Applied Soft Computing 111 (2021) 107827

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Ensemble-learning regression to estimate sleep apnea severity using at-home oximetry in adults

Gonzalo C. Gutiérrez-Tobal[a,b,*], Daniel Álvarez[a,b,c], Fernando Vaquerizo-Villar[a], Andrea Crespo[a,c], Leila Kheirandish-Gozal[d], David Gozal[d], Félix del Campo[a,b,c], Roberto Hornero[a,b]

[a] *Biomedical Engineering Group, Universidad de Valladolid, Valladolid, Spain*
[b] *Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, (CIBER-BBN), Madrid, Spain*
[c] *Pneumology Service, Río Hortega University Hospital, Valladolid, Spain*
[d] *Department of Child Health, and the Child Health Research Institute, The University of Missouri School of Medicine, Columbia, MO, USA*

ARTICLE INFO

Article history:
Received 1 November 2020
Received in revised form 15 July 2021
Accepted 12 August 2021
Available online 25 August 2021

Keywords:
Sleep apnea
Ensemble learning
Oximetry
Decision curve analysis
Sleep heart health study

ABSTRACT

Overnight pulse oximetry has shown usefulness to simplify obstructive sleep apnea (OSA) diagnosis when combined with machine-learning approaches. However, the development and evaluation of a single model with ability to reach high diagnostic performance in both community-based non-referral and clinical referral cohorts are still pending. Since ensemble-learning algorithms are known for their generalization ability, we propose a least-squares boosting (LSBoost) model aimed at estimating the apnea–hypopnea index (AHI), as the correlate clinical measure of disease severity. A thorough characterization of 8,762 nocturnal blood-oxygen saturation signals (SpO_2) obtained at home was conducted to extract the oximetric information subsequently used in the training, validation, and test stages. The estimated AHI derived from our model achieved high diagnostic ability in both referral and non-referral cohorts reaching intra-class correlation coefficients within 0.889–0.924, and Cohen's κ within 0.478–0.663 when considering the four OSA severity categories. These resulted in accuracies ranging 87.2%–96.6%, 81.1%–87.6%, and 91.6%–94.6% when assessing the three typical AHI severity thresholds, 5 events/hour (e/h), 15 e/h, and 30 e/h, respectively. Our model also revealed the importance of the SpO_2 predictors, thereby minimizing the 'black box' perception traditionally attributed to the machine-learning approaches. Furthermore, a decision curve analysis emphasized the clinical usefulness of our proposal. Therefore, we conclude that the LSBoost-based model can foster development of clinically applicable and cost saving protocols for detection of patients attending primary care services, or to avoid full polysomnography in specialized sleep facilities, thus demonstrating the diagnostic usefulness of SpO_2 signals obtained at home.

1. Introduction

Simplification of obstructive sleep apnea (OSA) diagnosis has become a major research priority in the past years. The OSA diagnostic gold standard, the nocturnal polysomnography (PSG), is intended to detect the typical overnight recurrence of apneas (complete cessations of airflow) and hypopneas (significant reductions of airflow) [1], which lead to inadequate gas-exchange, namely intermittent hypoxia, as well as fragmented sleep [2]. These events are associated in a severity-dependent fashion with a number of cardiovascular and metabolic morbidities [3], with frequency in co-morbidities estimated in 46.4%–59.5% for hypertension [4,5], 29.0%–56.8% for heart diseases [5,6], and 14.7%–19.5% for diabetes mellitus [4,5]. Moreover, high presence of other respiratory (6.4%–16.0%) and psychiatric (5.0%–14.5%) diseases has been also reported [4,6]. The overnight application of continuous positive airway pressure (CPAP) is the first and most widely used choice for treatment. It has been shown to reduce apneic events, blood pressure, and daytime somnolence, as well as improve some physical and sleep-related quality of life indicators [7]. However, PSG complexity, high costs, intensive labor, and patient discomfort [8–10], along with the large number of undiagnosed OSA cases estimated at nearly 1 billion worldwide [11], lead to limited availability of facilities and delayed access to diagnosis and treatment.

