Evaluation of Machine-Learning Approaches to Estimate Sleep Apnea Severity from at-Home Oximetry Recordings

Gonzalo C. Gutiérrez-Tobal*, *Member, IEEE*, Daniel Álvarez, Andrea Crespo, Félix del Campo, and Roberto Hornero, *Senior Member, IEEE*

Abstract— **Complexity, costs, and waiting lists issues demand a simplified alternative for sleep apnea-hypopnea syndrome (SAHS) diagnosis. The blood oxygen saturation signal (SpO_2) carries useful information about SAHS and can be easily acquired from overnight oximetry. In this study, SpO_2 single-channel recordings from 320 subjects were obtained at patients' home. They were used to automatically obtain statistical, spectral, non-linear, and clinical SAHS-related information. Relevant and non-redundant data from these analyses were subsequently used to train and validate four machine-learning methods with ability to classify SpO_2 signals into one out of the four SAHS-severity degrees (no-SAHS, mild, moderate, and severe). All the models trained (linear discriminant analysis, 1-vs-all logistic regression, Bayesian multi-layer perceptron, and AdaBoost), outperformed the diagnostic ability of the conventionally-used 3% oxygen desaturation index. An AdaBoost model built with linear discriminants as base classifiers reached the highest figures. It achieved 0.479 Cohen's κ in the SAHS severity classification, as well as 92.9%, 87.4%, and 78.7% accuracies in binary classification tasks using increasing severity thresholds (apnea-hypopnea index: 5, 15, and 30 events/hour, respectively). These results suggest that machine learning can be used along with SpO_2 information acquired at patients' home to help in SAHS diagnosis simplification.**

Index Terms— **At-home oximetry, ensemble learning, machine learning, neural networks, sleep apnea severity**

I. Introduction

THE Sleep Apnea-Hypopnea Syndrome (SAHS) has become a major focus of investigation over the last decades. The reasons for such interest include its severe consequences for health and quality of life of affected people, as well as its high prevalence [1]. Recent studies estimated that moderate to severe SAHS is present in 6% women and 13% men in the United States [2]. Patients suffer from recurrent episodes of complete absence of breathing (apneas) and significant airflow reduction (hypopneas) while sleeping, causing oxygen desaturations, arousals and, eventually, sleep fragmentation [3]. These undesirable effects lead to daytime symptoms such as hypersomnolence, cognitive impairment, and depression, which increase the risk for occupational accidents, absenteeism, and motor vehicle collisions [4], [5]. Furthermore, a significant number of pathological conditions have been related to SAHS, including hypertension, cardiac failure, and stroke [1]. Recently, an increase in cancer incidence has been also suggested [6].

In spite of its high prevalence, SAHS is considered an underdiagnosed condition [7]. Hence, its diagnostic protocol plays a key role to avoid time delays in reaching diagnosis and accessing treatment. Nocturnal in-lab polysomnography (PSG) is the gold standard to establish SAHS and its severity [1], [3]. It includes monitoring and recording multiple biomedical signals from patients (electroencephalogram, electrocardiogram, airflow, blood oxygen saturation, etc.) [3], which increases its complexity. PSG also requires an overnight stay of patients in a specialized sleep unit, outside their usual sleep environment, where clinicians attend them and ensure the proper functioning of the test. Therefore, the need for these dedicated facilities and human resources leads to increased costs [8], [9]. Once PSG is finished, SAHS is offline diagnosed by computing the apnea-hypopnea index (apneas and hypopneas per hour of sleep, AHI) [10]. Thus, all the biomedical signals recorded during the night need inspecting, which implies a significant time consumption.

Complexity, cost, and consumed time lead to a limited PSG availability, which is not able to cope with the high prevalence of SAHS [11]. This results in restricted access to diagnosis and treatment and, consequently, increased waiting lists [11]. In this regard, nocturnal pulse oximetry (NPO) has become a useful tool to overcome several PSG limitations. Single-channel blood oxygen saturation (SpO_2) from NPO measures the percentage of oxygen in the hemoglobin of blood, whose healthy value ranges between 96% and 100%. However, apneic events cause recurrent drops from these values (oxygen desaturations) [1], [3]. Moreover, SpO_2 can be easily acquired by using a single sensor placed in a finger. Hence, NPO is a simple, portable, and non-invasive test, which is widely used in clinical practice [12].

This research has been supported by the project VA037U16 from the Consejería de Educación de la Junta de Castilla y León, the project 265/2012 of the Sociedad Española de Neumología y Cirugía Torácica (SEPAR), the projects RTC-2015-3446-1 and TEC2014-53196-R from the Ministerio de Economía y Competitividad, and the European Regional Development Fund (FEDER). D. Álvarez was in receipt of a Juan de la Cierva grant from the Ministerio de Economía y Competitividad.

G. C. Gutiérrez-Tobal*, D. Álvarez, A. Crespo, F. del Campo, and R. Hornero, are with the Biomedical Engineering Group of the University of Valladolid, Spain (e-mail: gonzalo.gutierrez@gib.tel.uva.es). D. Álvarez, F. del Campo, and A. Crespo are with the sleep unit of Hospital Universitario Rio Hortega in Valladolid, Spain.

A number of studies have evaluated the SpO_2 signal acquired during in-lab PSG as diagnostic alternative. The 3% oxygen desaturation index (*ODI_3*), commonly used in clinical practice, as well as other univariate and multivariate automatic analyses, have been already tested [13]-[17], with the latter exhibiting higher performance [18], [19]. In spite of the promising results showed, none of these studies was focused on determining SAHS presence and its severity, nor was conducted involving SpO_2 recordings obtained at home. The simplification of the diagnostic test has as final goal to move it to patients' natural sleep environment, that is, their homes, while providing as accurate diagnostic information as possible. In this regard, some studies assessed univariate analyses applied to at-home SpO_2 recordings. Nevertheless, they were focused on *ODI_3* estimation and evaluation, and none of them conducted multivariate analyses [20]-[22]. Hence, there is still a need for further evaluation of machine-learning approaches applied to SpO_2 signals acquired under environmentally realistic conditions and focused on establishing both the presence and severity of SAHS.

In this study, we hypothesize that the SAHS diagnostic process may be simplified by the use of a machine-learning approach and the information contained in the at-home SpO_2 signal. Accordingly, our main goal is the assessment of machine-learning approaches with ability to automatically establish SAHS and its severity, with single-channel at-home SpO_2 as the only source of training information. Thus, first we propose a comprehensive characterization of SpO_2 by means of automatic extraction of spectral, non-linear, and statistical features already evaluated in previous in-lab studies, as well as *ODI_3*. As highlighted in preceding works, these features are expected to provide useful and complementary information about SAHS [13], [14], [18], [23]. However, such a comprehensive approach has not been already tested using at-home recordings and may lead to obtain features that offer similar information, that is, redundant features. A novel feature-selection step is included to avoid this issue. We propose a combination of the fast correlation-based filter (FCBF) and 'bootstrapping' to find an optimum set of features consistent through a variety of samples [24], [25]. The performance of the FCBF is independent of subsequent analyses or methodologies [25], which provides us with the additional advantage of conducting a fair evaluation of the optimum set of features regardless the different machine-learning algorithms adopted. Thus, we finally propose a comprehensive assessment of the at-home SpO_2 usefulness by training and validating up to four new machine-learning derived models, ranging from simple to complex ones: linear discriminant analysis (LDA), logistic regression (LR), multi-layer perceptron Bayesian neural network (BY-MLP), and the ensemble learning method adaptive boosting (AdaBoost), arranged along with LDA as base classifiers (AB-LDA). Preliminary studies of our own group have been already conducted regarding BY-MLP and AdaBoost, providing signs of the usefulness of the machine-learning approach for at-home SpO_2 recordings [26], [27]. Nonetheless, the comprehensive approach conducted in the current study has led to the use of different SpO_2 information to train the models, as well as LDA instead of classification and regression trees as base classifiers for AdaBoost. In addition, these works showed limitations such as lack of proper validation and the use of a BY-MLP model only trained for binary classification. By contrast, we follow a multinomial approach to find new models with ability to not only predict the presence of SAHS but also assign each subject under study into one of its four severity degrees (no SAHS, mild, moderate, and severe).

