FA rate in the photoplethysmogram may desensitize staff, reducing the reaction speed. The main reason for the high FA rates of critical care monitors is the unimodal analysis behaviour. In this work, we propose a multimodal analysis framework to reduce the FA rate in neonatal apnoea monitoring. Information about oxygen saturation, heart rate, respiratory rate and signal quality was extracted from electrocardiogram, impedance pneumogram and photoplethysmographic signals for a total of 20 features in the 5 min interval before a desaturation event. 1616 desaturation events from 27 neonatal admissions were annotated by two independent reviewers as true (physiologically relevant) or false (noise-related). Patients were divided into two independent groups for training and validation, and a support vector machine was trained to classify the events as true or false. The best classification performance was achieved on a combination of 13 features with sensitivity, specificity and accuracy of 100% in the training set, and a sensitivity of 86%, a specificity of 91% and an accuracy of 90% in the validation set.

Keywords: neonatal apnoea monitoring, oxygen desaturation, multimodal analysis, classification, support vector machines, signal quality indices, ECG, PPG, respiratory rate

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

1. Introduction

In premature infants, immaturity of respiratory control almost invariably results in respiratory pauses (apnoeas) of variable duration that may require pharmacological intervention or ventilatory support (Martin and Abu-Shaweesh 2005, Halbower 2008). A close temporal relationship between apnoea, bradycardia (slow heart rate) and desaturation (low oxygen saturation in arterial blood) has been described in these infants, although the succession of these three events is variable and complex (Poets 2010). This condition is known as apnoea of prematurity.

Conventional monitoring in the neonatal intensive care unit (NICU) analyses respiratory and electrocardiographic signals separately to detect cessations of breathing effort and large changes in the heart rate (HR) respectively. Additionally, pulse oximetry monitoring of the peripheral oxygen saturation (SpO_2) from the photoplethysmographic signal (PPG) provides an alarm trigger when the SpO_2 falls below a pre-defined value.

An important limitation of pulse oximetry monitors is the high rate of false alarms (FA), produced by bad connections, poor sensor contact and unimodal data analysis (Chambrin 2001). FA rates associated with pulse oximeters higher than 70% have been reported in the literature (Petterson *et al* 2007). Voluntary and involuntary movements in the neonate such as kicking, stretching, crying and imposed motion (Tobin *et al* 2002) cause motion artefacts which lower the signal-to-noise ratio of the SpO_2 series and produce inaccurate readings. In this work we present a method to distinguish whether low SpO_2 readings in NICU monitors are caused by motion artefacts, termed *FA* from here on, or have a true physiological origin, termed *apnoea-related events* from here on. (Of course, low O_2 saturation can be due to poor perfusion, but the rate of change of O_2 levels is extremely slow.)

In the literature on signal processing for apnoea analysis, most studies focus on the diagnosis of the obstructive sleep apnoea syndrome, either in adults (Marcos *et al* 2009, Alvarez *et al* 2010, Khandoker *et al* 2009) or in children (Gil *et al* 2009, 2010). Fewer studies focus on the detection or classification of apneic episodes. For example, several recent works (Acharya *et al* 2011, Bsoul *et al* 2011, Khandoker and Palaniswami 2011) applied machine learning techniques to classify apnoea/hypoapnoea and normal epochs, based solely on the electrocardiogram (ECG). Comparatively, the detection of FA in apnoea monitors has received scarce attention (Belal *et al* 2011).

The method proposed in this work is applicable to NICU monitors that include respiratory, electrocardiographic and photoplethysmographic signals. A support vector machine (SVM) classifies each desaturation event as *FA* or *apnoea-related* based on information about the instantaneous values and changes of the HR, respiration rate (RR), oxygen saturation (SpO_2), and also based on the information on the quality of the monitored signals, which can be particularly useful for dealing with noisy or missing data (Clifford *et al* 2009). Earlier works which employed signal quality measures (Zong *et al* 2004, Aboukhalil *et al* 2008) addressed the FA issue by using unimodal signal quality metrics to decide if the information from a given signal could be trusted. In this work we describe a framework which uses both features and signal quality metrics simultaneously from multiple channels, thereby taking advantage of the covariant structure of the noise and data to produce a more accurate FA reduction system.

2. Materials and methods

2.1. Data set

Data for this study were extracted from the Multi-Parameter Intelligent Monitoring for Intensive Care II (MIMIC II) database (Saeed *et al* 2011), which is available from the PhysioNet

archives (Goldberger *et al* 2000). The MIMIC II database contains bedside monitor trends and physiological waveforms from over 3500 NICU patients hospitalized at Beth Israel Deaconess Medical Center, Boston, USA between 2004 and 2007. In this study, we analysed the data recorded during 27 randomly selected stays in the NICU. Data consisted of four physiological waveforms sampled at 125 Hz—two leads of ECG, impedance pneumogram (IP) and pulse photoplethysmogram (PPG)—as well as two 1 Hz derived parameter time series provided by bedside monitors—the HR derived from the ECG and the peripheral oxygen saturation (SpO_2) derived from the PPG. These specific waveforms and time series are usually but not always available in individual MIMIC II recordings. Figure 1 presents excerpts from two patients' data, one of them showing a classic apneic episode, and the other one showing a FA. The MIMIC II database does not provide demographic or clinical information for neonatal patients.

2.2. Methods

First, a set of reference annotations was created as a gold standard to evaluate the performance of the FA detection algorithm. Then, various variables were computed from the data to characterize the trends and changes of physiological parameters and the quality of the signals. Each variable was independently analysed with a receiver operating characteristic (ROC) curve to find the optimum evaluation point in a 300 s interval preceding each desaturation event. Finally, a multivariate classification strategy based on SVMs was evaluated. Each of these steps is explained in the following sections.

