Multi-Class AdaBoost to Detect Sleep Apnea-Hypopnea Syndrome Severity from Oximetry Recordings Obtained at Home

G. C. Gutiérrez-Tobal[1], D. Álvarez[1,2], A. Crespo[2], C. A. Arroyo[2], F. Vaquerizo-Villar[1], V. Barroso-García[1], F. del Campo[1,2], and R. Hornero[1]

[1]Biomedical Engineering Group, Universidad de Valladolid, Valladolid, España

[2]Sleep Unit of Hospital Universitario Rio Hortega, Valladolid, España

Email: gonzalo.gutierrez@gib.tel.uva.es

Abstract — **This paper aims at evaluating a novel multi-class methodology to establish Sleep Apnea-Hypopnea Syndrome (SAHS) severity by the use of single-channel at-home oximetry recordings. The study involved 320 participants derived to a specialized sleep unit due to SAHS suspicion. These were assigned to one out of the four SAHS severity degrees according to the apnea-hypopnea index (AHI): no-SAHS (AHI<5 events/hour), mild-SAHS (5≤AHI<15 e/h), moderate-SAHS (15≤AHI<30 e/h), and severe-SAHS (AHI≥30 e/h). A set of statistical, spectral, and non-linear features were extracted from blood oxygen saturation (SpO_2) signals to characterize SAHS. Then, an optimum set among these features were automatically selected based on relevancy and redundancy analyses. Finally, a multi-class *AdaBoost* model, built with the optimum set of features, was obtained from a training set (60%) and evaluated in an independent test set (40%). Our *AdaBoost* model reached 0.386 Cohen's kappa in the four-class classification task. Additionally, it reached accuracies of 89.8%, 85.8%, and 74.8% when evaluating the AHI thresholds 5 e/h, 15 e/h, and 30 e/h, respectively, outperforming the classic oxygen desaturation index. Our results suggest that SpO_2 obtained at home, along with multi-class *AdaBoost*, are useful to detect SAHS severity.**

Keywords — **AdaBoost, at-home oximetry, feature extraction, feature selection, sleep apnea severity.**

I. Introduction

Sleep Apnea-Hypopnea Syndrome (SAHS) is a respiratory chronic disease that worsens both health and quality of life of affected people [1]. It is characterized by the recurrence of apneas and hypopneas during sleep, i.e., events of breathing cessation or significant airflow reduction, respectively [2]. These events lead to inadequate overnight gas exchange, which derives in blood oxygen saturation drops and arousals, causing fragmented and restless sleep [1], [2]. SAHS daytime symptoms include hypersomnolence, cognitive impairment, and depression [3]. Furthermore, SAHS has been related to severe illnesses such as myocardial infarction, cardiac failure, and stroke, as well as with an increase in cancer incidence [1], [4]. Moreover, SAHS is highly prevalent, affecting up to 2% women and 5% men in western countries [3].

SAHS diagnosis is derived from polysomnography test (PSG), which acts as "gold standard" [2]. Patients sleep in specialized facilities where up to 32 biomedical signals are monitored and recorded from them, including electroencephalogram, electrocardiogram, airflow, and blood oxygen saturation (SpO_2) [2]. After PSG, physicians inspect these recordings to obtain the apnea-hypopnea index (AHI), i. e., the number of apneas and hypopneas per hour of sleep. Then, one out of four severity degrees is assigned to the subject under study according to AHI: no SAHS (AHI < 5 events/hour), mild SAHS (5 ≤ AHI < 15 e/h), moderate SAHS (15 ≤ AHI < 30 e/h), and severe SAHS (AHI ≥ 30 e/h) [5].

