Systematic Comparison of Respiratory Signals for the Automated Detection of Sleep Apnea

Tom Van Steenkiste[1], Willemijn Groenendaal[2], Joeri Ruyssinck[1], Pauline Dreesen[3], Susie Klerkx[4], Christophe Smeets[3], Ruben de Francisco[2], Dirk Deschrijver[1], and Tom Dhaene[1]

Abstract—Sleep apnea is one of the most common sleep disorders. It is characterized by the cessation of breathing during sleep due to airway blockages (obstructive sleep apnea) or disturbances in the signals from the brain (central sleep apnea). The gold standard for diagnosing sleep apnea is performing an overnight polysomnography recording which contains, among others, a wide array of respiratory signals. Respiration information can also be extracted from other physiological signals such as an electrocardiogram or from a bio-impedance measurement on the chest. Studies have shown that algorithms can be developed for automated sleep apnea detection using one of these many respiratory signals. In this work, the predictive power of these different respiratory signals is analyzed and compared. The results provide useful insights into the comparative predictive power of the different respiratory signals in a realistic setting for automated sleep apnea detection and provide a basis for the development of less obtrusive measurement techniques.

I. INTRODUCTION

Sleep apnea is one of the most common sleep related disorders. It is characterized by the occurrence of breathing pauses of at least 10 seconds (apneic episodes) during the night [1] leading to frequent awakenings. If left untreated, the patient is at risk of hypertension, cardiac arrhythmia, heart attacks and strokes [2], and has an increased chance of motor vehicle collisions [3]. Sleep apnea is classified as either Obstructive Sleep Apnea (OSA) in which the airway is blocked by the throat muscles, Central Sleep Apnea (CSA) in which the signals to control the breathing are disturbed or hypopnea in which the breathing is becoming too shallow. Hypopnea can be further categorized as obstructive or central hypopnea. Although some studies report that an estimated 49.7% of male and 23.4% of female adults suffer from sleep-disordered breathing [4], many cases remain undiagnosed as patients are rarely aware of their condition.

The gold standard to detect and subsequently diagnose sleep apnea is an overnight polysomnography (PSG) in a specialized sleep laboratory [5]. During this PSG, multiple physiological signals pertaining to respiration, oxygen saturation, cardiovascular functioning and sleep status are recorded.

[1]Tom Van Steenkiste, Joeri Ruyssinck, Dirk Deschrijver and Tom Dhaene are with Ghent University - imec, IDLab Technologiepark-Zwijnaarde 15, B-9052 Ghent, Belgium tomd.vansteenkiste@ugent.be

[2]Willemijn Groenendaal and Ruben de Francisco are with imec Netherlands, imec, Eindhoven, The Netherlands

[3]Pauline Dreesen and Christophe Smeets are with Future Health, Ziekenhuis Oost-Limburg, Genk, Belgium and Mobile Health Unit, Faculty of Medicine and Life Sciences, Hasselt University, Hasselt, Belgium.

[4]Susie Klerkx is with Department of Pneumology, Ziekenhuis Oost-Limburg, Genk, Belgium

Then, trained sleep technicians evaluate the recorded signals using a standard reference such as the AASM guidelines [5] and annotate each part of the signal as either OSA, CSA, hypopnea or no apnea. These annotations are used to compute the Apnea-Hypopnea-Index (AHI) which is the number of apnea and hypopnea events per hour and which is used to categorize the severity of the condition from mild to severe.

Due to the large amount of recorded signals, PSG studies can be uncomfortable for the patient leading to an unrepresentative night of sleep. Furthermore, the shortage of beds and staff in medical centers typically results in long waiting times (2-10 months in the US, 7-60 months in the UK) [6].

Much research has been done into reducing the discomfort of the PSG recording and waiting times by developing portable home monitoring devices [7]. Such devices record a subset of the standard physiological signals in a PSG and often contain one or more respiration measurements originating from a chest band, pressure sensor, thermal sensor or other device. Respiration can also be accurately estimated from other physiological signals such as the electrocardiogram (ECG). This estimate is known as the ECG Derived Respiration (EDR) signal [8] and can be computed in various ways. A measurement of the chest bio-impedance can also be used to estimate the respiration as impedance changes reflect changes in the volume of air within the patient's lungs [9].