2.2.1. Annotation of desaturation events. There is no consensus among neonatologists as to what constitutes a safe SpO_2 level. Acceptable SpO_2 levels may vary with the developmental stage, and target values ranging from 85% to 95% can be found in the literature (Finer and Leone 2009). The intermediate value within this range, 90%, was considered as the limit to trigger desaturation alarms in this work. Desaturation events were thus defined as those intervals where $SpO_2 < 90\%$.

Two investigators independently annotated 1880 desaturation events from 27 NICU stays. For each event, the investigators decided among three options: (1) the desaturation was associated with an apnoea (positive event), (2) the desaturation was caused by noise or artefacts (negative event, that is, a FA), or (3) it could not be determined whether the desaturation is associated with an apnoea or not (unsure). Option (1) was chosen if the following conditions were fulfilled: within the interval of 300 s before the desaturation event (a) the HR decreased at least 10 beats per minute (bpm), (b) the minimum HR was < 130 bpm, and (c) on visual inspection the quality of ECG and PPG waveforms appeared to be high, so that one would expect the waveforms to provide reliable parameter estimates, and no obvious artefacts were present. Option (2) was chosen if high levels of noise and/or artefacts were clearly visible in the signals. Option (3) was chosen otherwise.

The two annotators agreed for 1616 (86%) events, which were then used as the reference set of annotations for classification: 316 positive (apnoea-related) events and 1300 negative (noise-related) events. This reference set was split into training and validation subsets for SVM analysis. The training subset comprised 14 NICU stays, with a total of 158 positive and 638 negative events (80% of events labelled as FA), and the validation subset comprised the other 13 stays, with 158 positive and 662 negative events (81% of events labelled as FA); in this way, the validation and training data were made independent.

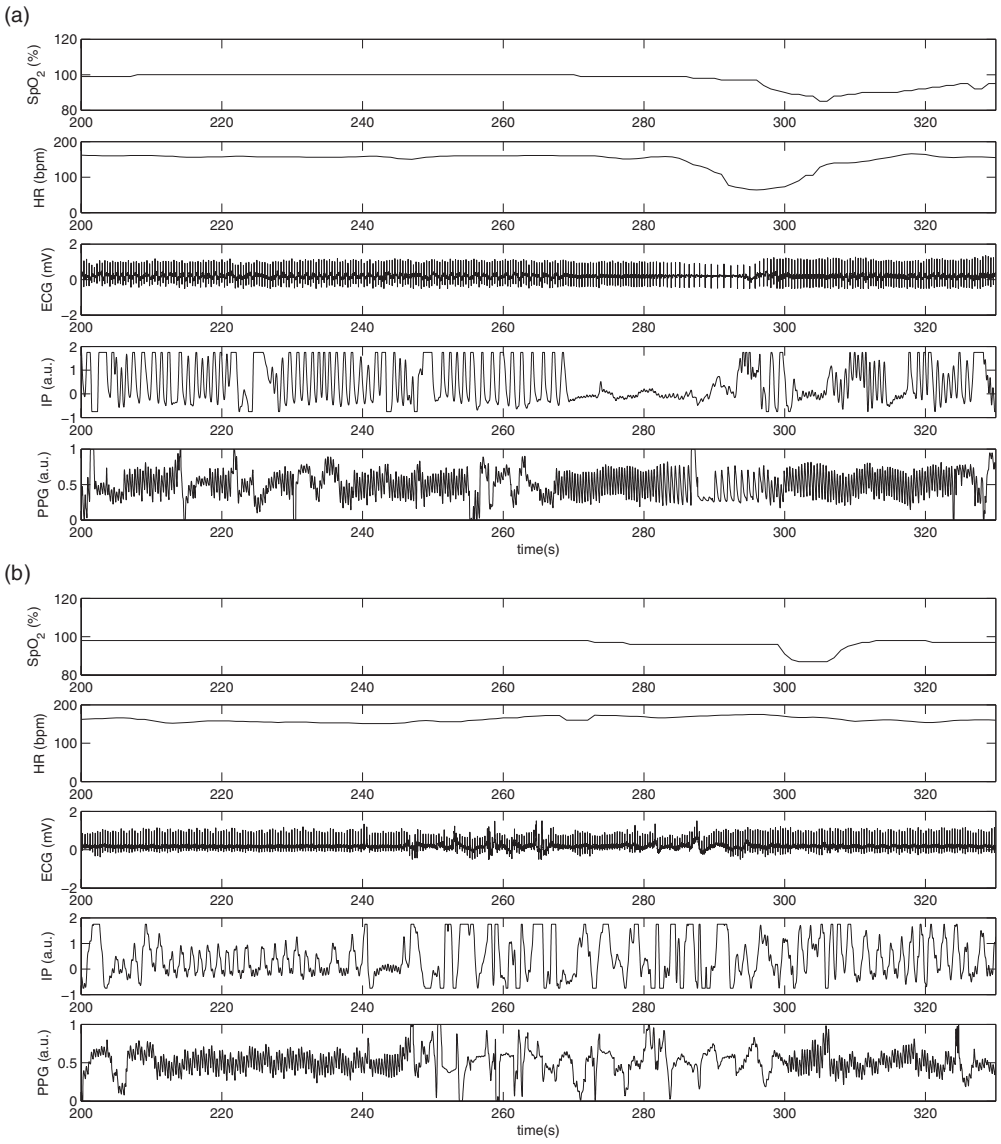

Figure 1. (a) Excerpt of SpO_2, HR, ECG, PPG and IP tracings during an apneic event. A cessation of respiration can be observed in IP signal at $t = 270$ s, followed by bradycardia around 20 s later; oxygen saturation falls below 90% at $t = 300$ s. (b) Excerpt of SpO_2, HR, ECG, PPG and IP tracings showing a noise-related desaturation (a FA) at $t = 300$ s. Tracings show no signs of bradycardia, and high amounts of noise are visible in IP and PPG signals before the desaturation (from $t = 250$ to $t = 300$ s) (a.u., arbitrary units).

