KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Deviaene M., Testelmans D., Buyse B., Borzee P., Van Huffel S., Varon C. (2017), **Automatic screening of sleep apnea patients based on the sp02 signal** IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, Mar. 2019, 607-617..
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	https://ieeexplore.ieee.org/document/8322241/
Journal homepage	https://jbhi.embs.org/
Author contact	Margot.deviaene@esat.kuleuven.be + 32 (0)16 32 89 60
Abstract	Objective: This paper presents a methodology to automatically screen for sleep apnea based on the detection of apnea and hypopnea events in the blood oxygen saturation (SpO 2) signal. Methods: It starts by detecting all desaturations in the SpO 2 signal. From these desaturations, a total of 143 time-domain features are extracted. After feature selection, the six most discriminative features are used to construct classifiers to predict if desaturations are caused by respiratory events. From these, a random forest classifier yielded the best classification performance. The number of desaturations, classified as caused by respiratory events per hour of recording, can then be used as an estimate of the apnea-hypopnea index (AHI), and to predict whether or not a patient suffers from sleep apnea-hypopnea syndrome (SAHS). All classifiers were developed based on a subset of 500 subjects of the Sleep Heart Health Study (SHHS) and tested

	on three different datasets, containing 8052 subjects in total. Results: An averaged desaturation classification accuracy of 82.8% was achieved over the different test sets. Subjects having SAHS with an AHI greater than 15 can be detected with an average accuracy of 87.6%. Conclusion: The achieved SAHS screening outperforms SpO 2 methods from the literature on the SHHS test dataset. Moreover, the robustness of the method was shown when tested on different independent test sets. Significance: These results show that an algorithm based on simple features of SpO 2 desaturations can outperform more elaborate methods in the detection of apneic events and the screening of SAHS patients.
IR	https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS1645437&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US

(article begins on next page)

Automatic Screening of Sleep Apnea Patients Based on the SpO_2 Signal

Margot Deviaene, *Member, IEEE,* Dries Testelmans, Bertien Buyse, Pascal Borzée, Sabine Van Huffel, *Fellow, IEEE,* and Carolina Varon, *Member, IEEE*

Abstract*—*Objective: **This paper presents a methodology to automatically screen for sleep apnea based on the detection of apnea and hypopnea events in the blood oxygen saturation (SpO_2) signal. *Methods:* It starts by detecting all desaturations in the SpO_2 signal. From these desaturations, a total of 143 time-domain features are extracted. After feature selection, the six most discriminative features are used to construct classifiers to predict if desaturations are caused by respiratory events. From these, a random forest classifier yielded the best classification performance. The number of desaturations, classified as caused by respiratory events per hour of recording can then be used as an estimate of the apnea-hypopnea index (AHI), and to predict if a patient suffers from sleep apnea-hypopnea syndrome (SAHS) or not. All classifiers were developed based on a subset of 500 subjects of the Sleep Heart Health Study (SHHS) and tested on three different datasets, containing 8052 subjects in total. *Results:* An averaged desaturation classification accuracy of 82.8 % was achieved over the different test sets. Subjects having SAHS with an AHI larger than 15 can be detected with an average accuracy of 87.6 %. *Conclusion:* The achieved SAHS screening outperforms SpO_2 methods from the literature on the SHHS test dataset. Moreover, the robustness of the method was shown when tested on different independent test sets. *Significance:* These results show that an algorithm based on simple features of SpO_2 desaturations can outperform more elaborate methods in the detection of apneic events and the screening of SAHS patients.**

***Index Terms*—Sleep apnea hypopnea syndrome, event detection, Oxygen saturation, Random forest**

Manuscript received December 22, 2017; revised February 22, 2018 .The work was supported by: Bijzonder Onderzoeksfonds KU Leuven (BOF) Center of Excellence (CoE) #: PFV/10/002 (OPTEC) SPARKLE Sensor-based Platform for the Accurate and Remote monitoring of Kinematics Linked to E-health #: IDO-13-0358 The effect of perinatal stress on the later outcome in preterm babies #: C24/15/036 TARGID - Development of a novel diagnostic medical device to assess gastric motility #: C32-16-00364. Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (FWO) Project #: G.0A5513N (Deep brain stimulation). Agentschap voor Innovatie door Wetenschap en Technologie (IWT) Project #: SWT 150466 - OSA+, O&O HBC 2016 0184 eWatch. imec funds 2017. imec ICON projects ICON HBC.2016.0167, 'SeizeIT'. Belgian Federal Science Policy Office IUAP #P7/19/ (DYSCO, 'Dynamical systems, control and optimization', 2012-2017). Belgian Foreign Affairs-Development Cooperation VLIR UOS programs (2013-2019). EU: European Union's Seventh Framework Programme (FP7/2007-2013) EU MC ITN TRANSACT 2012, #316679, The HIP Trial: #260777. ERASMUS + INGDIVS 2016-1-SE01-KA203-022114. European Research Council The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Advanced Grant: BIOTENSORS (n 339804). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. Carolina Varon is a postdoctoral fellow of the Research Foundation-Flanders (FWO).

M. Deviaene, S. Van Huffel, and C. Varon are with the Department of Electrical Engineering-ESAT, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, and imec, KU Leuven, B-3001 Leuven, Belgium, (e-mail: margot.deviaene@esat.kuleuven.be).

D. Testelmans, B. Buyse and P. Borzée are with the UZ Leuven, Department of Pneumology , Leuven, Belgium .

I. Introduction

SLEEP apnea-hypopnea syndrome (SAHS) is the most common sleep related breathing disorder. This disease causes complete or partial cessations of breathing during sleep, called respectively apneas and hypopneas. These events can be due to a blockage in the upper airway (obstructive) or due to a loss of respiratory drive (central) [1]. Events regularly result in a drop in blood oxygenation and / or an arousal, which causes normal breathing to be restored.

The recurrence of these events during the night causes sleep fragmentation, which results in daytime sleepiness, leading to, among others, a decrease in work performance and an increased risk of motor vehicle accidents. On top of the sleepiness, patients with SAHS have an increased risk of developing cardiovascular comorbidities [2]–[4]. Therefore, timely diagnosis and treatment of SAHS is important.

The gold standard for diagnosing SAHS is a full night polysomnography (PSG) in an attended hospital setting. Sleep stages and respiratory events are, in most countries, manually scored using these PSG recordings. In order to ensure uniformity among scoring of sleep studies, the American Academy of Sleep Medicine (AASM) developed a set of scoring rules [1], [5], [6]. According to the AASM 2012 rules, all respiratory events should last longer than 10 seconds, and when airflow amplitude decreases with more than 90 % the event is scored as an apnea. Hypopnea events, on the other hand, only need a decrease of at least 30 %, accompanied by either an arousal or an oxygen desaturation of at least 3 %. For the remainder of this article, the term apneic events will be used as a global term representing both apneas and hypopneas, unless a clear distinction is made between the two.

The apnea-hypopnea index (AHI) represents the number of annotated events divided by the hours asleep and is used to assess the severity of SAHS. As such, SAHS severity is divided in four categories: normal (AHI$<$5), mild (5$\leq$AHI$<$15), moderate (15$\leq$AHI$<$30) and severe SAHS (AHI$\geq$30) [1].

It is estimated that around 13 % of men and 6 % of women between 30 and 70 years old have an AHI larger than 15 and suffer from moderate to severe SAHS [7]. A lot of these people, however, remain undiagnosed. This is mainly due to the cumbersome and expensive diagnosis using PSG which discourages people to get tested and is not suited for screening of large populations. SAHS screening is therefore an important field of research as sleep studies need to be brought out of the hospital environment into the subject's home to reach a larger population in order to find undiagnosed SAHS

patients and treat them before they develop SAHS related comorbidities. Many researchers have focused on developing alternative screening and diagnostic tools based on signals that can easily be acquired in a home environment [8]. Detection algorithms based on electrocardiogram (ECG) [9], [10], photoplethysmography (PPG) [11], snoring sounds [12], respiration [13], blood oxygen saturation (SpO_2) [14]–[22], or a combination of the above mentioned signals have been proposed [23]–[25].