II. Subjects and signals

The study involved 320 adult subjects referred to the Hospital Universitario Rio Hortega in Valladolid (Spain) due to SAHS suspicion. All of them were diagnosed through an in-lab overnight PSG (E-series, Compumedics). A physician computed AHI following the rules of the American Academy of Sleep Medicine (AASM) [10], which was used as the gold standard. Participants with an AHI < 5 events per hour (e/h) were considered as no-SAHS subjects. Those showing an AHI in the ranges [5, 15) e/h, and [15, 30) e/h, were diagnosed as mild and moderate SAHS patients, respectively. Finally, subjects with AHI ≥ 30 e/h were diagnosed as severe. Each participant also conducted an at-home NPO. This was randomly carried out within the 24 hours before or after PSG to minimize the night-to-night variability effect [20]. Participants were divided into two sets: a training set composed of the first 60% consecutive subjects (n_{tr}=193, 19 no-SAHS, 31 mild, 35 moderate, 108 severe) and a test set composed of the remaining 40% (n_{test}=127, 10 no-SAHS, 24 mild, 21 moderate, 72 severe). All of them gave an informed consent. The Ethics Committee of the Hospital accepted the protocol (approval number: CEIC 7/13). Table I displays demographic and clinical data of the subjects (mean ± standard deviation). No statistically significant differences (*p*-value>0.01) were found in age, body mass index (BMI), or AHI.

SpO_2 signals were acquired during NPO (overnight length) by the use of a portable oximeter (Nonin WristOx2 3150, sampling rate 1 Hz). Artifacts due to movements were automatically removed during preprocessing. Thus, the SpO_2 values equal to zero, as well as the differences between consecutive SpO_2 samples ≥4%, were considered artifacts [15]. Figure 1 shows examples of SpO_2 recordings from each SAHS severity degree (no-SAHS, mild, moderate, and severe). Different patterns are observed, showing a tendency of higher amount of baseline falls (desaturations) in SpO_2 as SAHS severity increases. However, a comprehensive analysis is required to predict the class of each recording.

TABLE I
Demographic and clinical data of the subjects under study

	All	Training	Test
# Subjects	320	193	127
Age (years)	54.8 ± 13.5	54.2 ± 12.8	55.6 ± 14.4
Men (%)	74.1	76.7	70.0
BMI (kg/m^2)	29.2 ± 5.5	29.3 ± 5.4	29.1 ± 5.5
AHI (e/h)	39.2 ± 29.4	38.9 ± 28.7	39.6 ± 30.6

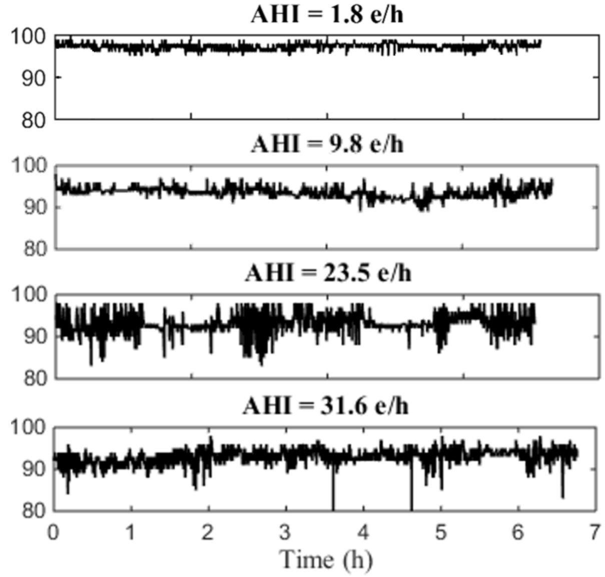


Fig. 1 Overnight SpO_2 (%) examples of no-SAHS (AHI < 5 e/h), mild SAHS (5 ≤ AHI < 15 e/h), moderate SAHS (15 ≤ AHI < 30 e/h), and severe SAHS (AHI ≥ 30 e/h) subjects.

III. Methodology

A three-stage methodology was conducted. First, statistical, spectral, and non-linear features were obtained from the SpO_2 recordings. These were used because of its reported usefulness in the in-lab SpO_2 evaluation. Additionally, 3% ODI was also computed due to its importance in clinical practice. Hence, 16 parameters composed the initial feature set. Then, an automatic selection stage was used to discard redundant features and obtain an optimum set of among them. The FCBF selection algorithm, along with a bootstrap technique, was used for this purpose. Finally, the optimum set of features fed four machine-learning approaches to obtain LDA, LR, BY-MLP, and AB-LDA models. Their performances were subsequently evaluated using a previously unseen test set.

A. Feature extraction

1) Common statistics: First-to-fourth order statistical moments were extracted from SpO_2 in time domain: mean (*Mt1*), standard deviation (*Mt2*), skewness (*Mt3*), and kurtosis (*Mt4*). These features characterize central tendency, dispersion, asymmetry, and peakedness of a given time series [13], [18], [19]. Previous in-lab studies reported statistically significant lower values of *Mt1*, *Mt3* and *Mt4*, as well as higher values of *Mt2*, in the SpO_2 signals from SAHS positive subjects [18].

2) Non-linear measures: Central tendency measure (*CTM*), Lempel-Ziv complexity (*LZC*), and sample entropy (*SampEn*) were also extracted from SpO_2 time series. These methodologies have previously shown its utility to characterize the stochastic components present in biomedical signals [23], [28]. Particularly, *CTM*, *LZC* and *SampEn* have shown promising results to respectively quantify the variability, complexity, and irregularity caused by SAHS in SpO_2 in-lab recordings [14], [23], [29]. Higher *LZC* and *SampEn* values [14], [29], as well as lower *CTM* values [23], have been reported in the SpO_2 of SAHS subjects.

3) Spectral analysis: The recurrence of the apneic events leads to analyze SpO_2 in the frequency domain too. The power spectral density (PSD) of the SpO_2 signals were estimated by the Welch's non-parametric periodogram [30]. Up to 8 spectral features were obtained from the PSDs of each SpO_2. Thus, first-to-fourth order statistical moments were also extracted from PSDs (*Mf1*-*Mf4*). As in time domain, these have been already used to characterize central tendency, dispersion, asymmetry, and peakedness in the PSDs of in-lab SpO_2 recordings [18], [19]. Higher *Mf1* and *Mf2*, as well as lower *Mf3* and *Mf4*, have been reported for SAHS subjects [18]. Two additional full-spectrum features were also obtained: median frequency (*MF*), and spectral entropy (*SpecEn*). *MF* is the frequency for which 50% of the spectral power is below it [23]. Consequently, upper values imply that the power is more concentrated in high frequencies. Significantly higher *MF* have been reported for SAHS patients in the case of SpO_2 recordings obtained in a laboratory [18]. *SpecEn* has been commonly used with biomedical signals to measure the flatness of the spectrum, with higher *SpecEn* values due to larger number of spectral peaks or higher dominance (or peakedness) of these [31]. Its past application to in-lab SpO_2 recordings showed statistically significant higher *SpecEn* values in SAHS patients [32]. Finally, two features were extracted from the spectral band of interest of the SpO_2 (0.014-0.033 Hz.) [17], [18]: peak amplitude (*PA*) and relative power (P_R). Both of them were found significantly higher in SAHS patients in previous in-lab studies [17], [18].

4) Oxygen desaturation index (ODI_3): ODI_3 counts the number of drops from the SpO_2 baseline greater than or equal to 3%, divided by the number of hours of recording. It is a clinical parameter widely used to help in SAHS diagnosis [15], [33]. Since desaturations are involved in the hypopnea definition [10], it is expected that higher values of ODI_3 be found in SAHS patients.