In spite of its effectiveness, PSG is technically complex, costly, time-consuming, and deprives patients from their natural sleep environment [6]. These drawbacks have led to the search for alternatives focused on simplifying SAHS diagnosis. In this regard, overnight pulse oximetry has been commonly investigated as single-channel diagnostic alternative for SAHS [7]-[10]. Pulse oximetry is a standard for monitoring and recording the SpO_2 signal [11], which is able to reflect blood oxygen desaturations caused by apneic events. A few works have analyzed the diagnostic ability of conventional clinical parameters obtained from SpO_2, such as oxygen desaturation index (ODI) or delta index [8], [12] whereas other studies have focused on automated methodologies based on signal processing and pattern recognition techniques [9], [10], [13]. However, SpO_2 signals are commonly acquired during in-hospital PSG and, consequently, there exists a lack of evaluation of these methodologies when using unattended at-home recordings.

The main purpose of this study is the assessment of an automated methodology to detect SAHS severity by the only use of SpO_2 data obtained from at-home overnight pulse oximetry. A set of features from different analytical approaches (statistical, spectral and non-linear), already used to characterize SAHS in in-lab SpO_2 signals [9], [10], was extracted from 320 at-home recordings. Moreover, an automatic feature selection stage, based on the fast correlation-based filter (FCBF) method [14], was implemented. This algorithm has been successfully used to discard features that provide similar information (redundant features) in biomedical applications, including SAHS diagnosis [15], [16]. Finally, the adaptive boosting (*AdaBoost*) method was chosen to perform the multi classification task, i.e., to establish SAHS severity, by the use of the non-redundant features previously selected. *AdaBoost* combines several classifiers of the same type in order to reach higher performance than each one separately [17]. It is known to be able to reach generalized models and has been already used in SAHS context applied to airflow signals [16].

Our hypothesis is that single-channel SpO_2 obtained at patient's home provides relevant information to help to determine SAHS severity reliably by the use of a generalizable model.

II. Subjects and Signals

The study involved 320 subjects derived to the specialized sleep unit of the Hospital Universitario Rio Hortega de Valladolid (Spain) due to SAHS suspicion. Participants underwent an overnight PSG test (E-series, Compumedics) to obtain their diagnosis. AHI from PSG was computed according to the rules of the American Academy of Sleep Medicine (AASM) [18], as well as used by the specialists to assign one out of the four SAHS severity degrees to subjects (no-SAHS, mild-SAHS, moderate-SAHS, and severe-SAHS). Participants also underwent an at-home nocturnal pulse oximetry in order to acquire the corresponding SpO_2 signals. Overnight pulse oximetry was carried out within the pre or post 24 hours from PSG to minimize the night-to-night sleep variability effect. All the subjects gave their informed consent and the Ethics Committee of the Hospital Universitario Rio Hortega accepted the protocol. Subjects were divided into two sets: training (60%, 193 first consecutive participants) and test (40%, 127 remaining consecutive participants). Table I shows demographical and clinical data (mean ± standard deviation) from participants according to their SAHS severity degree. Suspicion of SAHS biases the final diagnosis of people referred to sleep units. Hence, a clear imbalance can be observed in the number of subjects assigned to each class. Not balanced classes favor right classification of most frequent ones, affecting training of predictive models. Consequently, we used the synthetic minority oversampling technique (SMOTE) [19], to compensate for this imbalance.

SpO_2 signal was acquired at 1 Hz sample rate using a portable oximeter (Nonin WristOx2 3150). Artifacts due to movements were automatically removed during preprocessing. Thus, SpO_2 values equal to zero as well as differences between consecutive SpO_2 samples ≥4% were considered artifacts [8]. Removed samples were substituted by interpolated data.

II. Methodology

A three-step methodology was carried out during the study. First, we implemented a feature extraction stage in which up to 16 statistical, non-linear and spectral features were acquired from SpO_2 at-home recordings. Conventional 3% ODI (ODI3) was also computed for comparison purposes. Then, an automatic feature selection methodology was conducted to discard redundant features. Finally, *AdaBoost* algorithm was trained with the non-redundant ones and tested as multi-class classifier in order to obtain SAHS severity degree from the participants in the study.