Pulse oximetry is an ideal measurement technique for this as it is cheap, unobtrusive and easy to set up in a home environment. Lots of wearable pulse oximeter solutions are already available on the market that can be used to acquire the SpO_2 signal during sleep. Moreover, the SpO_2 signal is of particular interest for SAHS event detection because many apneic events are associated with an oxygen desaturation. Therefore, this study will focus on apneic event detection based on the SpO_2 signal. Previous studies based on SpO_2 used time and frequency domain features [16], [19]–[21], nonlinear parameters [14], [19], PSG parameters such as the oxygen desaturation index (ODI) [16], [18], [22], [25] and more elaborate methods, such as sparse representations [15], to detect SAHS. Most of these studies classify patients as having SAHS or not, based on features computed over the full night, without detecting the individual apneic events. For the computation of the AHI according to the definition, however, the detection of the events is crucial. Therefore, the present study focuses on detecting individual events. The detection of these events is not epoch based, as proposed by Koley et al. [21] and Burgos et al. [22], but based on detecting oxygen desaturations and deriving features from these desaturations. Using the most discriminative features, a classifier can then be built, which decides if the desaturation was caused by an apneic event or not. Based on the classification of desaturations, the AHI can then be estimated. We hypothesize that when the most discriminative SpO_2 features from different categories, e.g. desaturation severity and quasi-periodicity, are combined, higher classification performances can be achieved. Another point of novelty in this work is the fact that linear regression of the estimated AHI is used to inflate the AHI in order to account for respiratory events without oxygen desaturations and awake periods. The regressed AHI can then be used for the classification of patients as having SAHS or not.

Publicly available datasets are used in this study, in order to enable comparison with previous studies. The algorithm is also tested on a dataset of the UZ Leuven sleep laboratory to assess the generalization capabilities of the algorithm.

The remainder of this paper starts with the description of the datasets in Section II, followed by the explanation of the complete algorithm in Section III. Afterwards the results of the algorithm will be presented and analysed in Section IV, Section V discusses the algorithm further and compares the obtained results with methods from the literature. In the end, conclusions are presented in Section VI.

II. Materials

Three different datasets are used in this study. Two datasets are publicly available: the Sleep Heart Health Study (SHHS) dataset [26], [27] is made available by the NSRR [28] and the Apnea-ECG dataset [29] can be found on Physionet [30]. The third dataset was recorded at the sleep laboratory of the University Hospitals Leuven (UZ Leuven).

The SHHS is a multi-center cohort study focusing on the cardiovascular and other consequences of sleep-disordered breathing. Recordings from 5793 subjects undergoing unattended full night PSG at baseline are available. All subjects were 40 years or older and had no history of treatment for SAHS. 2651 of these subjects underwent a follow-up PSG, on average 5 years later. All available baseline and follow-up recordings have more than 3 hours of usable SpO_2 signal, and were therefore included in our study. The recordings were manually scored; annotations of sleep stages, arousals, oxygen desaturations and respiratory events are available. This enables the user to apply different hypopnea definitions. For this study, the AHI was computed taking into account all apneas and hypopneas with either an arousal or an oxygen desaturation of at least 3 %, in order to match the AASM 2012 rules [6]. This AHI was computed from the original polysomnography variables included in the dataset. For this study, annotations of individual apneic events are needed as well. Therefore, hypopnea annotations without either an arousal or desaturation of at least 3 % need to be ignored. It is, however, a known issue with the dataset that original oxygen desaturation annotations were indelibly changed due to software conversions over the years. Therefore, these annotations were not used, the oxygen desaturations were automatically recomputed and linked to hypopneas, as described in the methods section. Based on these linked desaturations, it was decided which hypopneas should be taken into account. The SpO_2 signals available in this dataset were recorded using a Nonin XPOD 3011 sensor and sampled at 1 Hz. The SHHS dataset is split in 3 subsets for this study. 500 recordings are selected from the baseline PSGs for training the algorithm, these are called *SHHS1 train*. The remaining baseline recordings will be called *SHHS1 test* and the follow-up recordings *SHHS2*. The training samples are selected as the centroids of k clusters, computed using k-means clustering of the patient characteristics (age, BMI, AHI and smoking status), while ensuring an equal number of males and females, and equal distribution over the four apnea severity classes. As such, the underlying distribution of the patient population in SHHS1 is fully represented in the training set.

The Apnea-ECG *Physionet* database consists of 8 full night recordings that include next to the ECG, the SpO_2 and 3 respiratory signals. The SpO_2 is sampled at 100 Hz, the AHI and minute-by-minute apnea annotations are available, depicting if any apneic event occurred during each 1-minute segment. Half of the subjects have no SAHS, the other half has severe SAHS.

The last dataset is the *UZ Leuven* dataset, which contains 100 PSGs of patients, suffering from moderate to severe SAHS (AHI$\geq$15). The dataset was scored by sleep specialists according to the AASM 2012 scoring rules [6]. The SpO_2 signals were recorded using a Nonin 8000J sensor and are stored at a sampling rate of 500 Hz.

An overview of the different datasets, and the patient

TABLE I
PATIENT CHARACTERISTICS FOR THE DIFFERENT DATASETS

Dataset	# Sub.	Age Years	BMI Kg/m^2	AHI Events/h	AHI $\geq$15	Male
SHHS1 train	500	63$\pm$11 (55, 74)	28$\pm$5 (25, 31)	20$\pm$20 (5, 30)	50%	50%
SHHS1 test	5293	63$\pm$11 (55, 72)	28$\pm$5 (25, 31)	18$\pm$16 (7, 23)	44%	53%
SHHS2	2651	67$\pm$10 (60, 76)	28$\pm$5 (25, 31)	18$\pm$16 (7, 25)	45%	54%
Physionet	8	43$\pm$8 (38, 52)	28$\pm$8 (22, 35)	32$\pm$36 (0, 70)	50%	88%
UZ Leuven	100	48$\pm$11 (39, 56)	30$\pm$4 (27, 33)	41$\pm$22 (24, 57)	100%	78%

Age, BMI and AHI are presented as mean $\pm$ standard deviation, with the 25 % and 75% quantile values underneath.

characteristics of each dataset are given in Table I.

III. METHODS

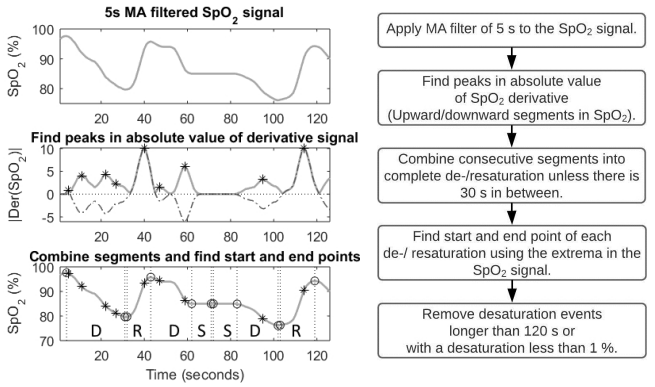
The algorithm starts by preprocessing the SpO_2 signals . Afterwards, desaturations are detected. Class labels are then computed for these desaturations, based on the presence of a preceding apnea or hypopnea that caused the desaturation. For each desaturation, 143 features are extracted, feature selection techniques are subsequently used to select the most discriminative features. Based on these features, different classifiers are then trained to detect apneic events. Based on the number of detected events, an estimation of the AHI can be made in the end. The discussed steps are explained in detail in the following sections.

A. SpO_2 preprocessing

Before analysing the signal, zero-level artifacts due to sensor disconnections are deleted by detecting desaturations in the SpO_2 that drop below 50 % and replacing them by linear interpolations. Afterwards a moving average (MA) filter with a 3 seconds duration [21] is used to get rid of sharp changes and ripples due to oversampling in the UZ Leuven dataset. This high sampling rate also causes the need to down sample the signals to 1 Hz, in order to equalize the sampling rates and speed up computations [21].