B. Feature selection: the fast correlation-based filter

An automated feature selection stage was implemented to avoid redundant information when training the machine learning models [25]. Our approach focused on the FCBF algorithm [24], which relies on symmetrical uncertainty (*SU*) as a normalization of the information gain (*IG*) between variables. FCBF is a filter method and, consequently, it is independent of the machine-learning algorithms later applied [25]. It is composed of two steps. First, a relevance analysis is carried out by ranking the 16 extracted features (F_i, i=1,2,…,16) according to the values of *SU* between each of them and a target variable Y, ($SU_{i,Y}$). *SU* is in the range 0-1, with *SU* = 0 indicating that the two variables are independent, and *SU* = 1 that knowing one it is possible to completely predict the other [24]. Hence, the higher the value of $SU_{i,Y}$, the more information shares a feature F_i with the target variable Y and the more relevant is that feature, i.e., the higher $SU_{i,Y}$ the higher rank of F_i. $SU_{i,Y}$ is computed as follows [24]:

$$SU_{i,Y}(F_i,Y) = 2\left[\frac{IG(F_i\,|\,Y)}{H(F_i)+H(Y)}\right], i = 1,2,\ldots 16, \qquad (1)$$

being *IG* (F_i | *Y*) = *H*(F_i) – *H*(F_i | *Y*), and *H* the well-known Shannon's entropy. *Y* has to be chosen in relationship with each specific problem. Therefore, in this study, it is a continuous variable composed of the AHI values of each subject (we only used the training group in the selection process). The second step is a redundancy analysis where *SU* between each pair of features, F_i, F_j, ($SU_{i,j}$, j=1,2,…16, $j \neq i$) is computed. The process starts from the most relevant features to the less relevant ones, that is, $SU_{i,j}$ is first computed between the most relevant feature and each of the remaining, which are also compared in the order of the ranking [24]. Then, if $SU_{i,j}$ (between two features) is higher or equal than $SU_{i,Y}$ (between the feature with the highest rank and the target variable) the corresponding less-ranked feature F_j is eliminated from the selection process due to redundancy with F_i. The optimum group of features is the set not discarded at the end of the process.

In order to find a more generalizable optimum set of features, FCBF was used along with a bootstrap procedure [34]. Thus, the feature values from the original training set were resampled with replacement and uniform probability to form *B*=1000 new training sets derived from the bootstrap process [35], [36]. Then, the FCBF algorithm was applied to each of them and, typically, different optimum sets of features were obtained. As might be expected, those features considered non-redundant more often than redundant formed the final optimum set, i. e., those selected more than 50% of the *B* times.

C. Multinomial classification

After feature selection, each subject under study (s_k, k=1,2,…K, K=320), was characterized by a pattern, $\mathbf{x}_k$, which is a vector whose components are the corresponding values of the optimum selected features. These patterns were used in a multinomial classification approach to predict SAHS severity, that is, to assign the subjects to one out of the four SAHS severity degrees: no-SAHS, mild, moderate, and severe. Four machine-learning models were obtained (LDA, LR, BY-MLP, and AB-LDA), which were trained and evaluated with the patterns from the training and test groups, respectively.

1) Linear discriminant analysis (LDA)

LDA is a pattern recognition technique that assigns a pattern $\mathbf{x}_k$ into one out of *l* classes, C_l. It relies on the assumption that the conditional class density function of each class, $p(\mathbf{x}_k \mid C_l)$, follows a multivariate normal distribution with identical covariance matrices, $\mathbf{\Sigma}$, for all the classes [37]. A discriminant score y_l is computed for each class using [38]:

$$y_l(\mathbf{x}_k) = \boldsymbol{\mu}_l^T \mathbf{\Sigma}^{-1}\mathbf{x}_k - \frac{1}{2}\boldsymbol{\mu}_l^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_l + \ln p(C_l), \qquad (2)$$

where $\boldsymbol{\mu}_l$ is the mean vector for class C_l and $p(C_l)$ its corresponding prior probability. Classification is conducted by assigning a pattern $\mathbf{x}_k$ to the class with the highest score y_l ($\mathbf{x}_k$).

2) Logistic regression (LR)

LR is a classic machine learning approach that computes the posterior probability of class membership for a given pattern, $\mathbf{x}_k$, by the use of the logistic function [39]:

$$p(C_l \mid \mathbf{x}_k) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{k1}+\ldots+\beta_M x_{kM})}}\ , \qquad (3)$$

where C_l represents all the possible classes, $\boldsymbol{\beta} = \beta_0, \beta_1, \ldots, \beta_M$ are the coefficients of the model corresponding to the interceptor and each feature, $\mathbf{x}_k = x_{k1},..., x_{kM}$, and *M* is the maximum number of independent variables (features) used. Coefficients $\boldsymbol{\beta}$ are estimated through the maximum likelihood method [39]. The model assigns a new pattern into the class with the highest posterior probability.

LR is a standard for classification methods that has been successfully tested in SAHS contexts involving in-lab SpO_2 features. However, its nature is essentially binary (l=1,2) [39]. Therefore, for our multiclass problem (l=1,2,3,4), it is evaluated following the well-known 1-vs.-all strategy.

3) Multi-layer perceptron neural network: the Bayesian approach (BY-MLP)

Artificial neural networks (ANN) are pattern recognition methods inspired by the human brain. MLP is an ANN whose architecture is arranged in several layers: input, hidden layers, and output [40]. Each layer is composed of computing units called perceptrons or neurons that are massively interconnected with units from other layers [40]. Particularly, each unit from one layer is connected with all the units from the following layer. The input layer was composed of one unit for each input feature. In our case, additionally, one single hidden layer has been used for simplicity, since it has been shown that this configuration can provide universal approximations [40]. Finally, as multinomial classification is intended, four units composed the output layer, which provided a posterior probability for each of the four classes.

The output units in our MLP architecture are represented by the next expression [40], [41]:

$$y_l(\mathbf{x},\mathbf{w}) = g_l\left(\sum_{n=1}^{N}\left(w_{nl}\, g_n\left(\sum_{m=1}^{M} w_{mn}\mathbf{x}_{ki} + b_n\right) + b_l\right)\right), \qquad (4)$$

where *M* is the number of features of the input pattern $\mathbf{x}_k$, *N* is the number of units in the hidden layer, w_{mn} is the weight that connects the feature *m* of the input pattern with the hidden unit *n*, b_n is the bias associated to hidden unit *n*, w_{nl} is the weight that connects the hidden unit *n* with the output unit *l*, b_l is the bias associated to output unit *l*, and $g_n(\cdot)$ and $g_l(\cdot)$ are the activation functions of the hidden and output units.

N is a tuning parameter to be optimized using the training set. Additionally, we have used the logistic expression for both $g_n(\cdot)$ and $g_l(\cdot)$, as a common choice for the activation functions in classification problems involving MLP [37]. Moreover, in the classic MLP approach, the weights and biases connecting the units in the different layers (w_{mn}, w_{nl}, b_l, b_n) are computed using the backpropagation algorithm during the training process, i.e., following a maximum likelihood optimization approach [40]. However, previous studies involving in-lab SpO_2 data showed higher performance of the Bayesian approach [41]. This alternative method models the posterior distribution of the whole set of weights and biases ($\mathbf{w}$), given a training set Tr_s, according to the Bayes' theorem:

$$p(\mathbf{w} \mid Tr_s) = \frac{p(Tr_s \mid \mathbf{w})\, p(\mathbf{w})}{p(Tr_s)}\ , \qquad (5)$$

where $p(Tr_s \mid \mathbf{w})$ is the likelihood of the training set, $p(\mathbf{w})$ is the prior probability function of the weights, and $p(Tr_s)$ is known as the evidence, which acts as a normalization factor [39]. The probability of membership of a pattern $\mathbf{x}_k$ to the class l can be obtained as follows [40]:

$$p(C_l \mid \mathbf{x}, Tr_s) = \int y_l(\mathbf{x}, \mathbf{w})\, p(\mathbf{w} \mid Tr_s)\, \mathrm{d}\mathbf{w}\,, \qquad (6)$$

which can be solved following the approximations and assumptions explained in the literature [36], [39]. Finally, the pattern $\mathbf{x}_k$ is assigned to the class with the highest probability.