2.2.2. Computation of physiological variables. Four groups of variables were computed: variables related to oxygen saturation, variables related to HR, variables related to RR and variables related to the quality of the signals. A total of 20 variables were computed every 5 s for the 300 s interval before each desaturation event as follows.

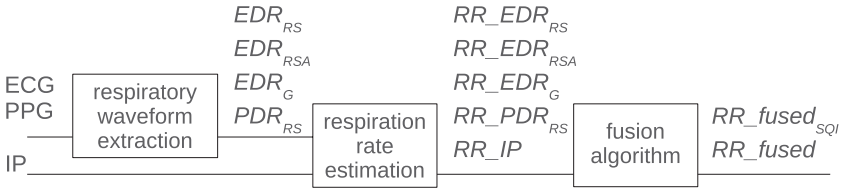

Figure 2. Computation of RR series: first, surrogate respiratory signals were derived from ECG and PPG waveforms; then, RRs were estimated from respiratory signals using an AR model; finally, robust RR estimates were computed by combining the individual RR estimates with a fusion algorithm.

Variables related to HR and oxygen saturation. Variables related to HR and SpO_2 were derived from HR and SpO_2 1 Hz series. The 300 s interval before each desaturation event was analysed with a running window of 20 s, with a sliding step of 5 s. In each 20 s window, the minimum value and the slope of HR and SpO_2 series were computed. These variables were denoted as $min\ HR$, ∇HR, $min\ SpO_2$ and ∇SpO_2 respectively. The slopes were computed using ordinary least-squares regression (LSR) over the 20 s window. Robust LSR was disregarded because it did not improve the results over ordinary LSR and entailed a higher computational cost.

Variables related to RR. Variables related to RR were computed in several steps (figure 2). First, respiratory signals were derived from ECG and PPG waveforms as follows. ECG beats were detected using an open source implementation of Hamilton and Tompkins' QRS detector (Hamilton and Tompkins 1986). Then, three widely used methods (Moody *et al* 1985) were applied for estimating a respiratory signal from the ECG-derived respiration (EDR): a method based on the QRS area summation (EDR_G), a method based on R-S amplitude tracking (EDR_{RS}) and a method based on the estimation of respiratory sinus arrhythmia (EDR_{RSA}). Since the PPG waveform exhibits amplitude fluctuations due to respiration, a similar approach to EDR_{RS} was considered, and the differences between successive peaks and valleys in the signal were computed to estimate a PPG-derived respiratory signal (PDR_{RS}). PPG peak detection was performed using an open source beat detector for arterial blood pressure signals (Zong *et al* 2003) with a time and amplitude threshold adjustment to fit PPG beat width and height (Li and Clifford 2012).

Second, RR was estimated from each derived respiratory signal and from IP signal using a RR extraction algorithm (Nemati *et al* 2010) based on the work of Mason and Tarassenko (Mason and Tarassenko 2001, Mason 2002), who used autoregressive (AR) modelling to estimate the respiratory frequency in adults. Since RR is usually higher in neonates than in adults, the upper bound for RR was increased from 55 bpm (original algorithm) to 70 bpm in this work. The resulting RR estimations were denoted as RR_EDR_{RS}, RR_EDR_{RSA}, RR_EDR_G, RR_PDR_{RS} and RR_IP. First and second steps were performed for the 300 s interval before each desaturation event using a running window of 20 s with a sliding step of 5 s. Figure 3 presents an excerpt of IP, ECG-derived and PPG-derived respiratory signals for the patient in figure 1(a), together with the corresponding RR estimates.

Third, an improved RR estimate was computed using the data fusion algorithm developed by Nemati *et al* (2010) and Li *et al* (2008). This method is an application of a modified Kalman filter (KF) framework for data fusion to the estimation of RR from multiple physiological sources. KF were employed to obtain independent RR estimates from the series of derived RR,