Sleep apnea diagnosis is challenging, as it requires the analysis of interactions within a wide array of physiological signals recorded by a full PSG. In this paper, only respiratory information is used to screen potential sleep apnea patients.

Various studies have demonstrated algorithms for the detection of sleep apnea using one of these signals. In this work, we perform a comparison of these different respiratory signals and their predictive power for the automated detection of sleep apnea using machine learning. We not only include OSA, but also CSA as well as hypopnea events. Furthermore, we perform this analysis on a real-life dataset gathered in a clinical setting without any manual cleaning to reflect a realistic measurement. The predictive power of the respiratory signals is analyzed and compared using three different machine learning models: Artificial Neural Networks (ANN), Logistic Regression (LR) and Random Forests (RF).

In Section II, we discuss the data gathering and various respiratory signals used in the study. In Section III, the experimental setup is discussed. Next, in Section IV, results are presented and their consequences examined. Finally, in Section V, conclusions are made.

II. RESPIRATION DATASET

The study uses real-life data gathered from 23 patients enrolled for an overnight PSG analysis in a clinical setting at Ziekenhuis Oost-Limburg, a hospital in Belgium. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethical Committee before study onset (CME ZOL, reference: 16/042U). All participants provided written informed consent before inclusion. The data, consisting of the gold-standard PSG recording and the chest bio-impedance measurement, was annotated by trained technicians following the AASM guidelines [5] for OSA, CSA and hypopnea. In the annotations, no distinction is made between obstructive or central hypopnea.

The set of patients consists of 6 female and 17 male subjects with mean age 57.5 years (std ± 13.5 years). There are 3 patients with normal AHI, 8 with mild AHI, 4 with moderate AHI and 8 with severe AHI.

The PSG recording consists of various respiratory signals recorded at 200Hz:

- Abd. belt: Abdominal respiratory belt below the lower edge of the left ribcage.
- Thor. belt: Thoracic respiratory belt below left armpit.
- V_{tot}: Total relative chest volume which is a combination of the abdominal and ribcage volume [10].
- Cannula: Pressure from nasal cannula positioned between nose and upper lip.
- Thermistor: Nasal thermistor positioned between the nose and upper lip, on top of the nasal cannula.

The PSG recording has a single-lead ECG (lead I) recorded at 200Hz. From this ECG signal, RR intervals are extracted using a Pan-Tompkins beat detector [11]. From these RR intervals, the instantaneous Heart Rate (HR) is extracted. These ECG, RR and HR signals are subsequently used to extend the respiration dataset with three commonly used EDR signals:

- EDR1: Cubic spline interpolation of the R-peak amplitudes as the amplitude of these peaks is modulated by the respiration.
- EDR2: Filtered ECG signal using a combination of a low-pass filter with cutoff frequency of 0.4 Hz and a high-pass filter with cutoff frequency of 0.2Hz [12]. Both filters have a zero phase shift.
- EDR3: Filtered HR signal using the same filtering method as in EDR2.

Finally, the respiration dataset also includes the bio-impedance (BioZ) signal measuring the impedance of the body at 1024Hz using a small sensor placed in the middle of the chest. As the impedance recordings are complex-valued, the magnitude of the bioZ signal is used in the next steps .

III. EXPERIMENTAL SETUP

The experimental setup consists of three different machine learning models to assess the predictive power of the various respiration signals for the prediction of sleep apnea events. The respiration signals are split into overlapping sections of 30 seconds, each 1 second apart. The annotation provided by the trained sleep technicians at the end of the section is used as the annotation label for that section. Before the signal sections are used in machine learning models, they are first preprocessed and features are extracted from them. To reflect a realistic measurement, all signals are processed automatically and no manual cleaning is performed.

A. Preprocessing

Both the directly measured as well as the derived respiratory signals contain noise due to subject movement, electrical inference, measurement noise and other disturbances. To extract all relevant respiration information, the signals are passed through a fourth-order low-pass zero-phase-shift Butterworth filter with a cutoff frequency of 0.7Hz. Next, the moving mean computed with a width of 4 seconds is subtracted from the original signal to remove baseline wander and large motion artifacts. Finally, the signals are resampled at 5Hz to reduce the computational burden in later steps. After the preprocessing stage, the signals are split up into overlapping 30 second sections as previously detailed. These are then individually preprocessed by rescaling them and subtracting a baseline to get the minimum section value to 0 and the maximum section value to 1. This ensures all characteristics in the respiration signal are equally visible regardless of the signal strength.