TABLE I
Demographic and clinical data of subjects under study

	no-SAHS	**mild**	**moderate**	**severe**
# Subjects	29	55	56	180
Age (years)	44.5 ± 15.1	55.9 ± 12.5	55.4 ± 14.5	56.5 ± 12.5
Men (%)	13 (44.8)	38 (69.1)	44 (78.6)	141 (78.3)
BMI(kg/m^2)	24.8 ± 3.9	26.8 ± 4.2	27.7 ± 3.9	31.2 ± 5.6
AHI (e/h)	2.8 ± 1.5	9.8 ± 2.8	21.7 ± 4.3	59.4 ± 23.4

A. Feature extraction and selection

1) Statistical features

First-to-fourth statistical moments were obtained in time domain. These were the well-known mean (M_{t1}), standard deviation (M_{t2}), skewness (M_{t3}), and kurtosis (M_{t4}), which measure central tendency, dispersion, asymmetry, and peakedness of data, respectively.

2) Non-linear features

SAHS has been proven to modify variability, complexity, and irregularity of SpO_2 [9], [12]. Hence, three non-linear features were also acquired from this signal in time domain. These were central tendency measure (*CTM*), Lempel-Ziv complexity (*LZC*), and sample entropy (*SampEn*). *CTM* quantifies variability in a time series based on first differences plots [9], [12]. It ranges between 0 and 1, with values closer to 0 indicating higher degree of variability. *LZC* provides a complexity measure of time series transformed into a finite sequence of symbols [9], [12]. This sequence is scanned to find different subsequences. The higher the number of different subsequences, the higher the complexity of the original time series. Finally, *SampEn* quantifies the irregularity of a time series based on similarities among vectors formed with its own samples. Thus, higher values of *SampEn* indicate less self-similarity in the times-series and, consequently, more irregularity [9], [16].

3) Spectral features

Recurrence of apneic events justifies conducting analyses in the frequency domain. Hence, up to 9 features were extracted from power spectral density (PSD) of the SpO_2 signals. First-to-fourth statistical moments were also obtained from PSD to analyze its data distribution (M_{f1}-M_{f4}). Additionally, spectral total power (P_T) was computed as the area comprised within the whole PSD. Two more features were directly derived from the typical frequency band of interest, 0.014-0.033 Hz: relative power (P_R), computed as the proportion of power falling within the band of interest with respect to the total power, and peak amplitude (*PA*), computed as the maximum PSD value in the band of interest. Finally, median frequency (*MF*) and spectral entropy (*SpecEn*) were also computed. The former is the frequency that splits PSD into two regions, each one containing 50% of the total spectral power [16], whereas the latter quantifies the flatness of the spectrum as a measure of regularity [16].

4) *Feature selection: the fast correlation-based filter*

FCBF is an automatic feature selection algorithm based on relevancy and redundancy analyses [14]. Symmetrical uncertainty (*SU*) is used to determine the information shared by the extracted features and AHI, which is taken as a reference variable. Additionally, *SU* between each pair of features is also computed [14]. Those features sharing more information with AHI are considered more relevant and ranked higher. Then, the features sharing more information with other higher ranked features than with AHI are considered redundant and, consequently, not selected for subsequent analyses [14].

B. Multi-class classification: AdaBoost.M2

Boosting methods are iterative algorithms used to combine models that complement one another [20]. This combination is carried out based on the weighted votes of the classifiers trained at each iteration, which are of the same type [17], [20]. *AdaBoost* is a boosting algorithm which is typically used along with simple or "weak" classifiers in order to reach generalized models [20]. In our case, we have chosen classification and regression trees models (CART) to act as weak classifiers. This combination of *AdaBoost* and CART has been successfully used in SAHS context applied to data from airflow signal [16].

AdaBoost relies on reweighting those instances that have been misclassified after each iteration. Thus, CART models trained during later iterations give more importance to these instances [17], being more likely to classify them rightly [20]. In this study, a multi-class classification is proposed. Hence, the *AdaBoost.M2* version of the algorithm has been used. All CART models iteratively trained using *AdaBoost.M2* are associated to an error based on their corresponding performance. A classifier weight is derived from this error so that the final classification task is conducted by returning the class with the highest sum of the weighted votes from all classifiers [17].