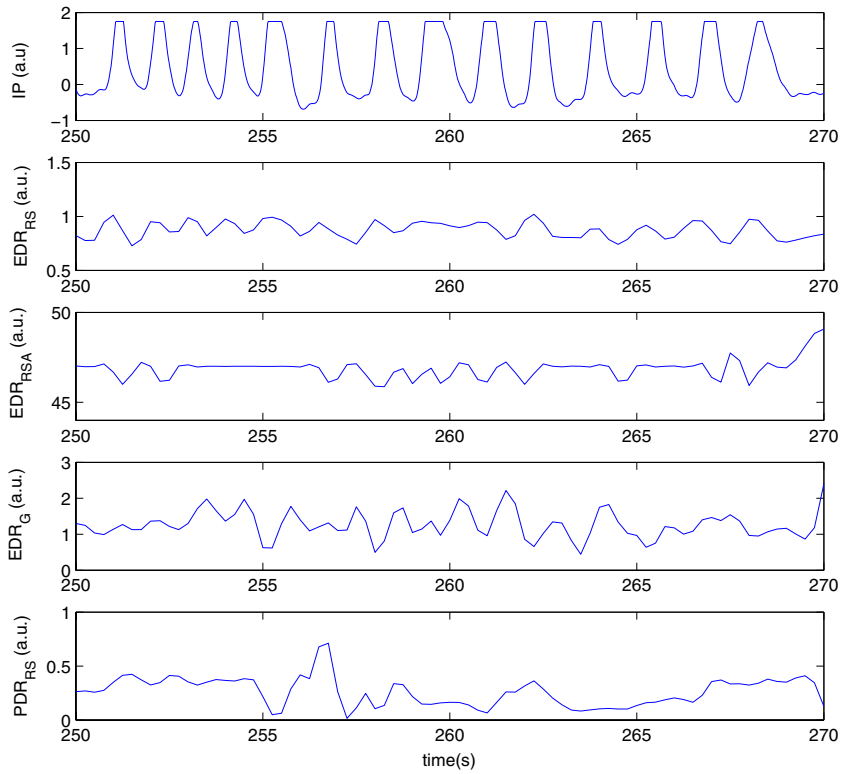

Figure 3. IP, ECG-derived and PPG-derived respiratory signals corresponding with the interval between 250 and 270 s in figure 1(a). RR estimates for segments shown above were 42 bpm for IP, 42 bpm for EDR_{RS}, 45 bpm for EDR_{RSA}, 44 bpm for EDR_{G} and 21 bpm for PDR_{RS} (a.u., arbitrary units).

and then the independent estimates were fused taking into account the uncertainty associated with each estimate. In this work, the fusion algorithm was applied to the series of derived RR for the 300 s interval before each desaturation event, and the result was denoted as *RR_fused*.

In Nemati *et al* (2010), the authors also proposed a variation of the fusion algorithm that makes use of signal quality indexes (SQI), which are explained in the following section. SQI are incorporated into the computation of the individual KF and into the fusion step to obtain a more robust RR estimation. In this work, we applied the fusion algorithm with SQI to the series of derived RR for the 300 s interval before each desaturation, and denoted the result as RR_fused_{SQI}.

Finally, we computed the minimum RR (*min RR*) and the slope of all RR series (∇RR) every 15 s for the 300 s interval before each desaturation event.

Variables related to signal quality. The selected index for determining the quality of PPG, IP and derived respiratory signals is the *spectral purity*, an approach proposed in Nemati *et al* (2010). The spectral purity of a signal is defined as (Sornmo and Laguna 2005)

$$\Gamma_s = \frac{\omega_2^2}{\omega_0 \omega_4}, \tag{1}$$

where ω_n is the nth-order spectral moment defined as

$$\omega_n = \int_{-\pi}^{\pi} \omega^n P(\mathrm{e}^{j\omega})\, \mathrm{d}\omega, \tag{2}$$

where $P(\mathrm{e}^{j\omega})$ is the power spectrum of the signal. In the case of a periodic signal with a single dominant frequency, Γ_s takes the value of one and approaches zero for non-sinusoidal noisy signals. Therefore, in an ideal respiratory waveform we would expect $\Gamma_s = 1$. The spectral purity was computed for PPG, IP and derived respiratory signals for the 300 s interval before each desaturation using a running window of 20 s with a sliding step of 5 s.

To determine the quality of the ECG, we followed the approach proposed in Li *et al* (2008) and computed the fourth moment (kurtosis) of the ECG signal using a running window of 20 s with a 5 s sliding step, and denoted the result as *kECG* .

2.2.3. Computation of features with univariate ROC analysis. The temporal relation between apnoea, desaturation and bradycardia is not completely understood, and significant changes in HR, RR and SpO_2 do not necessarily appear at the same time before an apnoea-related desaturation event. Therefore, we independently analysed each variable to find the evaluation interval that maximized the univariate classification of desaturation events for that variable. To do so, we defined 20 time windows within the 300 s interval before each event. The end point was defined for all windows as the beginning of the desaturation event (t_{end}), and the starting point of each window k was defined as $t_{\mathrm{end}} - 15k$ s (window 1 comprised the 15 s before desaturation, window 2 comprised the 30 s before desaturation, and so on). Within each window k, the minimum value of the variable was selected for all desaturation events and a ROC curve was constructed. The resulting area under the curve (AUC) was a measure of the classification performance of the variable at the selected interval k. This process was repeated for all k windows, and the window with the maximum AUC was selected as the optimum evaluation interval for the variable.

Finally, features for SVM classification were selected as the minimum value of each variable within its optimum evaluation interval. Features were named as the corresponding variables without the cursive; for example, 'min HR' denotes the feature computed as the minimum of variable '*min HR*' within its optimum evaluation interval.

2.2.4. Feature selection. Among the 20 features resulting from ROC analysis, it was not known which of them were most relevant, and which were irrelevant or redundant for FA detection. For classification purposes, reducing the number of input features by selecting only the relevant ones usually leads to higher performance with lower computational effort. Therefore, a feature selection algorithm was applied before performing SVM classification.

In general, two types of feature selection methods have been proposed in the literature: filter methods and wrapper methods. The essential difference between them is that a wrapper method depends on the algorithm that is used to build the final classifier, while a filter method does not (Saeys *et al* 2007). In this work we applied a filter method, the minimum redundancy maximum relevance (mRMR) method (Peng *et al* 2005), which computes a rank of the most relevant features using mutual information metrics. Mutual information methods for feature selection usually compute the utility of each feature by evaluating the feature's own mutual information, its correlation with the rest of existing features and a term which depends on class-conditional probabilities (Brown 2009). In particular, the class-conditional term is omitted in mRMR. We denoted as F_k the feature with the kth highest rank as computed by the mRMR algorithm, and then we defined 20 subsets of features as $S_k = \{F_1, \ldots, F_k\}$, that is, S_1

comprised the feature with the highest rank, S_2 comprised the features with the two highest ranks, and so on.