B. Desaturation event and baseline extraction

When an apneic event occurs, a drop in the blood oxygen concentration, or so-called desaturation (D), is often observed. Afterwards, possibly a stable period (S) of low SpO_2 occurs and then resaturation (R) normally brings the SpO_2 back to its baseline value, as can be seen in the lower graph of Fig. 1a. The term *desaturation event* will be used to depict the complete event starting from the start of the desaturation until the end of the resaturation, as depicted in Fig. 2. The goal of this study is to detect apneic events based on their desaturation event. The first step is thus to detect all desaturations in the SpO_2 signal, the different steps in this process are depicted in Fig. 1. In order to get a smoother signal and make the detection of desaturations easier, the first step is to apply an extra MA filter of 5 seconds to the preprocessed signal. This extra



(a) Graphical overview (b) Overview of steps

Fig. 1. Overview of the extraction of desaturations in the SpO_2 signal. (a) On top the MA filtered SpO_2 signal is plotted. In the middle, the derivative of this signal is shown, as well as its absolute value (solid line) , on which peaks are detected (*). On the bottom, the start and end points of the desaturations and resaturations corresponding to these peaks are detected and linked together (o). As such desaturation events are extracted, containing a desaturation (D), a possible stable period (S), and a resaturation (R).

MA filter was solely used for the detection of desaturations, as the extra smoothing would affect the calculation of SpO_2 features. Next, derivative filtering is performed on the resulting signal. Peaks in the absolute value of the derivative are detected, representing upward and downward parts in the SpO_2 signal [23] (Fig. 1a, middle). Consecutive downward segments are linked in order to create the complete desaturation, unless the detected peaks are more than 30 seconds apart [23]. In that case, it is unlikely that these desaturations are caused by the same physiological event, it is more probable that two independent desaturations occur without any resaturation in between [23]. Therefore these desaturations are seen as two separate events. Upward segments are linked in the same way to create complete resaturations. An example of an interrupted desaturation with a plateau in between can be seen in Fig. 1a, around the 60 s time point, 2 separate desaturation events are thus detected. The first event will have no resaturation, whereas the desaturation part of the second event starts below the baseline SpO_2 value. The ability to detect nearby desaturation events with incomplete desaturation or resaturation, is the main advantage of using the derivative of the SpO_2 signal instead of just taking the extrema in the signal. This is very important, since these desaturation events can now be detected as coming from two separate apneic events. Finally, the start and end points of these desaturations and resaturations are extracted, using the extrema in the SpO_2 signal. Desaturation events are retained if they have a drop in SpO_2 of at least 1 % and a total length, from the start of the desaturation to the end of the resaturation, shorter than 120 seconds, since this is typically the maximal duration of apneic events and their corresponding desaturations [21], [31].

Afterwards, a running SpO_2 baseline is computed as the 95$^{\text{th}}$ percentile value over the previous one minute recording that did not contain any detected artifacts or desaturations. A maximum baseline alteration of 1 % of oxygen saturation per second was allowed. When the computed baseline update had a change larger than 1 % with respect to the previous baseline

TABLE II
OVERVIEW OF THE EXTRACTED FEATURES

#	Extracted features	Source
16	Amplitude (differences), lengths and slopes	
1	Area below baseline * total length	[31]
18	Area and length below baseline, baseline - 2/3/4 %, start SpO_2 level	[32]
50	Min, max, mean, median and variance of SpO_2 and its 1st, 2nd and 3rd order derivative	[21]
16	Deviation from min and max for median and mean of SpO_2 and its 1st, 2nd and 3rd order derivative	[21]
40	PRSA features, upward and downward slopes with fragment lengths of 10 and 20 s	[33]
2	1st peak of the autocorrelation and its relative amplitude	
4	Age, BMI, smoking status, gender	

value, the baseline was updated with only 1 % in the direction of the computed baseline update value."

C. Compute class labels for the desaturation events

Sleep apnea scoring of PSG recordings is performed based on the respiratory signals; the start annotation of events represents the beginning of the flow limitation and the end annotation is placed when normal breathing is restored [6]. The start of the SpO_2 desaturation, however, has a delay of about 20-40 seconds with respect to the start of the apneic event [23]. For each annotated apneic event, therefore, a desaturation event will be sought within a window of 60 seconds after the start of the annotation. The apneic event is linked to the closest desaturation in this window, which is not yet linked to a previous apneic event. Sometimes apneic events cannot be linked to a desaturation, because there is no drop in SpO_2. In this case, this methodology will be unable to detect these apneic events. All remaining unlinked desaturations, will be labeled as non-apneic for training and testing, these might for example be due to short breathing disturbances which do not meet the AASM 2012 criteria to be scored as apneic [6].

As such, all desaturation events from the SHHS and UZ Leuven datasets are labeled as being linked to an apneic event or not. The Physionet dataset, however, only has 1-minute annotations. Therefore desaturation events cannot be labeled individually, and thus this dataset can only be used for testing with 1-minute based performance evaluation.

D. Feature extraction

A total of 143 features were extracted from each detected desaturation event, and an overview of these features is given in Table II. The features can be divided in simple time domain, desaturation severity, statistical and quasi-periodicity features.

1) Simple time-domain: A first group of parameters are simple time-domain features; amplitudes, amplitude differences, lengths and slopes are extracted from the desaturation and resaturation parts.

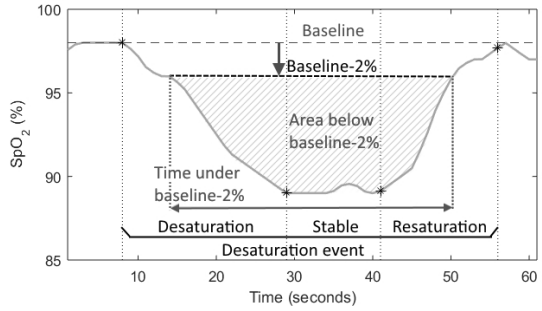


Fig. 2. Extraction of desaturation severity features. The computation of area and time below 2 % under baseline are shown. Start and end points of the desaturation and resaturation are depicted with *.

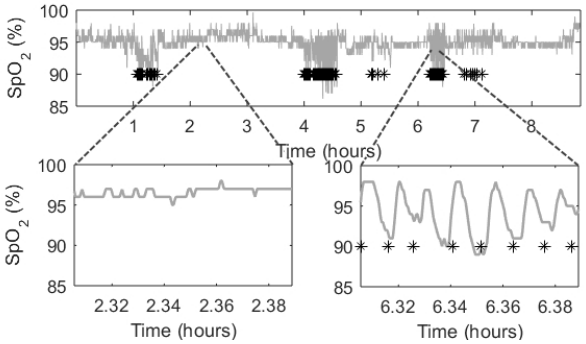


Fig. 3. Periodicity of the SpO_2 signal. Grouping of apneic events (*) during the night can be seen for a subject with an AHI of 19.5 on top. This leads to close succession of apneic events and periodicities in SpO_2 segments with apneas, as can be seen in the bottom right 5-minute segment. This periodicity is not present in unobstructed sleep as shown on the bottom left.

2) Desaturation severity: Next, desaturation severity parameters are computed, such as the obstruction severity, defined by Kulkas et al. as the area below baseline multiplied by the total length of the desaturation event [31]. In addition, the area and time below baseline or 2, 3, or 4 % under baseline during the event as introduced by Watanabe et al. and explained in Fig. 2 were extracted [32]. This area and time were also computed with respect to the starting SpO_2 level of the desaturation event.

3) Statistical: The third group of features are the minimum, maximum, mean, median and variance of the SpO_2 signal and its first, second and third order derivative during the complete desaturation event, the desaturation part and the resaturation part. Additionally the deviation from the means and medians by the minima and maxima are computed for the complete events, as proposed by Koley et al. [21].