C. Statistical analysis

Diagnostic performance of the multi-class *AdaBoost.M2* method was assessed in terms of Cohen's kappa (κ). Additionally, for each AHI cutoff involved in SAHS severity degrees (5 e/h, 15 e/h, and 30 e/h), the following diagnostic statistics were computed: sensitivity (Se, percentage of positive subjects rightly classified), specificity (Sp, percentage of negative subjects rightly classified), and accuracy (Acc, overall percentage of subjects rightly classified).

SMOTE was applied to the minority classes from the training set, no-SAHS (19 subjects), mild-SAHS (31 subjects), and moderate-SAHS (35 subjects), in order to compensate for imbalance with respect to severe-SAHS (108 subjects). Thus, synthetic samples were obtained to reach 114 no-SAHS subjects, 93 mild-SAHS subjects, and 105 moderate-SAHS subjects in the training set.

III. Results

A. Features automatically selected

The FCBF algorithm automatically selected 9 out of the 17 features by the only use of the 193 original training samples. According to their *SU* ranking, these were ODI3, *CTM*, *SampEn*, *LZC*, M_{t1}, P_R, M_{t4}, M_{t3}, and *SpecEn*. Therefore, features from all the analytical approaches proposed have been selected (statistical, non-linear, and spectral).

B. Model training

The 420 samples (both 193 original and 227 synthetic) of the training set, each one composed of the features previously selected, were used to feed the CART classifiers involved in the *AdaBoost.M2* algorithm. A low learning rate (α=0.1) and a high number of CART classifiers (L=2000) were used as strategy to deal with overfitting.

C. Diagnostic ability

Table II shows confusion matrices in the test set for the *AdaBoost.M2* method, as well as the classic clinic parameter ODI3. *AdaBoost.M2* rightly classified 74 out of 127 subjects

TABLE II. Confusion Matrices for *AdaBoost.M2* and ODI3 in the Test Set.

		AdaBoost.M2				ODI3			
	Estimated →	**no-SAHS**	**mild**	**moderate**	**severe**	**no-SAHS**	**mild**	**moderate**	**severe**
Actual	**no-SAHS**	8	2	0	0	3	7	0	0
	mild	6	10	7	1	2	17	5	0
	moderate	3	3	10	5	1	9	10	1
	severe	2	2	22	46	0	10	23	39

TABLE III. Diagnostic Ability of AdaBoost.M2 and ODI3 in the Test Set for AHI Cutoffs = 5 e/h, 15 e/h, and 30 e/h.

	AdaBoost.M2			**ODI3**		
	5	15	30	5	15	30
Se (%)	90.6	89.2	63.9	97.4	78.5	54.2
Sp (%)	80.0	76.5	89.1	30.0	85.3	98.2
Acc (%)	89.8	**85.8**	**74.8**	**92.1**	80.3	73.2
κ		**0.386**			0.351	

(58.3%), whereas ODI3 rightly classified 69 (54.3%). Table III displays diagnostic ability statistics derived from the previous confusion matrices, evaluated for the AHI cutoffs values 5 e/h, 15 e/h, and 30 e/h. *AdaBoost.M2* reached higher overall diagnostic performance for the multi-class classification task by achieving 0.386 Cohen's κ. Moreover, it also obtained higher Acc than ODI3 for 15 e/h and 30 e/h thresholds (85.68% and 74.8%, respectively), whereas the latter obtained higher Acc for the 5 e/h threshold (92.1 %). However, Acc of *AdaBoost.M2* is also high (89.8%) with a more balanced Se/Sp pair (90.6%/80.0% *vs*. 97.4%/30.0%).

IV. Discussion

A new *AdaBoost.M2* model combining CART classifiers has been proposed to detect SAHS severity degrees from unattended oximetry at home. The model combined statistical, non-linear, and spectral non-redundant features extracted from single-channel SpO_2 along with the classic clinical variable ODI3. It showed higher diagnostic performance than single ODI3 in the multi-class classification of subjects from a test set. Additionally, the model reached high overall Acc when evaluating the AHI thresholds 5 e/h (89.8%), 15 e/h (85.8%), and 30 e/h (74.8%). Acc of ODI3 was only higher in the case of 5 e/h (92.1%). However, ODI3 Se (97.4%) and Sp (30.0%) were highly unbalanced in that case, highlighting a poor discriminative ability.