2.2.5. SVM classification. Consider the problem of separating the set of training vectors belonging to two classes, $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i \in R^n$ is an input vector and $y_i \in \{+1, -1\}$ is a label that determines the class of $\mathbf{x}_i$. The objective of a SVM is to find the separating hyperplane with a maximal margin (Cortes and Vapnik 1995), which can be expressed as the following minimization problem:

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i, \tag{3}$$

$$\text{subject to } \ y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geqslant 1 - \xi_i, \tag{4}$$

$$\xi_i \geqslant 0, \tag{5}$$

where $\mathbf{w}$ and b define the separating hyperplane, ξ_i are 'slack' variables which allow for misclassified vectors and ϕ is a function that maps the training vectors $\mathbf{x}_i$ into a higher dimensional space. In the SVM literature, the term 'feature' may denote either the result of the mapping $\phi(\mathbf{x}_i)$, or each one of the n elements of the input vector $\mathbf{x}_i$. In this paper we adopted the second use.

An explicit definition of the mapping function ϕ can be avoided by using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. In this work we used a radial basis function (RBF) kernel, defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2), \quad \gamma > 0. \tag{6}$$

An RBF kernel has been found to improve classification results over a linear kernel in most cases (Chang and Lin 2011). However, to define an RBF kernel it is necessary to select an appropriate value for γ. In practice, suitable values for γ and C can be found empirically by means of a grid search.

In this work, we performed a grid search to find the optimum RBF parameters for each subset of features S_k as follows:

(i) Consider a grid space of (C, γ) with $\log_2 C \in \{-5, -4, \ldots, 15\}$ and $\log_2\gamma \in \{-15, -14, \ldots, 3\}$.
(ii) For each pair (C, γ) in the space, perform ten-fold cross validation (CV) on the training set.
(iii) Choose the pair (C, γ) that produces the maximum mean CV accuracy.

For each subset of features S_k, we used the selected pair (C, γ) to train the RBF–SVM with the whole training set, and tested the final performance of the classifier with the validation set.

Every time the classifier was trained and tested (either with data subsets for CV, or with the whole sets for the final evaluation), the corresponding training and validation inputs were normalized, and the penalty parameter C was scaled as follows. Training inputs were normalized so that each input feature had zero mean and unit variance, and validation inputs were scaled to the same scaling factors than training inputs. To account for the imbalance between positive and negative classes in the dataset, the penalty associated with misclassification (C) was multiplied by a factor r for positive events, and by a factor of $1/r$ for negative events, with r being equal to the ratio between negative and positive events in the training inputs.

Table 1. Results of ROC analysis: maximum AUC and optimal evaluation interval (window) for each variable.

Variable	AUC	Window	Sign
min SpO_2	0.59	7	–
∇SpO_2	0.68	3	–
min HR	0.96	2	–
∇HR	0.91	4	–
min RR_EDR_{RS}	0.55	2	–
min RR_EDR_{RSA}	0.69	2	–
min RR_EDR_{G}	0.56	2	–
min RR_PDR_{RS}	0.57	8	–
min RR_IP	0.58	19	–
min RR_fused	0.57	2	–
∇RR_fused	0.59	5	–
min RR_fused_{SQI}	0.58	2	–
∇RR_fused_{SQI}	0.61	4	–
kECG	0.78	2	+
SQI_PPG	0.68	3	+
SQI_IP	0.64	19	–
SQI_EDR_{RS}	0.57	4	–
SQI_EDR_{RSA}	0.72	5	–
SQI_EDR_{G}	0.52	4	–
SQI_PDR_{RS}	0.58	4	–

3. Results

3.1. Computation of features with univariate ROC analysis

Results of the univariate ROC analysis are presented in table 1, which contains the optimum evaluation window for each variable and the corresponding AUC. A positive (negative) sign in the third column indicates that values above (below) the discrimination threshold are classified as positive events.

The maximum AUC, 0.96, was obtained for the minimum HR within the interval of 30 s before the desaturation event (variable *min HR* at window 2). The second highest AUC was obtained for the minimum slope of HR within the interval of 120 s before the desaturation event (variable ∇HR at window 4) (table 1).

3.2. Feature selection

Prior to RBF–SVM classification, we computed the rank of most relevant features by applying the mRMR algorithm to the training set (table 2). The four most relevant features were min HR, SQI_PPG, SQI_EDR_{RSA} and ∇HR.

3.3. SVM classification

A grid search was conducted for each subset of features to find the optimum RBF parameters (table 3). The resulting classification performances in training and validation sets are presented in tables 4 and 5.

Not all features could be computed for every desaturation event for two reasons. First, there were intermittently missing data in all signals, and second, the appearance of successive desaturation events separated by less than 20 s (minimum interval for variable

Table 2. Feature ranking according to mRMR.

Feature	Rank
min HR	1
SQI_PPG	2
SQI_EDR_{RSA}	3
∇HR	4
∇SpO_2	5
SQI_PDR_{RS}	6
SQI_EDR_{RS}	7
SQI_IP	8
SQI_EDR_{G}	9
∇RR_fused_{SQI}	10
kECG	11
∇RR_fused	12
min SpO_2	13
min RR_fused	14
min RR_IP	15
min RR_PDR_{RS}	16
min RR_fused_{SQI}	17
min RR_EDR_{G}	18
min RR_EDR_{RSA}	19
min RR_EDR_{RS}	20

Table 3. Optimization parameters for RBF–SVM. The accuracy in the training set from cross validation is reported as mean ± standard deviation.