B. Feature Extraction

After preprocessing, each 30 second section is passed through the feature extraction phase. Typical discriminative features for sleep apnea, sleep studies and biomedical health in general, are extracted, both in the time-domain as well as the frequency-domain [13], [14]. The time-domain features include the signal mean, the signal standard deviation, the signal skewness and the area under the absolute value of the signal. Furthermore, peak detection is performed to locate respiratory peaks. These locations are then used to add extra features such as mean peak height, standard deviation of peak height and skewness of peak height together with the number of peaks and the mean inter-peak distance, standard deviation of the inter-peak distance, skewness of the inter-peak distance and the sum of the peak heights. The frequency-domain features are computed by computing the Power Spectral Density (PSD) of each section and extracting the peak frequency, mean frequency, central frequency and band-power.

C. Modeling

The features extracted via the previous processes are used in three different machine learning models for each respiratory signal: a two-layer Artificial Neural Network (ANN) model, a Logistic Regression (LR) model [15] and a Random Forest (RF) model [16] as these are commonly used in sleep apnea and other sleep or medical studies [17].

The ANN and RF models have built-in feature relevance detection. The LR model has an additional feature selection stage in which the most important features are selected to prevent overfitting. The model hyperparameters, i.e. the

TABLE I
OVERVIEW OF RESULTS WITH THE ANN MODEL FOR ALL RESPIRATORY SIGNALS. EACH SIGNAL HAS BEEN TESTED FOR THE PREDICTIVE PERFORMANCE FOR OSA (O), CSA (C), HYPOPNEA (H) OR A COMBINATION THEREOF.

	Abd. belt		Thor. belt		V_{tot}		Cannula		Thermistor		BioZ		EDR1		EDR2		EDR3	
	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp
o	0.76	0.72	0.77	0.62	0.81	0.71	0.46	0.81	0.86	0.75	0.59	0.74	0.69	0.44	0.53	0.52	0.62	0.52
c	0.62	0.82	0.88	0.69	0.85	0.71	0.85	0.74	0.65	0.74	0.76	0.70	0.73	0.78	0.62	0.50	0.76	0.55
h	0.67	0.59	0.47	0.60	0.60	0.66	0.58	0.63	0.72	0.63	0.61	0.62	0.56	0.48	0.61	0.41	0.54	0.54
oc	0.59	0.82	0.80	0.66	0.42	0.80	0.88	0.74	0.86	0.73	0.80	0.69	0.61	0.64	0.74	0.39	0.69	0.49
och	0.71	0.65	0.55	0.57	0.73	0.63	0.72	0.65	0.78	0.65	0.67	0.64	0.62	0.48	0.71	0.33	0.62	0.47

number of hidden nodes within each layer for ANN, the regularization parameter for LR and the number of trees and maximum number of features per split for RF, are optimized using Bayesian optimization (BO). For a detailed description of BO, we refer to [18]. The metric used for this optimization is discussed in the evaluation section. This hyperparameter optimization ensures the full predictive power of the model is harnessed and overfitting is reduced. To further reduce overfitting for the ANN model, early-stopping is employed based on 10% of the left-out training data for validation.

All models are validated using 5-fold cross-validation where care has been taken that no fold has data of a patient already represented in another fold to ensure the models are capable of accurate predictions for new, unseen patients.

As apnea and hypopnea events are relatively rare, special care needs to be taken with this imbalance during training. If the models are trained using the original data distribution, the models would be too inclined to always predict a non-apnea event. Instead, we balance the training dataset by randomly selecting as many non-apnea events from the training data as there are apnea events. However, for evaluation purposes, it is important that the original imbalanced data distribution is maintained during the test phase.

To analyze the strength of the model to detect certain apnea types, several models are built using subsets of the data. This includes only OSA events (o), only CSA events (c), only hypopnea events (h), only apnea events (oc) or any type of apnea or hypopnea events (och). All detected apnea events are considered sleep apnea.