4) Quasi-periodicity: The last group of features captures the quasi-periodicities in the SpO_2 signal in a five-minute segment around the start of the desaturation events. Considering that apneic events rarely occur alone, often there are periods where apneic events follow in close succession. This can be seen in the recording of a subject with an AHI of 19.5 on top in Fig. 3, all apneic events are grouped within certain periods during the night. As such, the desaturations occur one after the other, leading to periodicity in the SpO_2 signal, as can be seen on the bottom right of Fig. 3. This periodicity

will be characterized using two methods: Phase Rectified Signal Averaging (PRSA) [33] and the autocorrelation (AC) of the SpO_2 signal. The PRSA method searches for anchor points, which represent all time points where the signal goes downward (or upward) in the five-minute segments. Fragments of duration d are extracted around each anchor point, all these fragments are aligned and averaged. As such, an average curve of length d is computed that represents the average desaturation (or resaturation) of the five-minute segment. From this curve, the capacity [34], the amplitude difference, overall slope, slope before and after the anchor point can be extracted as features. For the fragment duration d, 10 and 20 seconds are considered. Another way to detect the periodicity in the signal is by means of the autocorrelation (AC). The relative amplitude and position of the first peak in the AC on the five-minute segments are extracted. The position of this peak represents the delay in between successive apneas, the relative amplitude gives a notion if a periodicity is indeed present.

From all 143 features, the logarithmic transformation is computed, which will also be included in the feature selection step. In addition, patient characteristics as age, BMI, smoking status and gender are considered. In total, 294 features are thus used as input for the feature selection described in the next section, from which 286 features are extracted from each SpO_2 signal and 8 from patient characteristics.

E. Feature selection and classification

In order to create a good classifier, the most discriminative feature set to separate desaturation events caused by an apneic event from those that are not, should be selected. An overview of the methodology used for the feature selection is given in Fig. 4, and the different steps will be explained below. Both the feature selection and classifier training are based solely on the training set SHHS1 train, whereas the other datasets are used as independent test sets.

The first feature selection step is to delete highly correlated features to avoid multicollinearity in the classification system. Feature pairs with a correlation larger than 0.75 are investigated, in decreasing order of correlation. The feature with the lowest F-test score is deleted, the other one is retained. The F-score computes the ratio between intra-group variability and inter-group variability. In order to avoid overfitting of the feature selection to one patient, this F-score is computed using the leave-one-patient-out (LOPO) method, and the overall F-score is then taken as the median of the LOPO F-scores. This resulted in a total of 68 retained features , with a correlation smaller than 0.75. This is done in order to reduce the collinearity between the features. The maximal correlation value of 0.75 was obtained as a trade-off between the number of retained features and the computational cost.

The next step was to apply the minimal relevance maximal redundancy algorithm (mRMR) [35]. This forward selection algorithm, based on mutual information between the classes and feature sets was used to select the optimal feature set of 10 features. This number of 10 features is a trade-off between having all of the most informative features and reducing the computational cost of the backwards wrapper in the next

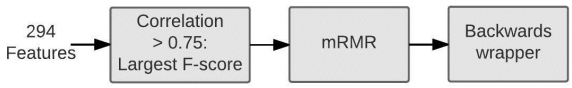


Fig. 4. Overview of the different steps in the feature selection methodology.

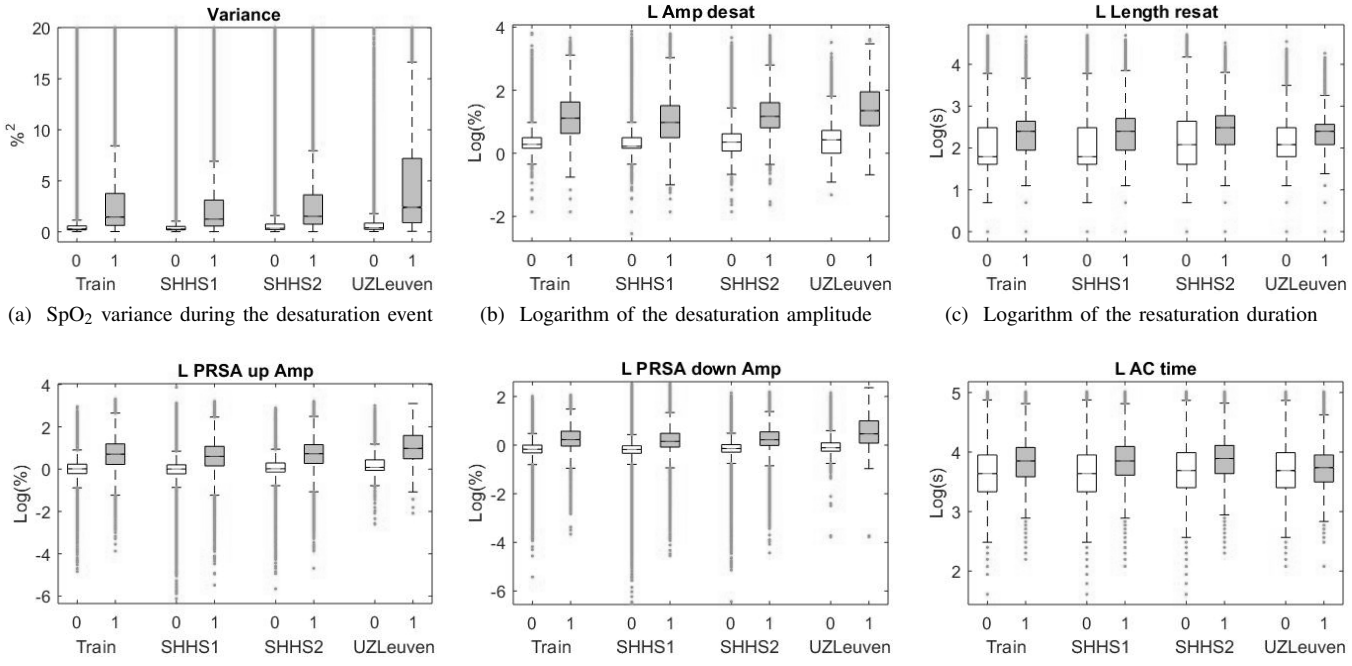
stage, which is needed to remove any redundant features that might be left in the mRMR feature subset, since the mRMR algorithm does not provide an optimal number of features needed [35]. This backwards wrapper was applied using RBF least-squares support vector machines (LS-SVM), using the `LS-SVMlab1.8` Toolbox [36] and 10-fold cross-validation (CV) for parameter tuning. Since the subjects have on average around 400 desaturations per night, and there are 500 patients in our dataset, the total training set consists of around 200,000 data points. This is too large to be used for LS-SVM classification, as the kernel matrix of n x n should be computed with n the number of training data points. Therefore a fixed-size LS-SVM classifier was trained on the 2000 most representative training data points, selected using the k-means algorithm on the features and class labels [37].

Using the selected feature set, classifiers were trained to decide if a desaturation was caused by an apneic event or not. The predictive value on the test sets was used to compare the different classifiers. Fixed-size LS-SVM classifiers using linear and RBF kernels , as explained above, were compared against k-nearest-neighbours (kNN) [38], linear discriminant analysis (LDA) [39] and random forest [40] . The kNN classification was implemented using the Matlab function `fitcknn` and the optimal k is obtained with 10-fold CV. LDA was implemented using the `fitcdiscr` function, whereas the random forest classifier was implemented using `TreeBagger` with 100 bagged trees. All classifiers, except for the LS-SVMs, are computed both on the full training dataset and the reduced set of 2000 desaturations, in order to compare the obtained results with the LS-SVM classifiers.

After training, these classifiers are tested on the SHHS1 test, SHHS2, Physionet and UZ Leuven datasets. The performance is computed by comparing the class labels of the desaturation events with the predicted labels from the classifiers. For the Physionet dataset, however, performance is computed by comparing 1-minute classification labels with the 1-minute annotations, as no event annotations are available. A 1-minute segment is classified apneic if a desaturation classified as caused by an apneic event starts within the start of the 1-minute and 15 seconds after the end of the 1-minute interval.