Features	$\log_2 C$	$\log_2 \gamma$	Accuracy
S_1	13	3	91.7 ± 1.6
S_2	5	3	88.7 ± 2.5
S_3	13	−9	88.8 ± 2.9
S_4	3	1	91.5 ± 2.8
S_5	11	−1	93.0 ± 3.4
S_6	9	−1	94.4 ± 2.5
S_7	3	−1	93.6 ± 2.5
S_8	1	−1	93.3 ± 2.8
S_9	3	−3	93.7 ± 2.0
S_{10}	3	−3	94.4 ± 3.1
S_{11}	7	−3	95.3 ± 2.7
S_{12}	3	−3	94.8 ± 2.6
S_{13}	5	−3	95.1 ± 2.3
S_{14}	13	−3	95.8 ± 1.2
S_{15}	5	−3	95.1 ± 2.0
S_{16}	11	−5	94.8 ± 1.9
S_{17}	5	−5	94.4 ± 1.5
S_{18}	3	−5	94.8 ± 2.3
S_{19}	3	−5	95.1 ± 3.1
S_{20}	11	−7	94.8 ± 2.0

computation) was frequent. Columns 'Positive' and 'Negative' in tables 4 and 5 show the number (percentage) of events in which all features of the corresponding combination could be computed.

The highest accuracy in the validation set (90.2%) was obtained with a subset of 13 features (S_{13}, that is, those features with ranks 1–13 in table 2), which included all features related to HR and SpO_2, the slope of the RR (∇RR_fused_{SQI} and ∇RR_fused), and all features related to the quality of the signals.

Table 4. RBF–SVM classification results for the training set. Features were selected according to the rank in table 2. Columns 'Positive' and 'Negative' show the number (percentage) of classified events. S: sensitivity, Sp: specificity, PPV: positive predictive value and NPV: negative predictive value.

Features	Positive	Negative	S	Sp	PPV	NPV	Accuracy
S_1	148 (93.7)	542 (85.0)	93.2	91.3	74.6	98.0	91.7
S_2	148 (93.7)	542 (85.0)	96.6	88.9	70.4	99.0	90.6
S_3	147 (93.0)	507 (79.5)	94.6	87.0	67.8	98.2	88.7
S_4	147 (93.0)	502 (78.7)	98.6	95.2	85.8	99.6	96.0
S_5	147 (93.0)	492 (77.1)	100.0	100.0	100.0	100.0	100.0
S_6	147 (93.0)	492 (77.1)	100.0	100.0	100.0	100.0	100.0
S_7	147 (93.0)	491 (77.0)	100.0	98.8	96.1	100.0	99.1
S_8	147 (93.0)	491 (77.0)	100.0	98.8	96.1	100.0	99.1
S_9	147 (93.0)	491 (77.0)	100.0	96.9	90.7	100.0	97.6
S_{10}	147 (93.0)	491 (77.0)	100.0	98.4	94.8	100.0	98.7
S_{11}	140 (88.6)	436 (68.3)	100.0	100.0	100.0	100.0	100.0
S_{12}	140 (88.6)	436 (68.3)	100.0	99.5	98.6	100.0	99.7
S_{13}	140 (88.6)	436 (68.3)	100.0	100.0	100.0	100.0	100.0
S_{14}	140 (88.6)	436 (68.3)	100.0	100.0	100.0	100.0	100.0
S_{15}	140 (88.6)	436 (68.3)	100.0	100.0	100.0	100.0	100.0
S_{16}	140 (88.6)	436 (68.3)	100.0	100.0	100.0	100.0	100.0
S_{17}	140 (88.6)	436 (68.3)	100.0	99.5	98.6	100.0	99.7
S_{18}	140 (88.6)	436 (68.3)	100.0	98.2	94.6	100.0	98.6
S_{19}	140 (88.6)	436 (68.3)	100.0	98.2	94.6	100.0	98.6
S_{20}	140 (88.6)	436 (68.3)	100.0	100.0	100.0	100.0	100.0

Table 5. RBF–SVM classification results for the validation set. Features were selected according to the rank in table 2. Columns 'Positive' and 'Negative' show the number (percentage) of classified events. S: sensitivity, Sp: specificity, PPV: positive predictive value, NPV: negative predictive value. The combination producing the best accuracy in the validation set is marked in bold.

Features	Positive	Negative	S	Sp	PPV	NPV	Accuracy
S_1	143 (90.5)	577 (87.2)	91.6	84.2	59.0	97.6	85.7
S_2	143 (90.5)	576 (87.0)	88.8	83.2	56.7	96.8	84.3
S_3	143 (90.5)	575 (86.9)	92.3	79.7	53.0	97.7	82.2
S_4	143 (90.5)	570 (86.1)	85.3	90.5	69.3	96.1	89.5
S_5	143 (90.5)	564 (85.2)	81.8	90.6	68.8	95.2	88.8
S_6	143 (90.5)	564 (85.2)	82.5	91.0	69.8	95.4	89.3
S_7	143 (90.5)	564 (85.2)	82.5	91.5	71.1	95.4	89.7
S_8	143 (90.5)	564 (85.2)	81.8	91.8	71.8	95.2	89.8
S_9	143 (90.5)	564 (85.2)	83.2	89.0	65.7	95.4	87.8
S_{10}	143 (90.5)	564 (85.2)	82.5	89.7	67.0	95.3	88.3
S_{11}	123 (77.8)	406 (61.3)	84.6	90.1	72.2	95.1	88.8
S_{12}	123 (77.8)	406 (61.3)	87.8	89.7	72.0	96.0	89.2
S_{13}	**123 (77.8)**	**406 (61.3)**	**86.2**	**91.4**	**75.2**	**95.6**	**90.2**
S_{14}	123 (77.8)	406 (61.3)	82.9	91.9	75.6	94.7	89.8
S_{15}	123 (77.8)	406 (61.3)	81.3	91.4	74.1	94.2	89.0
S_{16}	123 (77.8)	406 (61.3)	85.4	88.9	70.0	95.3	88.1
S_{17}	123 (77.8)	406 (61.3)	87.8	88.4	69.7	96.0	88.3
S_{18}	123 (77.8)	406 (61.3)	92.7	86.2	67.1	97.5	87.7
S_{19}	123 (77.8)	406 (61.3)	92.7	86.7	67.9	97.5	88.1
S_{20}	123 (77.8)	406 (61.3)	87.0	88.4	69.5	95.7	88.1