4) AdaBoost ensemble learning (AB-LDA)

AdaBoost is an ensemble learning method that combines multiple base classifiers of the same type to complement each other [36]. This combination relies on the weighted votes of the single classifiers, which are usually simple ones to preserve the generalization ability of the method [36], [37]. Hence, LDA base classifiers were used along with AdaBoost in this study. AB-LDA has already proven to be helpful to detect SAHS severity when used together with airflow features obtained during in-lab PSG [42].

AdaBoost is an iterative process. At each p iteration, it assigns a weight, w_k^p, to every training pattern, $\mathbf{x}_k$. The p^{th} base classifier is trained using the corresponding weighted patterns. Then, the performance of the classifier is assessed by the error, ε_p. This error is subsequently used to determine the corresponding weighted vote, α_p, of the p^{th} classifier [36]. Those classifiers with smaller ε_p contribute more to the final decision (higher α_p). At the end of each iteration, the weights of the misclassified patterns are updated (w_k^{p+1}) [36]. Then, the weights of all patterns are normalized to maintain their original distribution [43]. By reweighting those patterns that have been misclassified during a particular iteration, the base classifiers trained during the next ones give them more importance, being more likely to be rightly classified [36], [43]. AdaBoost.M2 is the algorithm version for multinomial classification. In such a case, ε_p is defined as [43]:

$$\varepsilon_p = \frac{1}{2} \cdot \sum_{k=1}^{N_{training}} \sum_{l \neq l_{true}} w_{k,l}^{p} \cdot (1 - h_p(\mathbf{x}_k, l_{true}) + h_p(\mathbf{x}_k, l)), \qquad (7)$$

where l is a categorical variable representing the multiple classes, l_{true} is the actual class of $\mathbf{x}_\mathrm{k}$, and h_p is the confidence of the prediction of the base learner for a pattern $\mathbf{x}_\mathrm{k}$ and a given class. The final classification task is conducted by returning the class l with the highest sum of the votes from all classifiers, taking into account the weight of their corresponding predictions α_p as follows [43]:

$$\alpha_p = \ln(\beta_p), \qquad (8)$$

where β_p is defined as (1 - ε_p)/ ε_p [43]. Additionally, the shrinkage regularization technique has been proposed to minimize overfitting [44]. It is based on adding a learning rate υ to the iterative process by redefining β_p as $(\beta_p)^\upsilon$, where υ ranges 0-1 and has to be experimentally chosen. The number of base classifiers (Q) to be used is another parameter to be experimentally chosen in the AdaBoost.M2 algorithm.

D. Statistical analysis

As features did not pass the Lilliefors normality test, the Kruskal-Wallis non-parametric method was used to evaluate differences among SAHS severity groups (p-value<0.01 to minimize the chances of type I errors). The overall performances of the proposed machine-learning methodologies in the multinomial classification task were assessed by the use of Cohen's kappa, κ, since it is able to measure the agreement between the actual SAHS severity levels and the predicted ones while avoiding the effect of agreement due to chance [36]. Additionally, the diagnostic performance for each of the AHI cutoffs that define the SAHS severity degrees where evaluated in a binary fashion in terms of sensitivity (Se, percentage of subjects above the cutoff rightly classified), specificity (Sp, percentage of subjects below the cutoff rightly classified), and accuracy (Acc, percentage of all subjects rightly classified).

As mentioned above, a bootstrap procedure was included in the feature selection stage for the sake of the generalization ability of the optimum set of features chosen. Moreover, the bootstrap 0.632 algorithm was also used in a similar way to optimize the tuning parameters of BY-MLP and AB-LDA methods. Thus, B = 1000 training groups are formed by resampling with replacement from the original training group, following a random uniform probability. Consequently, repeated patterns from $\mathbf{x}_k$ are most likely to be included in each new group; in the same number, several patterns are not used. While the former compose the bootstrap training groups, the latter form a bootstrap test group for each of them. To select the optimum parameter values (hidden neurons (N) in BY-MLP; learning rate (υ) and number of base classifiers in AB-LDA (Q)), a range of these were evaluated. It is known that using only results from the bootstrap test groups would lead to pessimistic estimations [36]. Therefore, Cohen's κ was computed for each model configuration as follows [36]:

$$\kappa = 0.368 \cdot \kappa_{Btraining} + 0.632 \cdot \kappa_{Btest} \qquad (9)$$

where $\kappa_{Btraining}$ and κ_{Btest} are Cohen's κ of each new bootstrap training and bootstrap test groups, both derived from the original training group. The final parameters were chosen according to the highest κ averaged over the 1000 groups.

IV. Results

A. Single features separability

Figure 2 displays the violin plots from all the extracted features in the training set split in the four SAHS-severity degrees. As in the case of boxplots, 25 percentile, 50 percentile (median), and 75 percentile are showed (horizontal gray lines). In addition, the data distribution (histogram) for each feature is also showed as the lateral outlines of each box, which are vertically symmetrical. The p-values after Bonferroni correction were also included. Only one feature (MF) did not reach statistically significant differences (p-value<0.01) among SAHS degrees. Consequently, all the proposed approaches (statistical, spectral, non-linear, and clinical) contributed with features that reached significant statistical differences, i.e.,