Several studies have recently focused on helping in SAHS diagnosis by the use of oximetry data. Alvarez et al. (2010) [9], developed a logistic regression model with four statistical, non-linear, and spectral features extracted from SpO_2. They reported 92.0% Se, 85.4% Sp, and 89.7% Acc (after leave-one-out cross-validation, loo-cv) when classifying 148 subjects into SAHS-positive or SAHS-negative classes (AHI threshold = 10 e/h), i.e., conducting a binary classification task. Similarly, Sánchez-Morillo and Gross [13] also conducted binary classification (AHI threshold = 10 e/h) by training a probabilistic neural network model with five statistical, non-linear, spectral, and clinical features from the SpO_2 of 115 subjects. They reached 92.4% Se, 95.9% Sp, and 93.9% Acc (after loo-cv). Additionally, Alvarez et al. (2013) carried out a multicenter study involving SpO_2 recordings from 320 subjects focused on evaluating several binary classifiers (AHI threshold = 10 e/h) [21]. The highest performance reported was reached by a logistic regression model, which was trained with four statistical, non-linear, and spectral features extracted from SpO_2. It reached 95.2% Se, 86.0% Sp, and 88.7% Acc after a hold-out cross-validation strategy. Marcos et al followed a different approach by training a multi-layer perceptron neural network to estimate AHI [10]. This model was built with fourteen features extracted from 240 SpO_2 signals. For AHI thresholds = 5 e/h and 15 e/h, they reported 91.8%/94.9% Se, 58.8%/90.9 Sp, and 84.0%/93.1% Acc, respectively, after a hold-out cross-validation procedure. Finally, *AdaBoost* has been already used in the context of multi-class classification of SAHS severity degrees involving 317 single-channel nasal pressure airflow recordings. Gutiérrez-Tobal et al. reached 0.381 κ and accuracies of 84.9%, 80.2%, and 83.3% (hold-out cross-validation) when evaluating an *AdaBoost.M2* model built with CART classifiers for the AHI thresholds 5 e/h, 15 e/h, and 30 e/h, respectively [16].

Coherent with the lack of studies aimed at the automatic determination of SAHS severity, most of the above mentioned studies focus on binary classification, for which AHI = 10 e/h is a common discriminative threshold. Only studies of Marcos et al. and Gutiérrez-Tobal et al. determined SAHS severity. Both of them reported lower diagnostic ability than our *AdaBoost.M2* model for the AHI threshold = 5 e/h. Gutiérrez-Tobal et al also reported lower Acc in the case of 15 e/h but higher in the case of 30 e/h. Marcos et al. reported higher diagnostic ability for 15 e/h. Finally, all the found studies that focus on helping in SAHS diagnosis by automatic analysis of signals use physiological recordings obtained from patients during in-lab supervised PSG. By contrast, our study used SpO_2 recordings obtained during unattended at-home oximetry. Consequently, our results are more likely to reflect the behavior of such methodologies in the natural sleep environment of patients.

In spite of the high diagnostic performance reached by our proposal, some limitations need to be mentioned. First, more participants would be required in order to compensate the imbalance among SAHS severity degrees. However, we used the SMOTE methodology to minimize its effect in the training of our *AdaBoost.M2* model. More subjects would be also helpful to give more statistic robustness to our results. Nevertheless, our subject's database is large comparing with those from the state-of-the-art studies. Additionally, further validation of our methodology would be required in order to evaluate different learning rates and number of CART classifiers for the *AdaBoost.M2* algorithm.

V. Conclusions

We have developed an automatic multi-class *AdaBoost.M2* model trained with single-channel SpO_2 data in order to determine SAHS severity. In contrast to state-of the-art studies, SpO_2 data were acquired at patient's home without supervision. The new model showed high diagnostic ability, particularly when discriminating no-SAHS subjects from the remaining severity degrees. It also outperformed the clinical variable ODI3. Our results suggest that SpO_2 signal obtained from at-home oximetry contains relevant information to help in SAHS severity detection by means of the *AdaBoost.M2* algorithm.