4. Discussion

The best classification performance in the validation set, an accuracy of 90%, was obtained with a subset of 13 features (row S_{13} in table 5). The accuracy in the training set was 100% for the same combination of features (table 4). The most useful feature for FA detection was the minimum HR within the 30 s interval before a desaturation. This was the variable with the maximum AUC (table 1), and it was ranked as the most relevant feature by the mRMR algorithm (table 2). Just adding the information on the minimum HR made it possible to detect 84% of the FA in the training set, while maintaining a high sensitivity to apnoea-related events (row S_1 in table 5).

Instantaneous RR were found to be less useful for FA detection. Several causes may explain this fact. One of them is that, unlike in central apnoea, in obstructive apnoea the respiratory effort does not cease completely, and therefore a fraction of all apneic events may remain unidentified when RR is analysed independently from other variables. A respiratory signal based on air flow measures would allow the identification of obstructive apnoeas, but such signal was not available for this study. A second possible cause is the limited accuracy of the algorithms for deriving respiratory signals and RRs. It has been shown that respiratory estimation algorithms depend heavily on the quality of the signals and on the actual (true) RR being estimated. In particular, the RR values computed with the AR model used in this work tend to be more inaccurate at lower RRs (Nemati *et al* 2010). This limitation can be partly overcome by combining RR from different sources into a robust RR estimate, and by evaluating the slope instead of the instantaneous value. Indeed, the contribution of the feature ∇RR_fused$_{SQI}$ to the classification performance was higher than the contribution of any individual RR estimate (see tables 1 and 2).

The RBF–SVM algorithm mainly relied on HR and SQI information to classify events. The best subset of features, S_{13}, contained all the features related to the quality of signals. We can interpret this behaviour as a replication of the rules for annotating desaturation events (see section 2.2.1). This was expected, since the results of any classification algorithm can only be as good as the quality of the reference dataset (in terms of correctness of annotations and representativity of cases). Indeed, the quality of the reference dataset is always a key issue in studies on patient monitoring (Clifford *et al* 2009). In the literature, we can find examples where a reference dataset is created using a preliminary mark-up system followed by rejection of data subsets by expert clinicians. For example in Belal *et al* (2011), expert annotators labelled clusters of patient events found by a heuristic algorithm, rather than individual events. In our work, on the other hand, each event was individually (double) labelled to ensure no algorithmic bias, and to most closely map to the human understanding of the event. No pre-screening bias was therefore introduced and the technique is thus expected to generalize to unseen new data, as long as our examples of artefacts and true events are sufficiently representative.

Several limitations of the study need to be acknowledged. First, it should be noted that although over 1500 events were used, only 27 patients were actually included in the study. A larger cross-section of patients is likely to be needed, perhaps divided by age and prematurity. Second, the technique proposed in this paper is not intended for the unambiguous identification of sleep apnoea. The aim of the algorithm is to rule out false monitoring alarms that have no relation to the physiological state of the patient. The term *apnoea-related* is used in the paper to denote those desaturation episodes which are related to bradycardia and/or changes in RR in the absence of apparent noise, and therefore are likely to have a physiological origin, but it should not be understood as a formal definition of apnoea.

Another important limitation is that the best classification results could only be obtained for a reduced subset of events (note the percentage of classified events, 'Positive' and 'Negative'

columns in table 5). A possible solution for the practical implementation of the algorithm would be to label all unclassifiable events as *apnoea-related*, so that true apnoea-related events are not missed. This approach would decrease the classification accuracy in the validation set from 90% to 63%. Even in that case, the proposed technique would reduce the percentage of FA from 81% to 67% in the validation set, so it could still be pertinent for real time use in NICU apnoea monitoring. However, we expect that a large number of these events could be further labelled with expert over-reading and so increase the performance.

We also note that the selected 13-feature subset was the best combination among the tested possibilities, but different combinations not listed in the subsets may outperform the chosen one. An exhaustive search over every possible combination might thus improve results. However, the relative small variations in the final accuracy of the algorithm (see table 5) and the feature selection results suggest that higher potential improvements could be achieved by exploring further signal quality parameters.

5. Conclusions

The work presented in this paper demonstrates a general framework for fusing both features and signal quality metrics simultaneously from multiple channels. Such an approach exploits the covariant information in the noise and the data, thereby producing a more accurate false alarm reduction system. Results from this work indicate that the analysis of oxygen saturation, heart rate, respiration rate and the quality of the monitored signals with a support vector machine significantly reduces the number of pulse oximetry false alarms. Conventional apnoea monitors can be therefore greatly improved with the use of multimodal analysis and machine learning techniques.