Studies predicated on the use of the single-channel blood oxygen saturation signal (SpO_2) have been popular approaches to

[*] Correspondence to: ETSI Telecomunicación, Campus Miguel Delibes, Paseo Belén 15, 47011, Valladolid, Spain.
E-mail address: gonzalo.gutierrez@gib.tel.uva.es (G.C. Gutiérrez-Tobal).

https://doi.org/10.1016/j.asoc.2021.107827
1568-4946/© 2021 Elsevier B.V. All rights reserved.

overcome these PSG drawbacks [12,13]. SpO_2 is easily recorded during the night using a pulse-oximeter, and the potential development of automated analyses contrasts with the labor-intensive visual inspection of multiple channels required in PSG [1]. The promising results of a combination of automatically extracted information from pulse oximetry data coupled with machine-learning approaches suggest that SpO_2 could serve as a suitable and valid candidate to simplify OSA diagnosis [14–22]. More specifically, recent studies have shown the usefulness of combining information from unsupervised at-home overnight oximetry with ensemble-learning methodologies [17,18], but have also exposed several serious limitations. Among these, there is a need for deeper evaluation of the SpO_2 information using additional analytical approaches, which was coupled with inability to generalize the findings due to relatively restricted datasets. Furthermore, none of the studies evaluated the diagnostic ability of their predictive models in both non-referral and clinical referral cohorts, i.e., in community dwellers and in symptomatic patients being referred to sleep specialists by their primary care physicians due to clinical suspicion of OSA, respectively.

The current study focused on automatically detecting OSA using at-home oximetry. We hypothesized that the information contained in the overnight SpO_2 signal can be used to diagnose OSA in all comers, i.e., including both referral and non-referral cohorts. Accordingly, our main objective was to obtain, and evaluate in both types of cohorts, a novel unified approach that can accurately estimate the apnea–hypopnea index (AHI), as it is commonly used to determine the presence of OSA and its severity [1]. Thus, our first step was to expand the analytical approaches previously applied to the SpO_2 signal to further characterize it. Then, we applied the least-squares boosting (LSBoost) ensemble-learning algorithm with stumps to derive a regression model that allows for estimation of the AHI from the SpO_2 information [23]. Ensemble-learning has been applied and shown promise in a wide range of fields, including OSA detection [17,18,24–27]. Here, we use LSBoost for the first time with this purpose as it retains the robustness against overfitting of boosting ensemble-learning algorithms, and has proven preserved effectiveness in regression problems without the cost of intensive computation requirements [28,29]. Additionally, the combined use of LSBoost and stumps as base classifiers enable us to explain its predictions based on the relative importance of the SpO_2 data used [23,30], thereby maximizing the interpretability of the model. Finally, the LSBoost regression model was assessed in terms of agreement with the actual AHI. Furthermore, the agreement between the OSA severity categories derived from the estimated AHI and the actual AHI was also evaluated. This latter evaluation included the diagnostic performance in common AHI thresholds as well as a decision curve analysis, thus providing a more complete view of the potential clinical usefulness of our proposal.

We think that a model that accurately assigns an estimate of AHI as a correlated of OSA severity in cohorts with both high (clinical-referred) and low (community-based) pre-test probability would be of great value to help physicians in their decisions. Indeed, such tool would enable its use in both primary and specialized healthcare centers, with increased efficiency, reduced costs, and most importantly maximal benefit for patients when combined with e-health systems [31,32].

2. Materials and methods

Next subsections detail the databases and methods used in our study. Fig. 1 shows a general flowchart of the methodology conducted.

2.1. Databases description and validation strategy

Two distinctly different databases were included in the study. Subjects and data from the Sleep Heart Health Study (SHHS) were used as a community-based non-referral sample [32,33]. SHHS is a publicly available dataset in which subjects are at least 40 years old and were recruited from several previously established cohorts aimed at assessing cardiovascular risks. A total of 5,804 individuals underwent unattended at-home overnight PSG, to then assess whether OSA is an independent risk factor for developing cardiovascular morbidity [33]. As part of these PSGs, data from 5,793 SpO_2 recordings were available for this study (SHHS1 subset). A follow-up PSG was then administered to 2,647 subjects from the original sample five years later, using the same procedures (SHHS2 subset) [33]. These PSGs and their corresponding SpO_2 recordings were also exploited. Further details on the composition and characteristics of the SHHS dataset can be obtained from the original studies [33,34].

A second dataset, a clinical cohort, was composed of 322 adult symptomatic patients referred to the sleep unit of the Rio Hortega University Hospital (Valladolid, Spain) due to clinical suspicion of OSA (RHUH dataset). They all underwent at-home PSG (Embletta MPR with the ST + proxy, Embla Systems, Natus Medical Inc. CA, USA) and a simultaneous overnight portable oximetry (Nonin WristOx2 3150, Nonin Medical, Inc,MN, USA), from which the SpO_2 recordings were obtained. The Ethics and Clinical Research Committee of the Hospital approved the protocol (CEIC 47/16).