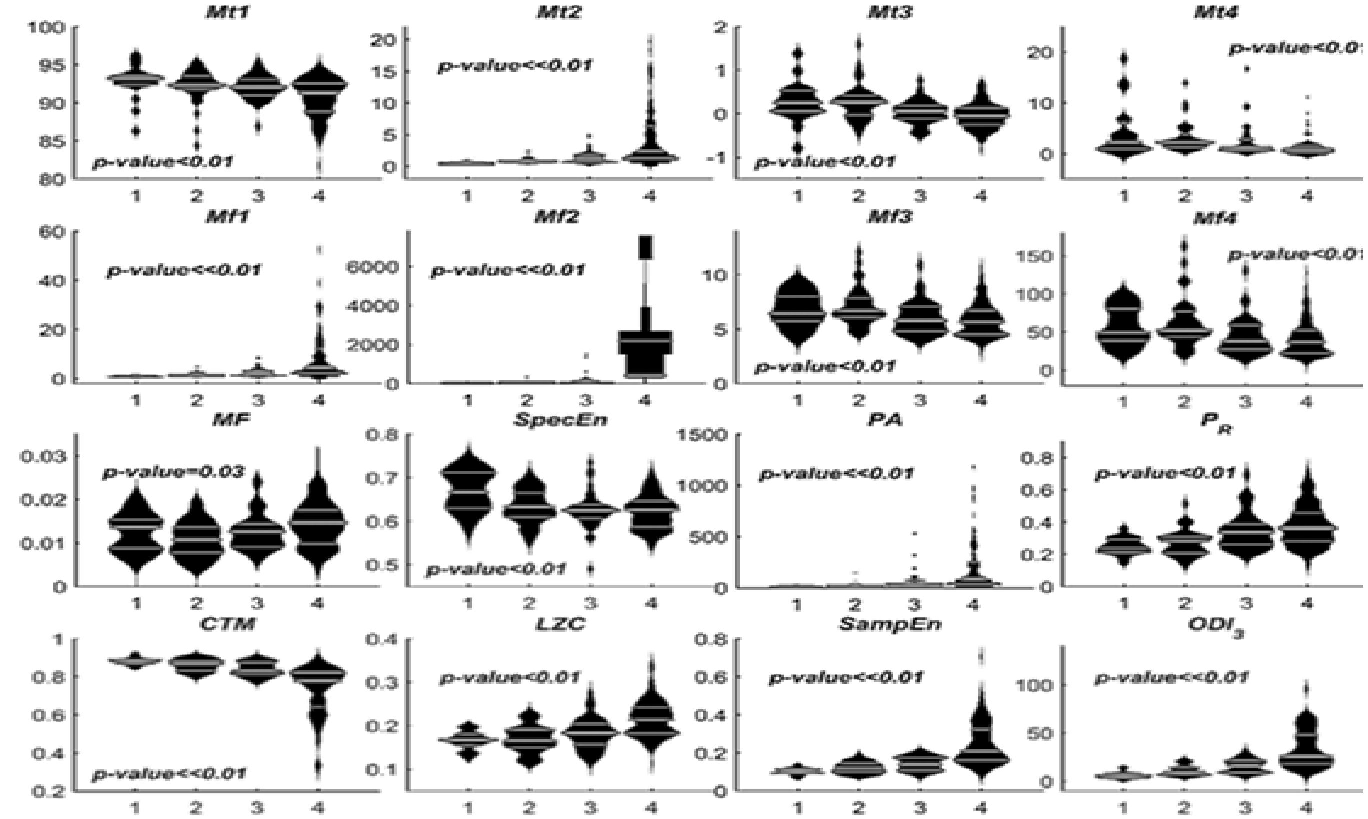


Fig. 2 Violin plots of each extracted feature divided by SAHS-severity degree (only training set). Numbers in x-axis represent the severity of SAHS: 1 stands for no-SAHS, 2 for mild, 3 for moderate, and 4 for severe. All *p*-values from Kruskal-Wallis test were corrected using the Bonferroni criterion.

showed separability among groups. Additionally, clear tendencies over the severity groups can be observed in those features with *p*-values$<10^{-15}$ (denoted as *p*-value << 0.01). Thus, *Mt2*, *Mf1*, *Mf2*, *PA*, *SampEn*, and ODI_3 showed higher figures as severity increased, whereas *CTM* was higher as severity decreased. However, similar distributions are present in these features in all severity groups, including *CTM*, which only differs in the direction of its tendency. By contrast, *Mt1*, *Mt3*, *Mt4*, *Mf3*, *Mf4*, *SpecEn*, P_R, and *LZC* showed higher *p*-values but different data distributions among and within groups.

B. Optimum set of features

Fig. 3 displays the histogram of features selected by the FCBF algorithm over the *B* = 1000 resampled bootstrap training sets. Only ODI_3 and *SpecEn* were selected more than half of the times, i.e., were considered non-redundant more often than redundant. ODI_3 and *SpecEn*, therefore, formed the optimum set of features used to train the machine-learning models. By contrast, *Mt1*, *Mt2*, *Mf1*, *Mf2*, *PA*, *PR*, and *LZC* were redundant all the times. Additionally, *MF* were only selected 0.5% of the times. Fig. 4 shows a scatter plot facing ODI_3 and *SpecEn* by

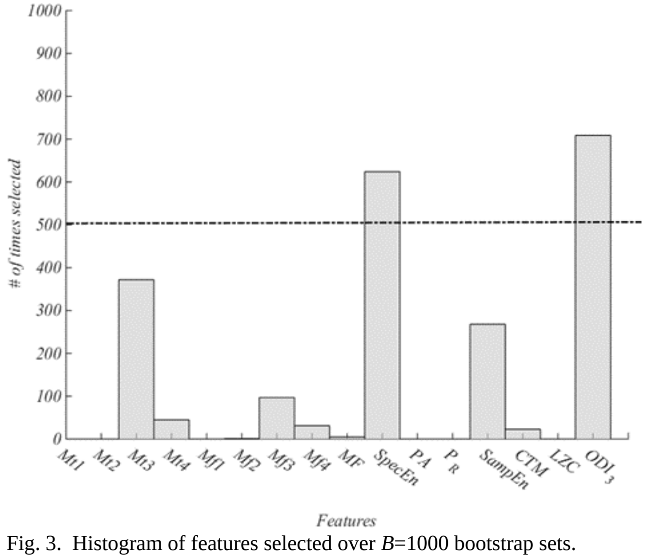


Fig. 3. Histogram of features selected over *B*=1000 bootstrap sets.

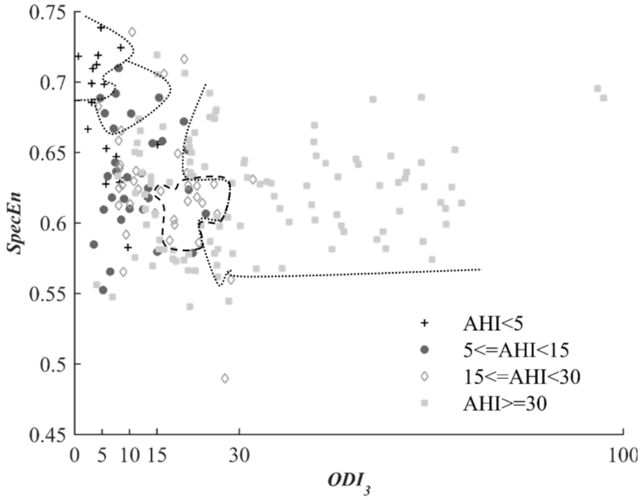
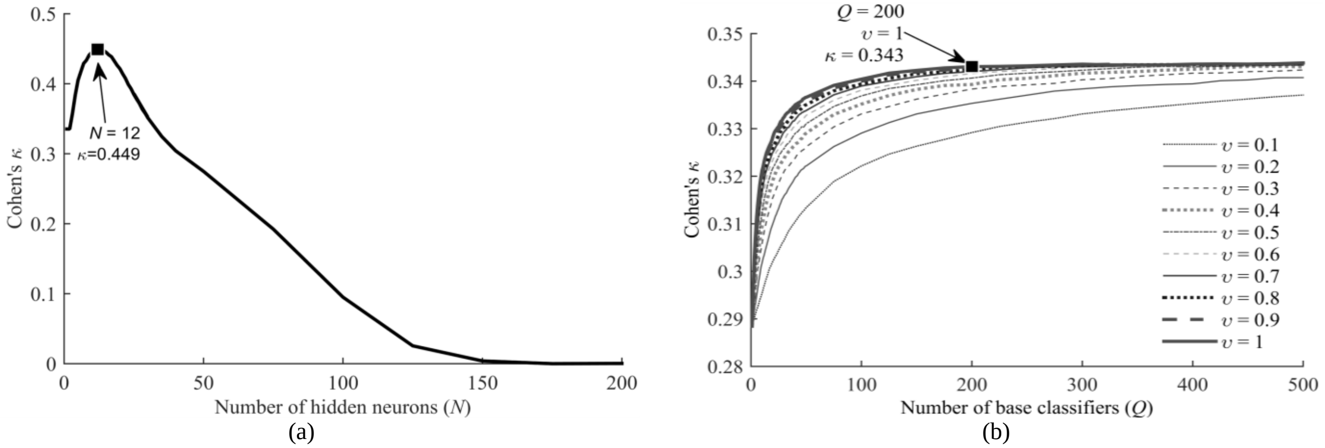


Fig. 4. Scatter plot facing ODI_3 and *SpecEn* for the four SAHS severity degrees. Dashed lines show examples of regions in which each of the classes prevails.



(a) (b)

Fig. 5. Averaged Cohen's κ for different (a) number of hidden neurons, N, (BY-MLP) and (b) number of base classifiers, Q, and learning rate, υ, (AB-LDA).