Acknowledgments

This work was generously supported by The John Fell Fund and the University of Oxford (UK), by CIBER de Bioingenieria, Biomateriales y Nanomedicina through Instituto de Salud Carlos III and Fondo Europeo de Desarrollo Regional (Spain), and by Projects TEC2010-21703-C03-02 of Ministerio de Ciencia e Innovacion (MICINN) and GTC T30 from DGA (Spain). The authors would like to thank M Shahid and S Nemati for their code sharing and helpful collaboration.

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

Acharya U R, Chua E C P, Faust O, Lim T C and Lim L F B 2011 Automated detection of sleep apnea from electrocardiogram signals using nonlinear parameters *Physiol. Meas.* **32** 287–303

Alvarez D, Hornero R, Marcos J and del Campo F 2010 Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis *IEEE Trans. Biomed. Eng.* **57** 2816–24

Belal S Y, Emmerson A J and Beatty P C W 2011 Automatic detection of apnoea of prematurity *Physiol. Meas.* **32** 523–42

Brown G 2009 A new perspective for information theoretic feature selection *Proc. 12th Int. Conf. on Artificial Intelligence and Statistics* pp 49–56

Bsoul M, Minn H and Tamil L 2011 Apnea MedAssist: real-time sleep apnea monitor using single-lead ECG *IEEE Trans. Inf. Technol. Biomed* **15** 416–27

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

Chang C C and Lin C J 2011 LIBSVM: A library for support vector machines *ACM Trans. Intell. Syst. Technol.* http://www.csie.ntu.edu.tw/ cjlin/libsvm **2** 1–27

Clifford G D, Long W J, Moody G B and Szolovits P 2009 Robust parameter extraction for decision support using multimodal intensive care data *Phil. Trans. R. Soc.* A **367** 411–29

Cortes C and Vapnik V 1995 Support-vector networks *Mach. Learn.* **20** 273–97

Finer N and Leone T 2009 Oxygen saturation monitoring for the preterm infant: the evidence basis for current practice *Pediatr. Res.* **65** 375–80

Gil E, Bailon R, Vergara J M and Laguna P 2010 PTT variability for discrimination of sleep apnea related decreases in the amplitude fluctuations of PPG signal in children *IEEE Trans. Biomed. Eng.* **57** 1079–88

Gil E, Mendez M, Vergara J M, Cerutti S, Bianchi A M and Laguna P 2009 Discrimination of sleep-apnea-related decreases in the amplitude fluctuations of PPG signal in children by HRV analysis *IEEE Trans. Biomed. Eng.* **56** 1005–14

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

Halbower A C 2008 Pediatric home apnea monitors *Chest* **134** 425–9

Hamilton P S and Tompkins W J 1986 Quantitative investigation of QRS detection rules using the MIT/BIH Arrhythmia Database *IEEE Trans. Biomed. Eng.* **33** 1157–65

Khandoker A H and Palaniswami M 2011 Modeling respiratory movement signals during central and obstructive sleep apnea events using electrocardiogram *Ann. Biomed. Eng.* **39** 801–11

Khandoker A H, Palaniswami M and Karmakar C K 2009 Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings *IEEE Trans. Inf. Technol. Biomed.* **13** 37–48

Li Q and Clifford G D 2012 Dynamic time warping and machine learning for signal quality assessment of pulsatile signals *Physiol. Meas.* **33** 1491–1501

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

Marcos J V, Hornero R, Alvarez D, Del Campo F and Zamarron C 2009 A classification algorithm based on spectral features from nocturnal oximetry and support vector machines to assist in the diagnosis of obstructive sleep apnea *EMBC 2009: Proc. Engineering in Medicine and Biology Society* pp 5547–50

Martin R J and Abu-Shaweesh J M 2005 Control of breathing and neonatal apnea *Neonatology* **87** 288–95

Mason C L 2002 Signal processing methods for non-invasive respiration monitoring *PhD Thesis* (Engineering Science) University of Oxford, Oxford, UK

Mason C L and Tarassenko L 2001 Quantitative assessment of respiratory derivation algorithms EMBC 2001: *Proc. Engineering in Medicine and Biology Society* vol 2 pp 1998–2001

Moody G B, Mark R G, Zoccola A and Mantero S 1985 Derivation of respiratory signals from multi-lead ECGs *Proc. Computers in Cardiology* pp 113–6

Nemati S, Malhotra A and Clifford G D 2010 Data fusion for improved respiration rate estimation *EURASIP J. Adv. Signal Process.* **2010** 1–10

Peng H, Long F and Ding C 2005 Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy *IEEE Trans. Pattern Anal. Mach. Intell.* 1226–38

Petterson M T, Begnoche V L and Graybeal J M 2007 The effect of motion on pulse oximetry and its clinical significance *Anesth. Analg.* **105** S78–84

Poets C F 2010 Apnea of prematurity: What can observational studies tell us about pathophysiology? *Sleep Med.* **11** 701–7

Saeed M, Villarroel M, Reisner A T, Clifford G D, Lehman L W, Moody G, Heldt T, Kyaw T H, Moody G B and Mark R G 2011 Multiparameter intelligent monitoring in intensive care: II. A public-access intensive care unit database *Crit. Care Med.* **39** 952–60

Saeys Y, Inza I and Larrañaga P 2007 A review of feature selection techniques in bioinformatics *Bioinformatics* **23** 2507–17

Sornmo L and Laguna P 2005 *Bioelectrical Signal Processing* (Amsterdam: Elsevier)

Tobin R M, Pologe J A and Batchelder P B 2002 A characterization of motion affecting pulse oximetry in 350 patients *Anesth. Analg.* **94** 54–61

Zong W, Heldt T, Moody G B and Mark R G 2003 An open-source algorithm to detect onset of arterial blood pressure pulses *Proc. Computers in Cardiology* pp 259–262

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