D. Evaluation

The main evaluation metrics used in the comparison are the Sensitivity (Se), also known as recall, and Specificity (Sp) metrics as these are commonly used in sleep apnea and other biomedical literature. These two metrics can also be summarized into an Informedness metric.

$$Se = TP/(TP + FN) \quad (1)$$
$$Sp = TN/(TN + FP) \quad (2)$$
$$Informedness = Se + Sp - 1 \quad (3)$$

The sensitivity and specificity metrics are influenced by a decision threshold τ. The various combinations of sensitivity and specificity at a specific τ can be summarized in a Receiver Operator Characteristic (ROC) and the area under this curve (AUROC). This AUROC metric is often used as

a hyperparameter optimization metric for machine learning models. However, it is not suited for problems with a large data imbalance. In such cases, the Area Under the Precision Recall Curve (AUPRC) metric is often preferred as an alternative. This AUPRC metric is used in this work for tuning the model hyperparameters.

IV. RESULTS AND DISCUSSION

The sensitivity and specificity scores for all experiments with the ANN model are shown in Table I. They can also be summarized using the informedness metric as shown in Fig. 1. The experiments were also run with an LR model and an RF model, both of which yielded similar results.
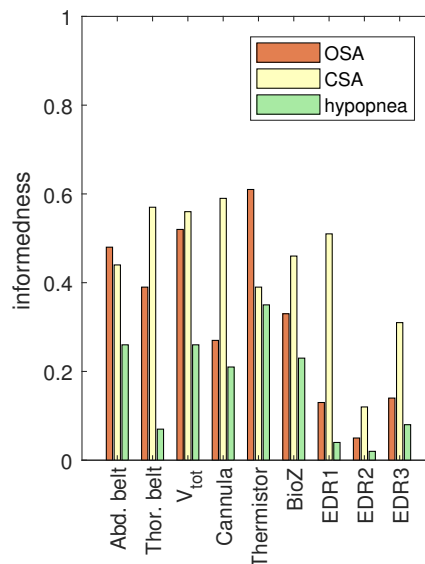

Fig. 1. Informedness metric for detection of OSA, CSA and hypopnea events in respiratory signals. The nasal cannula and nasal thermistor yield the best overall performance.

The results show that OSA and CSA events are in general the easiest to automatically detect whereas hypopnea events are more difficult. Constructing one model to detect several different types of apnea such as a combination of OSA and CSA or a combination of OSA, CSA and hypopnea reduces the predictive power of the model. This is due to the different origins of the events and resulting differences in features. As the hypopnea annotation is a combination of obstructive and central hypopnea, improvements could be made by creating a separate model for these two types of event.

The best performing signals across all apnea types are the direct measurements of respiration using the nasal cannula and the nasal thermistor. Each of the respiration belts also offers a good performance individually, but the combination of abdominal and thoracic respiration into the V_{tot} signal produces even better results. The performance of EDR1 for the detection of CSA is similar to the direct measurements. However, for the other cases, the performance of the EDR signals is substantially lower. Although literature suggests a strong performance, it is very susceptible to noise in realistic clinical settings. The bio-impedance measurement demonstrates promising performance both with OSA and CSA, as well as with the harder-to-detect hypopnea events.

On top of the predictive power of the different respiratory signals, the form factor also has to be taken into account. Although the directly measured respiratory signals provide a high predictive power, they are less comfortable and restrict the patient during the night. The EDR and bio-impedance signals on the other hand can be measured using a small sensor without inflicting too much discomfort on the patient.

This study represents a first step towards developing home monitoring devices. As mentioned in the introduction, a full PSG recording is uncomfortable and might lead to an unrepresentable night of sleep. However, the full PSG setup is required by sleep technicians to accurately annotate the signals. In future steps, the results from this study can be validated using small unobtrusive devices that only measure relevant respiratory signals.

The comparison in this work is useful to guide development of sleep apnea detection hardware and software. However, it is important to note that as these classification models only use respiratory information, they cannot correctly diagnose sleep apnea as other factors (such as oxygen saturation) have to be taken into account. The limitation of only using respiratory information leads to false positives caused by the erroneous detection of short events. The AASM guidelines only allow the annotation of events longer than 10 seconds. Nevertheless, short events can easily be filtered out in a later stage after which the larger events can be used to compute the AHI.