In both samples, apneas and hypopneas were scored by specialized personnel following the current recommendations of the American Academy of Sleep Medicine (AASM) [1,35], based on which, AHI was determined for each subject, and subsequently classified into one of the four OSA severity categories: no OSA (AHI $<$ 5 events/hour), mild OSA (5 e/h $\leq$ AHI $<$ 15 e/h), moderate OSA (15 e/h $\leq$ AHI $<$ 30 e/h), and severe OSA (AHI $\geq$ 30 e/h).

The datasets were divided into five groups, as delineated in Fig. 2 and in accordance with following criteria:

(i) *The training and validation sets ($SHHS1_{tr}$ and $SHHS1_{v}$, respectively) are only composed of subjects from the SHHS1 subset that was not a participant in the follow-up sleep study 5 years later*. Thus, we avoid the bias caused by the inclusion of recordings from the same subject in the development and the testing of the model.

(ii) *$SHHS1_{tr}$ and $SHHS1_{v}$ are composed of the same number of subjects in each of the four OSA severity categories*. This approach served to ensure balanced distribution of severity groups, thereby not favoring a model with biased final diagnosis towards one of the categories.

(iii) *There are three test sets, two non-referral and one referral.* $SHHS1_{t}$ (non-referral), composed of those subjects from SHHS1 not included in $SHHS1_{tr}$ or $SHHS1_{v}$; all the recordings from SHHS2 subset (non-referral); and all the patients from the Rio Hortega University Hospital (RHUH, clinical referral).

Table 1 summarizes sociodemographic and clinical data of the subjects. As expected, all non-referral groups showed statistically significant differences in AHI, age, sex, and race distribution with those in referral RHUH cohort ($p < 0.01$ after Bonferroni's correction). Furthermore, all non-referral groups except SHHS2 showed statistically significant differences with referral RHUH in body mass index (BMI). Minor differences were found in the proportion of black subjects of $SHHS1_{tr}$ compared with $SHHS1_{t}$ and SHHS2, as well as of $SHHS1_{v}$ compared with SHHS2. Finally, all groups showed statistically significant differences in age with SHHS2, as anticipated by the SHHS study design.

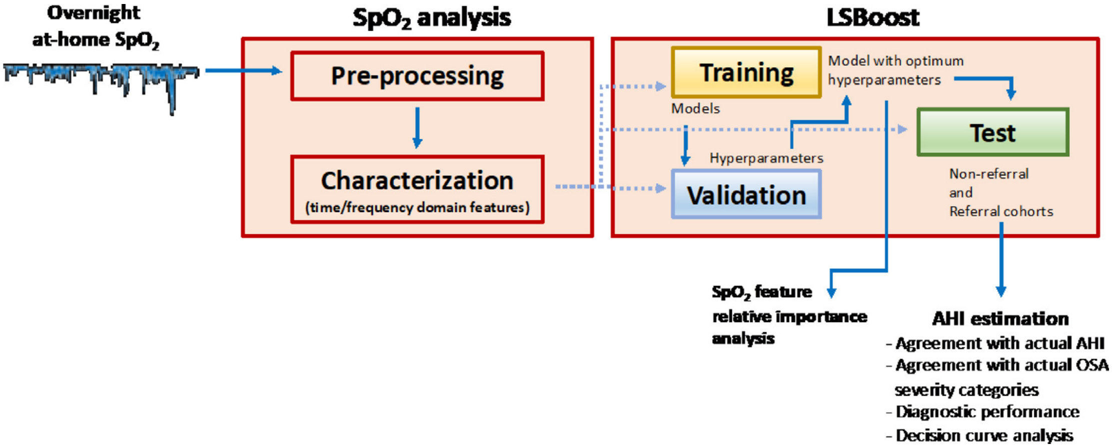
Overnight at-home SpO₂
SpO₂ analysis
Pre-processing
Characterization
(time/frequency domain features)
LSBoost
Training
Model with optimum hyperparameters
Models
Hyperparameters
Validation
Test
Non-referral and Referral cohorts
SpO₂ feature relative importance analysis
AHI estimation
- Agreement with actual AHI
- Agreement with actual OSA severity categories
- Diagnostic performance
- Decision curve analysis