TABLE II. Confusion matrices for the multinomial machine learning classifiers and ODI_3 in the test set. ***1***: No SAHS (AHI < 5 e/h); ***2***: Mild SAHS (5 ≤ AHI < 15 e/h); ***3***: Moderate SAHS (15 ≤ AHI < 30 e/h); ***4***: Severe SAHS (AHI≥30 e/h)

Estimated severity →		***ODI₃***				**LDA**				**LR (1 vs. all)**				**BY-MLP**				**AB-LDA**			
		1	***2***	***3***	***4***	***1***	***2***	***3***	***4***	***1***	***2***	***3***	***4***	***1***	***2***	***3***	***4***	***1***	***2***	***3***	***4***
Actual severity	***1***	3	7	0	0	2	8	0	0	8	2	0	0	8	2	0	0	5	5	0	0
	2	2	17	5	0	1	12	2	9	5	9	0	10	4	9	1	10	1	14	2	7
	3	1	9	10	1	0	5	0	16	2	2	0	17	2	1	4	14	1	2	6	12
	4	0	10	23	39	0	4	0	68	2	2	0	68	2	2	9	59	2	2	4	64

TABLE III. Binary diagnostic ability of the machine-learning classifiers and ODI_3 in the test set for the AHI cutoffs = 5 e/h, 15 e/h, and 30 e/h.

	ODI₃			**LDA**			**LR (1 *vs.* all)**			**BY-MLP**			**AB-LDA**		
	5	15	30	5	15	30	5	15	30	5	15	30	5	15	30
Se (%)	97.4	78.5	54.2	99.1	90.3	94.4	92.3	91.4	94.4	93.2	92.5	81.9	96.6	92.5	88.9
Sp (%)	30.0	85.3	98.2	20.0	67.6	54.5	80.0	70.6	50.9	80.0	67.6	56.4	50.0	73.5	65.5
Acc (%)	92.1	80.3	73.2	**92.9**	84.3	77.2	91.3	85.8	75.6	92.1	85.8	70.9	**92.9**	**87.4**	**78.7**
κ		0.351			0.341			0.468			0.363			**0.479**	

SAHS-severity degree. No simple boundaries are observed. However, it can be appreciated examples of regions in which each of the classess prevails.

C. Classification models training

BY-MLP and AB-LDA models needed optimum tuning parameters to be chosen (number of neurons in the hidden layer, N; and number of base classifiers, Q, and learning rate, υ, respectively). The bootstrap 0.632 method, applied to the ODI_3 and *SpecEn* values from the original training set, was used for this purpose. Fig. 5 (a) and (b) show Cohen's κ according to equation (9), and averaged over the 1000 derived bootstrap training and test sets. The optimum value for N (BY-MLP) was 12, whereas optimum Q and υ (AB-LDA) were 200 and 1, respectively. Q = 200 was chosen for the sake of model simplicity since the third decimal place did not change onwards. After the optimization of these tuning parameters, the ODI_3 and *SpecEn* values from the whole original training set were used to obtain the specific new LDA, LR, BY-MLP, and AB-LDA models, whose diagnostic performance were evaluated using the test set only.

D. Diagnostic performance

Table II shows the confusion matrices for the LDA, LR, BY-MLP, and AB-LDA models and ODI_3, evaluated in the test set. Using ODI_3, 69 subjects (54.3 %) were rightly assigned to their actual SAHS-severity group (main diagonal of the matrix). By contrast, 82 (64.6%), 85 (66.9%), 80 (63.0%), and 89 (70.1%) subjects were rightly classified by the LDA, LR, BY-MLP, and AB-LDA models, respectively. Hence, all of them outperformed single ODI_3 when considering the exact number of subjects rightly classified. ODI_3 showed clear overestimation of severity in no-SAHS group (7 out of 10 subjects) as well as underestimation of SAHS degree in severe patients (33/72 subjects). Interestingly, the LR, BY-MLP, and AB-LDA models significantly decreased these two undesirable effects. However, LR performed poorly when predicting moderate SAHS.

A complementary analysis can be derived from Table III. It displays Cohen's κ to show the overall performance of ODI_3 and the machine-learning models in the four-class classification. It also shows Se (%), Sp (%), and Acc (%) values obtained from confusion matrices when conducting a binary evaluation of the 3 clinically useful cutoffs that define the SAHS-severity groups (5 e/h, 15 e/h, and 30 e/h). AB-LDA reached the highest κ (0.479), as well as the highest Acc values for all AHI cutoffs: 92.9%, 87.4%, and 78.7%, respectively. Moreover, all machine-learning models but LDA outperformed ODI_3 in terms of κ. LDA, however, reached higher binary Acc than ODI_3 in the three AHI cutoffs. Additionally, ODI_3 showed higher Sp (%) than the models for the AHI cutoffs 15 e/h and 30 e/h.

V. Discussion

We have evaluated four new machine-learning models trained to automatically stablish both SAHS presence and its severity by the only use of non-redundant SpO_2 information obtained at patient's home. All of them outperformed the commonly used clinical index ODI_3 when assessing the total number of subjects rightly classified into one out of the four SAHS-severity degrees. Additionally, our AB-LDA proposal achieved the highest diagnostic ability in the overall performance (70.1% accuracy, 0.479 κ), as well as when conducting a binary evaluation of the AHI cutoffs 5 e/h (92.9% Acc), 15 e/h (87.4% Acc), and 30 e/h (78.7% Acc).

According to Fig. 2, all the extracted features but *MF* reached statistically significant differences among SAHS severity. These features were initially chosen due to the usefulness reported in previous in-lab studies [13], [14], [18], [23]. It is not surprising, therefore, that the vast majority of them showed statistically significant separability over the four classes, while highlighting the convenience of the three analytical approaches adopted (statistical, spectral, and non-linear). However, contrary to previous studies mainly focused on binary evaluations, a tendency can be observed over the severity groups in some features. Particularly, those showing p<<0.01. Thus, *Mt2*, *Mf1*, *Mf2*, *PA*, *SampEn*, and ODI_3 showed increasingly higher values with SAHS severity. By contrast, *CTM* reduced its values as SAHS severity got worse. In order to explain these behaviors notice that SpO_2 is ideally constant over time around 96%-100% saturation. Deviations are due to non-desirable drops in blood saturation, which are commonly present in SAHS [10]. In this regard, *Mt2* (standard deviation in time domain), *SampEn*, and *CTM* would be respectively characterizing the increasing degree of dispersion, irregularity, and variability of the SpO_2 data as more apneic-related desaturations are present. On the other hand, desaturations, as amplitude variations, increase the total spectral power of the SpO_2 signal, whereas a high degree in its recurrence may imply that this increased spectral power affects more discrete frequencies. Consequently, *Mf1* (mean of the PSD) and *Mf2* (standard deviation of PSD) would be measuring higher spectral power, and higher amount of frequencies affected by it, as more apneic events were present. In addition, the spectral band of interest of the SpO_2 signal (0.014-0.033 Hz.) has been defined as the range of frequencies in which recurrence of desaturations are more likely to happen [17], [18]. Therefore, the more apneic-related desaturations, the higher the value of the PSD in that band, i.e., the higher the *PA* feature. Finally, desaturations are involved in hypopnea definition [10]. Hence, increasing of ODI_3 is a natural tendency as more of these events happen.

The bootstrap FCBF method showed that, over 1000 repetitions of the algorithm, most of the features were found redundant more often than non-redundant. Hence, only ODI_3 and *SpecEn* were selected. This criterion to choose the final optimum set may be relaxed by including features with a selection rate below 50%. However, 8 out of the 16 features were selected less than 0.5% of times and 12 were less than 10% of times. Therefore, regardless the selection rate chosen, it has been shown that a high degree of redundancy in SAHS information is present in the features usually extracted from SpO_2 recordings.