F. Statistical analysis of classification performance

Classification performance will be analysed using the Accuracy (Acc), Sensitivity (Se), Specificity (Sp), Positive predictive value (PPV), Area under the ROC curve (AUC) and Cohen's Kappa value (κ). The performance parameters are computed for the point on the ROC curve that maximizes the product of Se and Sp.



(a) SpO_2 variance during the desaturation event (b) Logarithm of the desaturation amplitude (c) Logarithm of the resaturation duration

(d) Logarithm of the upwards PRSA amplitude (e) Logarithm of the downwards PRSA amplitude (f) Logarithm of the AC first peak time

Fig. 5. Selected features for desaturations linked to an apneic event (1) and not (0), for the SHHS and UZ Leuven datasets.

G. AHI estimation

Based on the number of desaturations classified as caused by an apneic event, an estimation can be made of the AHI for each patient. This is done by dividing the number of apneic classified events by the total recording time in hours. The annotated AHI, however, uses sleeping time instead of total recording time. This means that our computed AHI will be an underestimation of the annotated AHI. In addition, not all apneic events will result in a desaturation, and thus not all apneic events can be detected. In order to counteract these limitations, a robust linear regression is performed [41], using the `robustfit` Matlab function, between the annotated and computed AHI of the training set. This regression will then also be applied to inflate the computed AHI of the test sets. Based on these regressed AHIs, classification of patients in SAHS severity groups can be performed.

IV. Results

The results of the apneic event detection algorithm on the training and test sets are discussed in this section. The selected features will be discussed and the performance of the different classifiers will be compared. The estimate of the AHI values based on the best performing classifier will be presented, as well as the SAHS severity group classification.

A. Feature selection

The previously discussed feature selection techniques, applied on the training set SHHS1 train, resulted in the selection of six features. Three time domain features were selected: the amplitude of the desaturation, the variance of the SpO_2 signal and the duration of the resaturation. The three other features are quasi-periodicity features: the AC time and the amplitude difference for upward and downward PRSA of 20 second fragments. All features, except for the SpO_2 variance, were selected in their logarithmically transformed form.

Boxplots these features can be found in Fig. 5 for desaturations produced by an apneic event (1) or not (0), and for each dataset separately. A statistically significant difference was found between the apneic and non-apneic desaturation groups for all features and all datasets using the Wilcoxon rank sum test. Due to the large sample size, each group contains at least 25,000 samples, these differences might, however, represent a small shift in median values. When the boxplots of Fig. 5 are studied, one can nevertheless see a clear difference in median values for all cases, except for the AC time within the UZ Leuven dataset. Moreover, the inter quartile ranges have comparable values for the same group over the different datasets, and do not overlap between the apneic and non-apneic desaturations for four features (Fig. 5a, 5b,5d and 5e). No data is available for the Physionet dataset, as no ground truth annotation of desaturations is available. The general trend is that desaturations produced by an apneic event have larger amplitude differences of desaturation and in PRSA curves, a longer resaturation time, more variance in the SpO_2 signal during the desaturation and a larger time until the first AC peak.

B. Classification performance

Performance measures for classifying desaturation events caused by an apneic event or not are given in Table III for the different classifiers, averaged over all test datasets, using the number of subjects per test set as a weighting factor. The suffix *(2000)* represents the fact that these classifiers were

TABLE III
OVERVIEW OF THE AVERAGED CLASSIFICATION PERFORMANCE ON THE TEST DATASETS.

Classifier	Acc	Se	Sp	PPV	AUC	κ
Linear LS-SVM	79.0	75.8	79.9	54.6	85.3	49.1
RBF LS-SVM	78.7	76.6	79.3	54.1	85.4	48.9
k-NN	79.7	76.1	80.8	55.8	**86.0**	50.5
k-NN (2000)	79.5	76.1	80.4	55.4	85.7	50.0
LDA	78.7	**76.7**	79.3	54.1	85.3	48.9
LDA (2000)	79.3	75.2	80.5	55.1	85.3	49.5
Random Forest	**82.8**	64.3	**88.6**	**64.2**	85.4	**52.7**
Random Forest (2000)	77.8	74.4	78.7	52.7	83.7	46.5

TABLE IV
OVERVIEW OF THE AVERAGED CLASSIFICATION PERFORMANCE OF THE RANDOM FOREST CLASSIFIER ON EACH DATASET.

Dataset	Acc	Se	Sp	PPV	AUC	κ
SHHS1 train	82.8	64.6	89.3	68.4	85.7	54.9
SHHS1 test	83.3	60.8	90.2	65.4	85.0	52.2
SHHS2	82.0	70.7	85.7	61.1	86.1	53.5
UZ Leuven	78.1	75.5	80.8	79.5	85.4	56.3
Physionet	91.3	98.9	86.6	82.2	95.5	82.4

trained only on the subsample of 2000 training desaturations and not on the complete training set. As can be seen from Table III, the accuracy of the random forest classifier on the complete training dataset outperforms all the others. Its accuracy is at least 3 % higher. The AUC is similar for all datasets, except Random Forest (2000). Therefore, only the random forest classifier is used in the remainder of this work. When using the random forest classifier, a trade-off had to be made between the number of trees and the number of training samples used, due to computational limitations. Simulations have shown that it was better to use the full training set with a limited number of 100 trees.

Table IV shows separately the averaged performance parameters for each dataset. Mean out-of-bag performance is given for the training set, which yields an accuracy of 82.8 %.

C. Estimation of the AHI

The AHI is computed as the number of desaturations caused by an apneic event divided by the total recording time. In Fig. 6, on top, the computed AHI is plotted against the annotated AHI for each dataset. As expected, the computed AHI underestimates the annotated AHI due to the fact that subjects are on average only sleeping 75.9 % of the recording and 11.5 % of the apneic events is on average not accompanied by a desaturation. The linear regression between the two AHIs was computed for each dataset separately, resulting in good fits giving high R^2 coefficients of determination between 0.86 and 0.95. In order to ensure the independence of the test set, robust linear regression was computed on the training set and then applied to the test sets. This regression estimation of the AHI will be used in the remainder of this study. Fig. 6 contains on the bottom, the Bland-Altman plots between the estimated and the annotated AHI for each of the datasets. These plots show an overestimation of the AHI for high AHI, most clearly in the UZ Leuven and Physionet datasets. But to a lesser extent also in the SHHS test sets. The 95 % limits of agreement only slightly increase from the SHHS training set to the SHHS test sets, a limit of 10.9 is found for the training set and 14.8 and 14.7 respectively for the SHHS1 and SHHS2 test sets. The median absolute errors for the AHI estimate are for these three datasets, respectively, 1.19, 3.14 and 2.94. For the other two datasets, however, the limits of agreement double due to the overestimation of large AHIs.

D. SAHS classification

Using the estimated AHI, the subjects can be classified as having no (AHI$<$5), mild (5$\leq$AHI$<$15), moderate (15$\leq$AHI$<$30) or severe (AHI$\geq$30) SAHS. Fig. 7 represents the confusion matrix of the estimated and true class labels for the SHHS datasets. Excellent classification performance is achieved on the training dataset, 83.2 % of the subjects is classified in the correct class, the other 16.8 % are classified in a neighbouring class. Accuracies of 67.0 % and 71.9 % are achieved on, respectively, SHHS1 test and SHHS2, which is still quite good but more subjects with mild and moderate SAHS tend to get an underestimation of their SAHS severity. For SAHS screening, an AHI threshold of 15 is often applied. To optimize this classification, the optimal point in the ROC curve was sought for classification of subjects with an AHI larger than 15 on the training dataset, and a threshold for the estimated AHI of 14.52 was obtained. This is slightly lower than 15 and will counteract the underestimation of SAHS severity in the test sets. Classification of SAHS patients based on this threshold yielded an accuracy of 96 % on the training set and 87.5 % and 88.1 %, respectively, on the SHHS1 and SHHS2 test sets. SAHS classification of the Physionet test set yielded 100 % accuracy, as there is a large gap between the AHIs of the two groups. SAHS screening of the UZ Leuven dataset is not useful as all subjects have an AHI larger than 15.