ACKNOWLEDGMENT

This research has been partially supported by Consejería de Educación de la Junta de Castilla y León under project VA059U13 and Pneumology and Thoracic Surgery Spanish Society (SEPAR) under project 265/2012.

REFERENCES

[1] F. Lopez-Jiménez et al, “Obstructive Sleep Apnea,” Chest, vol. 133, pp 793-804, 2008.

[2] S. P. Patil, et al, “Adult Obstructive Apnea,” *Chest*, vol. 132, pp. 325-337, 2007.

[3] T. Young et al, “Epidemiology of Obstructive Sleep Apnea: A Population Health Perspective,” *Am. J. Respir. Crit. Care. Med.*, vol. 165, pp. 1217-1239, 2002.

[4] F. Campos-Rodriguez et al, “Association between obstructive sleep apnea and cancer incidence in a large multicenter spanish cohort,” *Am. J. Respir. Crit. Care Med.*, vol. 187, pp. 99-105, 2013.

[5] A. Qureshi et al, “Obstructive sleep apnea,” *J. Allergy Clin. Immunol.*, vol. 112, pp. 643-651, 2003.

[6] W. W. Flemons et al, “Home Diagnosis of Sleep Apnea: A Systematic Review of the Literature,” *Chest*, vol. 124, pp. 1543-1579, 2003.

[7] N. A. Collop et al, “Obstructive sleep apnea devices for out-of-center (OOC) testing: technology evaluation,” *J Clin Sleep Med*, vol. 7, pp.531–548, 2011.

[8] U. J. Magalang et al, “Prediction of the apnea-hypopnea index from overnight pulse oximetry,” *Chest*, vol. 124, pp. 1694–1701, 2003.

[9] D. Álvarez, et al, “Multivariate Analysis of Blood Oxygen Saturation Recordings in Obstructive Sleep Apnea Diagnosis,” *IEEE Trans Biomed Eng*, vol. 57, pp. 2816–2824, 2010.

[10] J. V. Marcos et al, “Automated prediction of the apnea-hypopnea index from nocturnal oximetry recordings,” *IEEE Trans Biomed Eng*, vol. 59, pp. 141-49, 2012.

[11] R. R. Kirby, R. W. Taylor, and J. M. Civetta, *Handbook of Critical Care*, 2nd ed. Philadelphia, PA: Lippincott-Raven, 1997.

[12] D. Álvarez, et al, “Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection,” *Phys Meas*, vol. 27, 399-412, 2006.

[13] D. S. Morillo and N. Gross, “Probabilistic neural network approach for the detection of SAHS from overnight pulse oximetry,” *Med Biol Eng Comput*, vol. 51, pp. 305-315, 2013.

[14] L. Yu and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” *J. Mach. Learn. Res.*, vol. 5, pp. 1205-1224, 2004.

[15] G. C. Gutiérrez-Tobal et al., “Pattern recognition in airflow recordings to assist in the sleep apnoea–hypopnoea syndrome diagnosis,” *Med. Biol. Eng. Comput.*, vol. 51, pp. 1367-80, 2013.

[16] G. C. Gutierrez-Tobal, et al., “Utility of AdaBoost to Detect Sleep Apnea-Hypopnea Syndrome from Single-Channel Airflow,” *IEEE Trans Biomed Eng*, vol. 63, pp. 636-646, 2015.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.

[18] 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(5), pp. 597-619, 2012.

[19] N. V. Chawla et al, “SMOTE: synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16(1), pp. 321-357, 2002.

[20] I. H. Witten, E. Frank and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann/Elsevier, 2011.

[21] D. Alvarez et al., “Assessment of feature selection and classification approaches to enhance information from overnight oximetry in the context of apnea diagnosis,” *Int J Neural Syst*, vol. 23, pp. 1350020, 2013.