As previously mentioned, those features considered non-redundant using a selection rate of 50% (ODI_3 and *SpecEn*) fed four new machine-learning models that outperformed the diagnostic ability of single ODI_3 when determining SAHS severity. This highlights the utility of the joint use of these two selected features. Violin plots from ODI_3 and *SpecEn* (Fig. 2), as well as its scatter plot (Fig. 4), are also coherent with this idea. While ODI_3 violin plot showed low class separability between class 1 (no-SAHS) and class 2 (mild SAHS), clear differences could be found for these classes in *SpecEn*. Similarly, the ODI_3 *vs*. *SpecEn* scatter plot showed clear regions in which class 1 and class 4 (severe–SAHS) prevailed, which are the classes where the machine-learning approaches most improved single ODI_3 classification. ODI_3, indeed, underestimated SAHS degree in severe patients (Table II). Therefore, a diagnostic test only supported by ODI_3 would not be effective in a significant proportion of subjects (46%) that would benefit more from a quick diagnosis and access to treatment. Furthermore, single ODI_3 also overestimated SAHS degree of 70% of no-SAHS subjects. By contrast, our machine-learning models were able to minimize both undesirable effects. The AB-LDA model reduced the ODI_3 overestimation to 50% of no-SAHS subjects, and only underestimated SAHS in 11.1% of severe ones.

Table IV summarizes previous studies aimed at automatically diagnosing SAHS by using SpO_2 signals acquired at patients' home. Olson et al. used a large sample (793 subjects) to conduct a direct validation of several univariate clinical indexes derived from at-home NPO as a surrogate for AHI [20]. The highest performance was reached using delta index, which achieved only moderate diagnostic ability. Chung et al. also conducted a direct evaluation of ODI_3 obtained at home, using a sample size of 475 subjects [21]. However, these were surgical patients rather than common SAHS suspects. They reported 87.0%,

TABLE IV. Comparison with state-of-the-art studies focused on at-home automatic detection of SAHS by the use of the SpO_2 signal.

Studies	#of subjects	Method	Valid.	AHI (e/h)	Se (%)	Sp (%)	Acc (%)
Olson *et al*. [20]	793	Delta index (univariate)	*vs*[a]	5	82.7	54.2	*nd*[b]
				15	88.5	39.6	67.1*
				30	92.6	34.1	*nd*
Chung et al. [21]	475	ODI_3 (univariate)	*vs*	5	96.3	67.3	87.0
				15	70.0	92.5	84.0
				30	76.0	97.2	93.7
Schlotthauer *et al.* [22]	996	ODI_3 estimation (univariate)	bootstrap+ hold-out	15	83.8	85.5	*nd*
This study	320	ODI_3 (univariate)	*vs*	5	97.4	30.0	92.1
				15	78.5	85.3	80.3
				30	54.2	98.2	73.2
		AB-LDA	bootstrap+ bootstrap+ hold-out	5	96.6	50.0	92.9
				15	92.5	73.5	87.4
				30	88.9	65.5	78.7

[a] *vs: validation study, direct comparison of a metric and the gold standard;* [b] *nd: not enough data to estimate;* **Estimated from reported data.*

84.0%, and 93.7% Acc for AHI thresholds of 5, 15, and 30 e/h, respectively. Schlotthauer et al. used empirical mode decomposition (EMD) to carry out an automatic estimation of ODI_3 in 996 SpO_2 recordings obtained during at-home polysomnography [22]. They used an initial set (40 recordings) to optimize EMD, and a validation set (669 recordings with 100 bootstrap repetitions) to conduct a receiver-operating characteristics analysis. An unseen test set was finally used to estimate the diagnostic performance of the proposal, reaching 83.8% Se and 85.5% Sp (AHI cutoff = 15 e/h). Our machine-learning approach is fully validated using a first bootstrap stage for feature selection, a second bootstrap stage for model design, the whole training set for model training, and a previously unseen test set for diagnostic ability estimation. Under these conditions, our new AB-LDA model outperformed the overall diagnostic ability of all the works of the state of the art. ODI_3 from the study of Chung et al. reported the closest diagnostic ability to the machine-learning proposal. Nonetheless, patients involved in their study were not regular but surgical ones. In addition, our ODI_3 computed in the same database showed evident lower diagnostic ability.

According to the results in the confusion matrices (Table II), several screening procedures could be derived from our models to show its clinical usefulness. As an illustrative example using AB-LDA, regard this protocol: *i*) if our model predicts class 3 (moderate SAHS) or class 4 (severe SAHS) a clinician could consider the application of treatment, since 100% of subjects predicted as moderate or severe (95 out of 95) have mild SAHS at least; *ii*) if our model predicts class 1 (no SAHS) or class 2 (mild SAHS), by following a conservative strategy a clinician could directly send the patients to undergo conventional PSG. A less conservative approach could consider PSG only at the persistence of symptoms since, most probably, the subjects predicted as class 1 or class 2 are no SAHS or mild SAHS (25 out of 32). It would maximize avoided PSGs while still taking into account the patients that initially missed the treatment, especially the 7.5% that is moderate or severe (7 out of 93). In either of these two options, this protocol would avoid a minimum of 74.8% full PSGs (95 out of 127). Comparing to the use of ODI_3 alone, only a minimum of 61.4% (78 out of 127) would be avoided, while potentially missing 21.5% of moderate and severe patients (20 out of 93).

Several limitations need to be addressed regarding our work. First, although our sample size is larger than most of in-lab studies [14-16], [18], [19], [45], more subjects would enhance the statistical generalization of our results as well as would equal the sample size of the few at-home studies found. A second limitation concerns to a historical lack of consensus in the AHI threshold to determine SAHS and its severity [46]-[48]. The criterion has changed over the last years, which makes the comparison with past works difficult. In this regard, our methodology is flexible, since it can be evaluated both in the four-class classification task as well as in three out of the most common AHI thresholds historically used in a binary way (5 e/h, 15 e/h, and 30 e/h). However, it has to be taken into account that AHI itself present several drawbacks to accurately predict the true health state of SAHS patients and, despite it is accepted as the main one among the current diagnostic options [47], it is still suboptimal for accurately establishing relationships with pathological conditions associated with SAHS [47]. In this regard, recent studies have proposed oximetry features as clinically helpful to gain insight into the severity of each patient [48]. Another constrain of our study is associated with the use of the SpO_2 signal. Despite its common assessment as alternative to PSG, according to the AASM it is possible that some apneic events do not cause a specific response in SpO_2 [10]. This could be a reason why both our machine-learning models as well as ODI_3 are not as accurate as PSG in SAHS diagnosis, and why its use should be recommended as screening tool. In addition, central apneas have not been differentiated from obstructive ones in this study, which could be an important future goal. Not using clinically oximetry-based parameters such as time under 90% of saturation (CT90) could be another limitation. However, previous at-home studies reported very low diagnostic ability of CT90 [20], and preliminary tests in this study showed that it was highly redundant with the features used. Moreover, regarding the high redundancy showed in the oximetry-based features, the use of other analyses could be useful to find helpful data. Thus, recent studies have pointed to cardiac information as promising complementary information [49], [50]. Therefore, a future objective could focus on evaluating the use of non-redundant SpO_2 features along with oximetric-based cardiac information. Regarding the selection of the features, a new limitation arises. Other feature selection or dimensionality reduction methods might obtain other data as the optimum choices. Similarly, the use of other machine-learning models, or the combination of those described in this study, might increase the final classification task, which could be another interesting future goal. Finally, the AHI of each subject under study was obtained from in-lab PSG, which is the gold standard for SAHS diagnosis. However, the SpO_2 signals were recorded at-patients' home in a different night. Such a protocol may be affected by the night-to-night variability effect [20]. In order to minimize it, all the SpO_2 recordings were acquired within the 24 hours before or after PSG (randomly assigned).

VI. Conclusions

In summary, we have shown and explained SAHS-severity related tendencies in statistical, spectral, non-linear, and clinical features extracted from SpO_2 recordings obtained at patients' home. Moreover, we have exposed excessive redundancy with ODI_3 in the information typically extracted from SpO_2 in the context of SAHS. Accordingly, we have identified *SpecEn* as a robust non-redundant complement for ODI_3. Our new machine-learning AB-LDA model, rigorously validated, reached the highest diagnostic ability comparing with the works of the state of the art. It can be applied to both multiclass and binary classifications using different AHI thresholds, which highlights its potential as a SAHS screening tool. These results suggest that the machine-learning approach can be used along with SpO_2 information acquired at patients' home to help in SAHS diagnosis simplification.