V. DISCUSSION

A. Apnea event detection

This study proposes a method for apneic event detection based on desaturations in the SpO_2 signal which can be used to estimate the AHI and to screen for SAHS. The oxygen desaturations are extracted by detecting upward and downward segments in the SpO_2 signal using its derivative. As such, all SpO_2 changes larger than 1 % should be detected, this was visually confirmed in a random subset of subjects. No annotations of desaturations were available. Hence, it is not possible to prove that all events of interest were detected for every subject. This can be seen as a limitation of the study.

From these desaturations, features were extracted and the most discriminating features were then selected. This feature selection included a backwards wrapper using the RBF LS-SVM classifier. This might have introduced a bias towards this classifier. However, only the final feature selection step included this classifier, which started from 10 features, selected independently from the classifier using the mRMR algorithm.

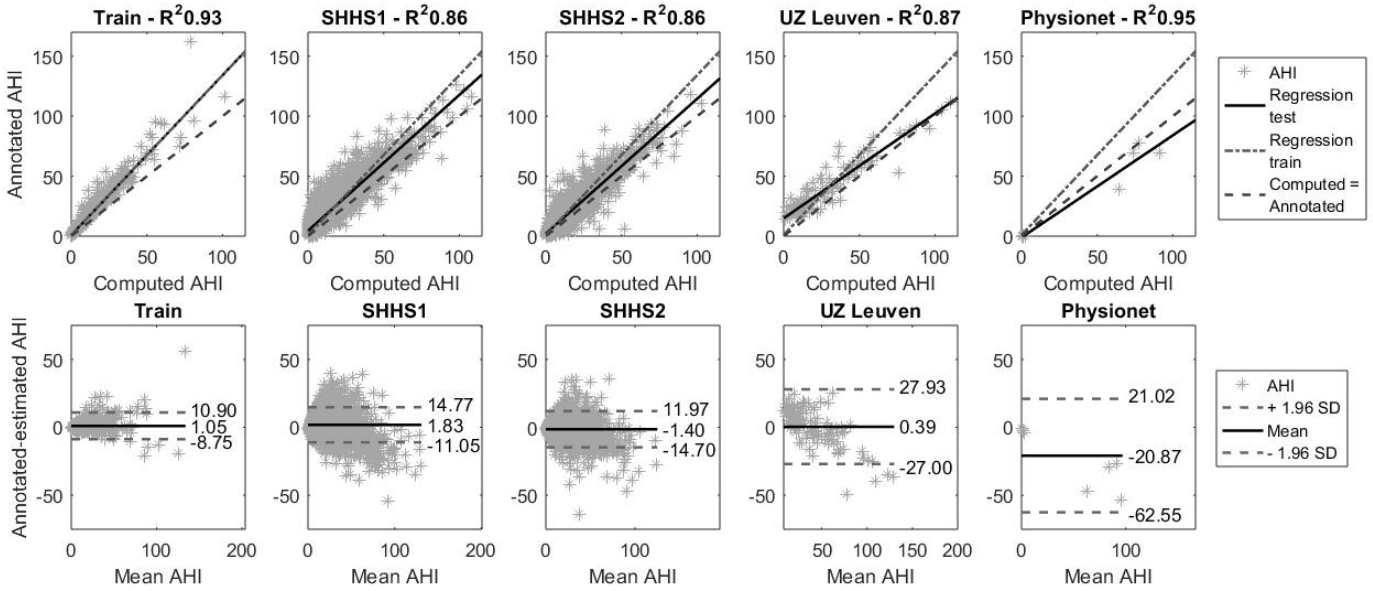


Fig. 6. Top: scatter plot of the computed and annotated AHI for all data sets, R^2 values are given. Regressions based on the training and test data are also plotted. Bottom: Bland-Altman plot between the estimated and the annotated AHI.

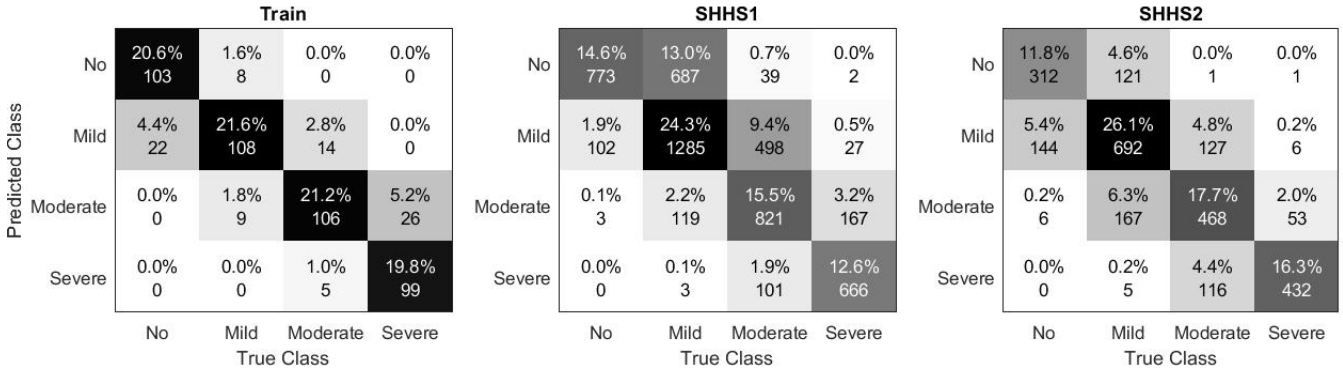


Fig. 7. Confusion matrix for SAHS severity classification based on the estimated AHI of the SHHS datasets.

Moreover, when comparing the test set performances of the different classifiers in Table III, no bias towards this classifier was observed. Using only 2000 training samples the k-NN and LDA classifiers perform comparable. The random forest classifier, however, performs worse with 2000 training samples, but when the extra training samples were added this classifier outperformed the others.

The random forest classifier achieved an averaged accuracy of 82.8 %, which is lower than the performance achieved by Koley et al. [22], but that study was performed on other datasets and the method to annotate desaturations is not clearly given. Therefore, it is difficult to compare the results of both methods. The accuracy of 82.8 % is a good performance taking into account the fact that the algorithm was tested on a large population from different datasets. Moreover, the gold standard manual scoring of respiratory events suffers from intra and inter scorer variability, which has a negative impact on the degree of agreement [42].

When comparing the performances on the different datasets in Table IV, the Physionet dataset stands out with the highest performance, namely an accuracy of 91.3 %. This can partly be explained by the 1-minute based evaluation. If a 1-minute segment contains 3 apneic events, and only 1 is detected by the algorithm, the segment will still be classified correctly. In contrast to the other datasets, where performance is evaluated for each desaturation separately and the previous scenario would only achieve 33 % sensitivity.

Due to the 1-minute annotations instead of individual apneic event annotations, this physionet data might be seen as not perfectly fitting this study. The Apnea-ECG dataset is, however, a publicly available and heavily studied dataset, which makes it possible to compare our results with methods from the literature on the assessment of SAHS severity. This comparison showed that our results were similar to those from the literature. An accuracy of 91.3 % was achieved in this study, whereas the SpO_2 based method proposed by Burgos et al. [22] yielded a performance of 93.0 % on the same dataset. Despite the fact that the classifier in our study was not specifically trained on the Physionet dataset or on 1-minute segment classification as is the case in the study of Burgos

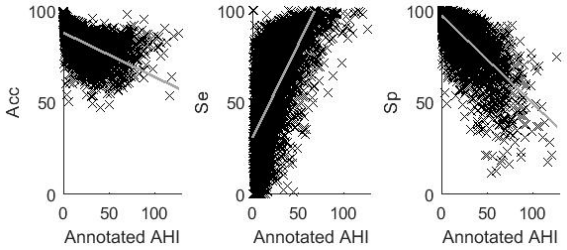


Fig. 8. Accuracy (Acc), Sensitivity (Se) and Specificity (Sp) plotted for each patient of the test sets in function of the patients AHI with the linear regression line plotted on top. Accuracy and Specificity decrease with increasing AHI, whereas Sensitivity increases.