V. Conclusion

Sleep apnea is one of the most common sleep disorders and detection is important as the health risks can be severe. However, beds and trained staff are limited. In this work, we investigated and compared the predictive power of respiratory signals to either quickly indicate interesting sections to trained staff in a full PSG or to screen potential patients in their home. The results show that typical PSG signals have the largest predictive power. However, signals from less obtrusive sensors such as the ECG derived respiration can also be used although the predictive power is significantly less due to the influence of several types of noise. The bio-impedance signal, consisting of only a small sensor placed on the patients chest, shows promising performance. These results can be used to further guide algorithm and hardware

developments for the automated detection of sleep apnea in home environments and to provide practical insights into which respiratory signals might be useful.

References

[1] C. Guilleminault, A. Tilkian, and W. C. Dement, "The sleep apnea syndromes," *Ann. review of medicine*, vol. 27, no. 1, pp. 465–484, 1976.

[2] V. K. Somers, D. P. White, R. Amin, W. T. Abraham, F. Costa, A. Culebras, S. Daniels, J. S. Floras, C. E. Hunt, L. J. Olson, and Others, "Sleep apnea and cardiovascular disease," *Circulation*, vol. 118, no. 10, pp. 1080–1111, 2008.

[3] A. Sassani, L. J. Findley, M. Kryger, E. Goldlust, C. George, and T. M. Davidson, "Reducing motor-vehicle collisions, costs, and fatalities by treating obstructive sleep apnea syndrome," *Sleep*, vol. 27, no. 3, pp. 453–458, 2004.

[4] R. Heinzer, S. Vat, P. Marques-Vidal, H. Marti-Soler, D. Andries, N. Tobback, V. Mooser, M. Preisig, A. Malhotra, G. Waeber, *et al.*, "Prevalence of sleep-disordered breathing in the general population: the hypnolaus study," *The Lancet Resp. Medicine*, vol. 3, no. 4, pp. 310–318, 2015.

[5] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, "The AASM manual for the scoring of sleep and associated events," *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 2012.

[6] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to diagnosis and treatment of patients with suspected sleep apnea," *American journal of respiratory and critical care medicine*, vol. 169, no. 6, pp. 668–672, 2004.

[7] W. W. Flemons, M. R. Littner, J. A. Rowley, P. Gay, W. M. Anderson, D. W. Hudgel, R. D. McEvoy, and D. I. Loube, "Home diagnosis of sleep apnea: a systematic review of the literature: an evidence review cosponsored by the American Academy of Sleep Medicine, the American College of Chest Physicians, and the American Thoracic Society," *CHEST Journal*, vol. 124, no. 4, pp. 1543–1579, 2003.

[8] G. B. Moody, R. G. Mark, M. A. Bump, J. S. Weinstein, A. D. Berman, J. E. Mietus, and A. L. Goldberger, "Clinical validation of the ecg-derived respiration (edr) technique," *Group*, vol. 1, no. 3, 1986.

[9] C. Poon, Y. Chung, T. Choy, and J. Pang, "Evaluation of two noninvasive techniques for exercise ventilatory measurements," in *Engineering in Medicine and Biology Society, 1988. Proceedings of the Ann. Int. Conference of the IEEE*. IEEE, 1988, pp. 823–824.

[10] B. A. Staats, H. W. Bonekat, C. D. Harris, and K. P. Offord, "Chest wall motion in sleep apnea 1, 2," *American Review of Respiratory Disease*, vol. 130, no. 1, pp. 59–63, 1984.

[11] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE transactions on biomedical engineering*, no. 3, pp. 230–236, 1985.

[12] C. Gari D., A. Francisco, and M. Patrick E., *Advanced Methods and Tools for ECG Data Analysis*, C. Gari D., A. Francisco, and M. Patrick E., Eds. Artech House, 2006.

[13] C. Song, K. Liu, X. Zhang, L. Chen, and X. Xian, "An obstructive sleep apnea detection approach using a discriminative hidden markov model from ecg signals," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1532–1542, 2016.

[14] L. S. Correa, E. Laciar, V. Mut, A. Torres, and R. Jané, "Sleep apnea detection based on spectral analysis of three ecg-derived respiratory signals," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 4723–4726.

[15] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[16] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier," *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 10–19, 2012.

[17] D. Alvarez, R. Hornero, J. V. Marcos, and F. del Campo, "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.

[18] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.