References

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

[2] P. E. Peppard et al, "Increased prevalence of sleep-disordered breathing in adults. *Am J Epidemiol*," vol. 177(9), pp. 1006-14, 2013.

[3] S. P. Patil, et al, "Adult Obstructive Apnea," *Chest*, vol. 132, pp. 325-37, 2007.

[4] C. J. Egea and F. del Campo, "Work-related accidents, absenteeism and productivity in patients with sleep apnea. A future consideration in occupational health assessments?," *Arch Bronconeumol*, vol. 51(5), 2015.

[5] A. Sassani, et al, "Reducing Motor-Vehicle Colissions, Costs, and Fatalities by Treating Obstructive Sleep Apnea Syndrome," *Sleep*, vol. 27, pp. 453-458, 2003.

[6] 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.

[7] J. Durán et al, "Obstructive sleep apnea–hypopnea and related clinical features in a population-based sample of subjects aged 30 to 70 yr," *Am. J. Respir. Crit. Care Med.*, vol. 163(3), pp. 685-89, 2001.

[8] J. A. Bennet and W. J. M. Kinnear WJM, "Sleep on the cheap: the role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome," *Thorax*, vol. 54, pp. 958-59, 1999.

[9] W. W. Flemons et al, "Home Diagnosis of Sleep Apnea: A Systematic Review of the Literature," *Chest*, vol. 124, pp. 1543-79, 2003.

[10] 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.

[11] W. W. Flemons et al, "Access to diagnosis and treatment of patients with suspected sleep apnea," *Am J Respir Crit Care Med*, vol. 169, pp. 668-72, 2004.

[12] K. E. Bloch, "Getting the most out of nocturnal pulse oximetry," *Chest*, vol. 124(5), pp. 1628-1630, 2003.

[13] D. Álvarez 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.

[14] R. Hornero et al, "Utility of approximate entropy from overnight pulse oximetry data in the diagnosis of the obstructive sleep apnea syndrome," *IEEE Trans Biomed Eng*, vol. 54(1), pp. 107-13, 2007.

[15] U. J. Magalang et al, "Prediction of the apnea-hypopnea index from overnight pulse oximetry," *Chest*, vol. 124(5), pp. 1694-1701, 2003.

[16] D. Sánchez-Morillo et al, "Novel multiclass classification for home-based diagnosis of sleep apnea hypopnea syndrome," *Expert Systems with Applications*, vol. 41(4), pp. 1654-62, 2014.

[17] C. Zamarrón et al, "Oximetry spectral analysis in the diagnosis of obstructive sleep apnoea," *Clin Sci*, vol. 97(4), pp. 467-73, 1999.

[18] D. Álvarez et al, "Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis," *IEEE Trans Biomed Eng*, vol. 57, no. 12, pp. 2816–24, 2010.

[19] J. V. Marcos et al, "Automated prediction of the apnea-hypopnea index from nocturnal oximetry recordings," *IEEE Trans Biomed Eng*, vol. 59(1), pp. 141-49, 2012.

[20] L. G. Olson et al, "Prediction of sleep-disordered breathing by unattended overnight oximetry. *J Sleep Res*, vol. 8(1), pp. 51-55, 1999.

[21] F. Chung et al, "Oxygen desaturation index from nocturnal oximetry: a sensitive and specific tool to detect sleep-disordered breathing in surgical patients," *Anesthesia & Analgesia*, vol. 114(5), pp. 993-1000, 2012.

[22] G. Schlotthauer et al, "Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry," *Med Eng Phys*, vol. 36(8), pp. 1074-1080, 2014.

[23] D. Álvarez et al, "Improving diagnostic ability of blood oxygen saturation from overnight pulse oximetry in obstructive sleep apnea detection by means of central tendency measure," *Artif Intell Med*, vol. 41(1), pp. 13-24, 2007.

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

[25] Y. Saeys et al, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23(19), pp. 2507-2517, 2007.

[26] D. Álvarez et al, "Automated analysis of unattended portable oximetry by means of Bayesian neural networks to assist in the diagnosis of sleep apnea," in *Proc*. GMEPE/PAHCE, Madrid, Spain, 2016, pp. 1-4.

[27] G. C. Gutiérrez-Tobal et al, "Multi-class adaboost to detect Sleep Apnea-Hypopnea Syndrome severity from oximetry recordings obtained at home," in *Proc*. GMEPE/PAHCE, Madrid, Spain, 2016, pp. 1-5.

[28] M. Costa et al, "Multiscale entropy analysis of biological signals," *Physical review E*, vol. 71(2), pp. 021906, 2005.

[29] D. Álvarez et al, "Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection," *Phys Meas*, vol. 27(4), pp. 399.

[30] P. D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on time Averaging Over Short, Modified Periodograms," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-15, pp. 70-73, 1967.

[31] T. Inouye et al, "Quantification of EEG irregularity by use of the entropy of the power spectrum," *Electroencephalography and clinical neurophysiology*, vol. 79(3), pp. 204-210, 1991.

[32] A. Garde et al, "Development of a screening tool for sleep disordered breathing in children using the phone Oximeter™," PloS one, vol. 9(11), pp. e112959, 2014.

[33] N. Netzer et al, "Overnight pulse oximetry for sleep-disordered breathing in adults: a review," *Chest*, vol. 120(2), pp. 625-633, 2001.

[34] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J Mach Learn Res*, vol. 3, pp. 1157-1182, 2003.

[35] B. Efron and R. J. Tibshirani, An introduction to the bootstrap. CRC press, 1994.

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

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

[38] J. V. Marcos et al, "Automated detection of obstructive sleep apnea syndrome from oxygen saturation recordings using linear discriminant analysis," Med. Eng. Phys., vol. 59, pp. 141-49, 2010.

[39] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression. New York, NY: John Wiley and Sons, 2000.

[40] C. M. Bishop, Neural networks for pattern recognition. Oxford university press, 1996.

[41] J. V. Marcos et al, "The classification of oximetry signals using Bayesian neural networks to assist in the detection of obstructive sleep apnoea syndrome," *Phys Meas*, vol. 31(3), pp. 375, 2010.

[42] G. C. Gutiérrez-Tobal et al, "Utility of AdaBoost to Detect Sleep Apnea-Hypopnea Syndrome From Single-Channel Airflow," *IEEE Trans Biomed Eng*, vol. 63(3), pp. 636-646, 2016.

[43] Y. Freund, and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", J. Comput. System Sci., vol. 55(1), pp. 119-139, 1997.

[44] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, 1189-1232, 2001.

[45] H. M. Al-Angari and A. V. Sahakian, "Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier," *IEEE Trans Inf Technol Biomed*, vol. 16, pp. 463-468, 2007.

[46] S. M. Caples et al, "Obstructive sleep apnea," *Annals of internal medicine*, vol. 142(3), pp. 187-197, 2005.

[47] T. Penzel et al, "Revise respiratory event criteria or revise severity thresholds for sleep apnea definition?," *Journal of clinical sleep medicine*, vol. 11(12), pp. 1357, 2015.

[48] 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(6), pp. 697-708, 2013.

[49] J. R. Williamson et al, "Individualized apnea prediction in preterm infants using cardio-respiratory and movement signals," *In Proceedings of the IEEE International Conference on Body Sensor Networks* (BSN), pp. 1-6), May 2013.

[50] G. C. Gutiérrez-Tobal et al, "Assessment of Time and Frequency Domain Entropies to Detect Sleep Apnoea in Heart Rate Variability Recordings from Men and Women," *Entropy*, vol. 17, pp. 123-141, 2015.