TABLE V
NUMBER OF DETECTED DESATURATIONS PER HOUR OF RECORDING AND THE PERCENTAGE OF THESE DESATURATIONS THAT IS LABELED APNEIC FOR EACH DATASET AND SAHS SEVERITY CATEGORY (C=CONTROL, MI=MILD, MO=MODERATE, S=SEVERE SAHS).

	Desaturations/hour				% apneic			
Dataset	C	Mi	Mo	S	C	Mi	Mo	S
SHHS1 train	35	45	56	71	6	15	27	45
SHHS1 test	35	46	57	72	6	15	26	44
SHHS2	35	46	54	70	6	15	27	43
UZ Leuven	-	-	50	74	-	-	34	57
Physionet	42.1	-	-	88	-	-	-	-

et al. [22]. SpO_2 based algorithms seem to outperform ECG based algorithms on this dataset, which achieve accuracies between 85 % and 90 % on the full Apnea-ECG database [9], [10].

The SHHS test sets have good performances, comparable to the one achieved in the training set. An accuracy of 83.3 % and 82.0 % is achieved for those datasets, however, the sensitivity is rather low. Fig. 8 shows that this low sensitivity is mainly due to subjects with a low AHI, the Acc, Se and Sp are plotted for each test patient in function of the AHI. A decreasing trend can be found for the accuracy and the specificity if the AHI increases, whereas the sensitivity increases. In the SHHS 1 and 2 test data sets, only 17 % of the patients have severe SAHS (AHI$\geq$30), most patients thus have a lower sensitivity and a higher specificity.

Moreover, when looking at the number of desaturations per hour of recording and the percentage of these desaturations that are labeled apneic, which are given in Table V, it can be seen that subjects with larger AHI have more desaturations per hour and a larger percentage of these desaturations are apneic. To avoid a bias in the classifier, it thus might be a good idea to resample the training set to obtain the same number of apneic and non-apneic desaturations per subject. This was not done and is a limitation of the study.

Another important point is the fact that approximately 73 % of the events in the SHHS test sets are hypopneas, which are harder to detect than apneas. If only the apneas would be considered, the sensitivities would increase to, respectively, 77.8 % and 83.9 %. The remaining difference in performance between the SHHS1 and SHHS2 test sets might be caused by the age difference and the further progression of the SAHS in between the two polysomnographies.

The accuracy of the UZ Leuven dataset is 78.1 % which is slightly worse than the SHHS datasets. This is probably due to the fact that SHHS scoring of respiratory events is slightly different from the AASM 2012 [6] rules used in the UZ Leuven dataset. In the SHHS scoring, apneas are already scored when a reduction of respiration of 75 % occurs instead of the standard 90 % nowadays. The study population in the UZ Leuven dataset also differs, whereas the other studies are investigating the general population, the UZ Leuven recordings are conducted on a patient population with identified SAHS symptoms. This difference probably also explains the higher percentage of apneic desaturations for UZ Leuven subjects in Table V. Moreover, all subjects had an AHI larger than 15, so moderate to severe SAHS. When taking into account the tendency seen in Fig. 8, one can conclude that in these categories the accuracy and specificity will be lower but a higher sensitivity is seen. This dataset also contains a high number of hypopneas (66 %), hence, if only the apneas are taken into account, the sensitivity for the UZ Leuven dataset raises to 92.6 %.

These results are obtained without using any form of sleep staging, some of the detected desaturations, however, occur while the patient is awake. When desaturations during awake are deleted from the analysis, the performance only slightly increases, by 0.5 % on average. As the added value of adding sleep staging is minimal, this was not further investigated.

B. AHI computation and SAHS screening

From the detected apneic events, the AHI was computed using robust regression on the training dataset. As can be seen from Fig. 6, this regression on the training set has for all test sets, a steeper slope than the regression on the training set. For the SHHS datasets the deviation between the two regressions is limited, but more prone for higher AHIs. This could be due to the under representation of severe SAHS patients in the test sets. For the UZ Leuven dataset, the number of apneic events not associated with a desaturation is much lower than the average, only 5.7 %. This causes the regression on this dataset to be much closer to the identity line, therefore the inflation of the AHI due to the training regression causes a large overestimation of the AHI, mostly in severe SAHS patients. This difference in the number of apneic events associated with desaturations might be due to the fact that this dataset was recorded on a patient population with SAHS related complaints. It might be possible that these subjects are more prone to more severe apneic events associated with desaturations than subjects with a high AHI without any symptoms. Similar results are seen on the Physionet dataset, although no information is available on the individual apneic events, only 5.3 % of the apneic minutes do not contain any desaturation.

This AHI regression difference between the separate datasets indicates that it might be useful to create different AHI regression estimates dependent on the dataset at hand. This could for example be done by retraining the regression

TABLE VI
COMPARISON OF SAHS PATIENT CLASSIFICATION PERFORMANCE FOR AHI >15 ON SHHS2.

Method	Acc	Se	Sp	AUC
Our Study	**88.08**	**90.12**	86.39	**0.953**
Morales et al. [14]	85.28	84.75	85.81	0.936
Rolón et al. [15]	85.78	85.65	85.92	0.937
Schlotthauer et al. [16]	85.02	84.11	85.94	0.922
Vázquez et al. [17]	84.17	80.84	**87.50**	0.909
Chiner et al. [18]	77.15	78.12	76.17	0.795

based on a subset of the population at hand, or in case of a clinical trial where no ground truth is available the regression can be based on a similar dataset. For example, if the focus of the clinical trial is on a population with SAHS symptoms, AHI regression based on the clinical UZ Leuven dataset should be preferred compared to the SHHS, which is a general population dataset. Another option is to try to identify patient markers that could be responsible for altering this regression, which should be further investigated.

From the computed AHI, SAHS severity classification of subjects was conducted, which achieved good results on the SHHS datasets. The results of the SAHS subject classification with an AHI threshold of 15 on the SHHS2 dataset are compared with methods from the literature [15]–[18], as computed by Rolón et al. on a subset of 995 subjects of the SHHS2 dataset in Table VI. Additionally, the method by Morales et al. [14] was applied on the SHHS2 dataset and the results are included in the table as well. As observed, our methodology outperforms all other methods on this dataset. The accuracy improves by more than 2 % to 88.08 % and AUC also improves to 0.953. These results are even more remarkable since no data of the SHHS2 dataset was used for training of this classifier. Only 500 subjects of the similarly recorded SHHS1 dataset were used for training. But, 223 SHHS1 recordings of SHHS2 subjects are included in the training set. This means that 8.4 % of the subjects in this dataset are not independent of the training set. The other 91.6 % of the subjects are, however, independent of the training set. It was investigated if the inclusion of these previous recordings of the same subject improved the results. This, however, did not show any positive bias towards subjects included in the training set, these subjects only had a severity classification accuracy of 85.2 % in SHHS2 , with a Se, Sp and AUC of respectively 84.7 %, 85.6 % and 95.0 % .

Since most of the used features for classification of desaturations have to do with desaturation severity, the added value of this method compared to the conventional ODI 3 score was investigated. The ODI 3 parameter represents the number of oxygen desaturations of at least 3 % with respect to baseline per hour of sleep. Unfortunately this ODI 3 parameter is not available for the used datasets. As an alternative, however, the desaturation classification was repeated using only the percentage of desaturation with respect to baseline as a feature. A cut-off value of 3 % was used to mimic the ODI 3 parameter. This resulted in an averaged classification accuracy of 79.5 % on the test sets. Our method has a 3.3 % increase in performance compared to this conventional parameter. Moreover, the ODI 3 method has a very low averaged sensitivity of 41.6 %. Our method has an increase in sensitivity of more than 20 %, thus extra apneic events are detected, without compromising the specificity of the classifier. This also results in a lower SAHS severity classification. This is to say, when only the desaturation from baseline is considered, accuracies of 83.5 and 81.9 are obtained, respectively on the SHHS1 and SHHS2 test sets. The computations are repeated for cut-off values of 2 and 4 %, but the results did not improve. Moreover, the use of the conventional ODI parameters has as drawback that no standard definitions of desaturation and baseline are available.

The results of the proposed processing methods show that SAHS screening based on SpO_2 signals can achieve high performances. Moreover, the methods are fully automated, can handle noisy signal segments, do not need any sleep scoring and use only simple features extracted from the desaturations, therefore, they could easily be implemented in a wearable device for screening. Since the screening is, however, only based on the SpO_2 signal, sleep staging and detection of apneic events without desaturations will be lost compared to the full PSG.

C. Future developments

The presented results could still be improved if extra features are considered, for example non-linear or frequency features [14], [19]. This, however, probably will lead to an increase in computational cost, which is a drawback for wearable systems. Moreover, the considered data segments are too short for the computation of nonlinear parameters.

A multimodal apneic event detection algorithm could also be created by for example adding information from the PPG signal, which is already available from the pulse oximeter.

Further investigation into the proposed method is needed to see if a differentiation can be made between the different types of apneic events. Features should be sought which can differentiate central from obstructive events and hypopneas from apneas, as this gives clinicians crucial information on how to treat the SAHS.

VI. CONCLUSION

A methodology for the automatic detection of respiratory events using SpO_2 signals based on the detection and classification of desaturations caused by apneic events was developed. A random forest classifier based on six desaturation severity and SpO_2 periodicity features achieved the best performance for classifying desaturations. An averaged accuracy of 82.8 % was achieved over different independent test sets. Robust linear regression was used to estimate the AHI from the number of desaturations classified as apneic. It was shown that the optimal regression differed between the datasets, to estimate the AHI, the dataset characteristics should be taken into account. The estimated AHI was used for SAHS severity classification and SAHS screening with an AHI threshold of 15. This screening achieved an accuracy of 88 % on the SHHS2 dataset, outperforming all SpO_2 based methods from

the literature tested on the SHHS2 dataset. These results show that this computationally cheap methodology can be very useful in a SAHS home monitoring system based on pulse oximetry.

REFERENCES

[1] AASM Task Force, "Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research," *Sleep*, vol. 22, no. 5, pp. 667–689, 1999.

[2] T. Young et al, "Epidemiology of obstructive sleep apnea: a population health perspective," *American journal of respiratory and critical care medicine*, vol. 165, no. 9, pp. 1217–1239, 2002.

[3] T. D. Bradley and J. S. Floras, "Obstructive sleep apnoea and its cardiovascular consequences," *The Lancet*, vol. 373, no. 9657, pp. 82–93, 2009.

[4] J. M. Marin et al, "Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study," *The Lancet*, vol. 365, no. 9464, pp. 1046–1053, 2005.

[5] C. Iber et al, "The aasm manual for the scoring of sleep and associated events: rules, terminology and technical specifications," *American Academy of Sleep Medicine*, 2007.

[6] R. B. Berry et al, "Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events," *J Clin Sleep Med*, vol. 8, no. 5, pp. 597–619, 2012.

[7] P. E. Peppard et al, "Increased prevalence of sleep-disordered breathing in adults," *American journal of epidemiology*, vol. 177, no. 9, pp. 1006–1014, 2013.

[8] J. Verbraecken, "Applications of evolving technologies in sleep medicine," *Breathe*, vol. 9, no. 6, pp. 442–455, 2013.

[9] C. Varon et al, "A novel algorithm for the automatic detection of sleep apnea from single-lead ecg," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 9, pp. 2269–2278, 2015.

[10] P. De Chazal et al, "Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 6, pp. 686–696, 2003.

[11] J. Lázaro et al, "Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 1, pp. 240–246, 2014.

[12] N. Ben-Israel et al, "Obstructive apnea hypopnea index estimation by analysis of nocturnal snoring signals in adults," *Sleep*, vol. 35, no. 9, pp. 1299–1305, 2012.

[13] H. Nakano et al, "Automatic detection of sleep-disordered breathing from a single-channel airflow record," *European Respiratory Journal*, vol. 29, no. 4, pp. 728–736, 2007.

[14] J. F. Morales et al, "Sleep apnea hypopnea syndrome classification in spo 2 signals using wavelet decomposition and phase space reconstruction," in *Wearable and Implantable Body Sensor Networks (BSN), 2017 IEEE 14th International Conference on*. IEEE, 2017, pp. 43–46.

[15] R. Rolón et al, "Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea–hypopnea detection," *Biomedical Signal Processing and Control*, vol. 33, pp. 358–367, 2017.

[16] G. Schlotthauer et al, "Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry," *Medical engineering & physics*, vol. 36, no. 8, pp. 1074–1080, 2014.

[17] J.-C. Vázquez et al, "Automated analysis of digital oximetry in the diagnosis of obstructive sleep apnoea," *Thorax*, vol. 55, no. 4, pp. 302–307, 2000.

[18] E. Chiner et al, "Nocturnal oximetry for the diagnosis of the sleep apnoea hypopnoea syndrome: a method to reduce the number of polysomnographies?" *Thorax*, vol. 54, no. 11, pp. 968–971, 1999.

[19] D. Alvarez et al, "Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2816–2824, 2010.

[20] J. V. Marcos et al, "Automated prediction of the apnea-hypopnea index from nocturnal oximetry recordings," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 1, pp. 141–149, 2012.

[21] B. L. Koley and D. Dey, "On-line detection of apnea/hypopnea events using spo signal: A rule-based approach employing binary classifier models," *IEEE journal of biomedical and health informatics*, vol. 18, no. 1, pp. 231–239, 2014.

[22] A. Burgos et al, "Real-time detection of apneas on a pda," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 995–1002, 2010.

[23] V. Moret-Bonillo et al, "Intelligent approach for analysis of respiratory signals and oxygen saturation in the sleep apnea/hypopnea syndrome," *The open medical informatics journal*, vol. 8, p. 1, 2014.

[24] A. Yadollahi et al, "Sleep apnea monitoring and diagnosis based on pulse oximetery and tracheal sound signals," *Medical & biological engineering & computing*, vol. 48, no. 11, pp. 1087–1097, 2010.

[25] P. De Chazal et al, "Multimodal detection of sleep apnoea using electrocardiogram and oximetry signals," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1887, pp. 369–389, 2009.

[26] S. F. Quan et al, "The sleep heart health study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[27] S. Redline et al, "Methods for obtaining and analyzing unattended polysomnography data for a multicenter study," *Sleep*, vol. 21, no. 7, pp. 759–767, 1998.

[28] D. A. Dean et al, "Scaling up scientific discovery in sleep medicine: the national sleep research resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016.

[29] T. Penzel et al, "The apnea-ecg database," in *Computers in cardiology 2000*. IEEE, 2000, pp. 255–258.

[30] A. L. Goldberger et al, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[31] A. Kulkas et al, "Novel parameters indicate significant differences in severity of obstructive sleep apnea with patients having similar apnea–hypopnea index," *Medical & biological engineering & computing*, vol. 51, no. 6, pp. 697–708, 2013.

[32] E. Watanabe et al, "Prognostic importance of novel oxygen desaturation metrics in patients with heart failure and central sleep apnea," *Journal of Cardiac Failure*, 2016.

[33] A. Bauer et al, "Phase-rectified signal averaging detects quasi-periodicities in non-stationary data," *Physica A: Statistical Mechanics and its Applications*, vol. 364, pp. 423–434, 2006.

[34] J. W. Kantelhardt et al, "Phase-rectified signal averaging for the detection of quasi-periodicities and the prediction of cardiovascular risk," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 17, no. 1, p. 015112, 2007.

[35] H. Peng et al, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[36] J. A. Suykens et al, *Least squares support vector machines*. World Scientific, 2002.

[37] C. Varon et al, "Noise level estimation for model selection in kernel pca denoising," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 11, pp. 2650–2663, 2015.

[38] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[39] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of human genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[40] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics-theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[42] C. W. Whitney et al, "Reliability of scoring respiratory disturbance indices and sleep staging," *Sleep*, vol. 21, no. 7, pp. 749–757, 1998.