Fig. 1. Flowchart with the main methodological steps of the study. At-home overnight SpO$_2$ recordings are pre-processed and comprehensively characterized using different analytical approaches. The features extracted from SpO$_2$ are used to train multiple LSBoost-based regression models with ability to estimate AHI. The validation step is used to select the model with the optimum hyperparameters, which is subsequently assessed in independent test sets from both non-referral and referral cohorts. The AHI estimated by this selected model is used to measure the agreement with actual AHI, as well as the actual OSA severity categories that are commonly derived from it. Its diagnostic performance is also assessed, and decision curves are provided. Finally, the relative importance of each SpO$_2$ extracted feature in the AHI estimation is analyzed.

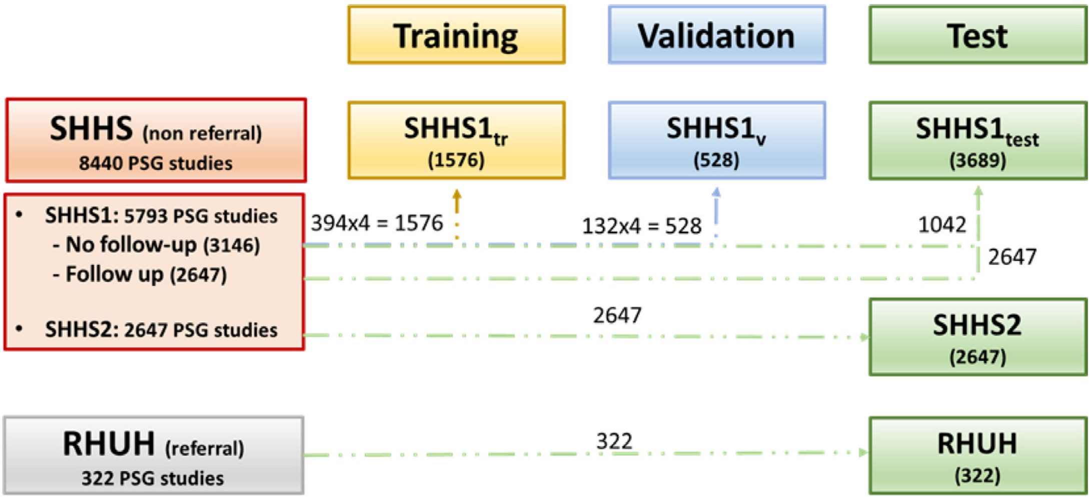

Fig. 2. Distribution of the subjects involved in the study to conduct the validation strategy. The total number of subjects included in SHHS1$_{tr}$or SHHS1$_v$ is limited by the criterion of same size of the four OSAS severity degrees and the size of the OSAS degree less represented (no OSAS: 526 subjects). Among those subjects included in SHHS1$_{tr}$ or SHHS1$_v$, 75% (1,576) were heuristically assigned to training and 25% (528) to validation. The remaining subjects from SHHS1 (3,689), all SHHS2 (2,647), and all RHUH (322), were assigned to test sets. SHHS: Sleep Heart Health Study dataset; RHUH: Rio Hortega University Hospital dataset.

Table 1
Sociodemographic and clinical data of each subgroup. Median and interquartile range for Age, BMI, and AHI (p-values obtained with Mann–Whitney non-parametric U test). Number of subjects for males (M), females (F), white (W), black (B), and other (O) (p-values obtained with Fisher's exact test). Statistical differences in race refer only to the proportion of black subjects, as no differences were found in the proportion of the 'other' category. Category names for race were those originally reported in the SHHS dataset.

Data	SHHS1$_{tr}$ (N = 1,576)	SHHS1$_v$ (N = 528)	SHHS1$_t$ (N = 3,689)	SHHS2 (N = 2,647)	RHUH (N = 322)	p<0.01*
Age (years)	63.0 (55.0, 73.0)	64.5 (55.0, 74.0)	63.0 (55.0, 71.0)	68.0 (60.0, 76.0)	57.0 (46.0, 66.0)	c,d,f,g,h,i,j
Sex (M/F)	811/765	255/273	1,967/1,722	1,422/1,225	220/102	d,g,i,j
Race (W/B/O)	1,248/174/118	432/62/34	3,183/278/228	2,307/181/159	322/0/0	b,c,d,f,g,i,j
BMI (kg/m^2)	27.4 (24.5, 30.8)	27.1 (24.5, 30.5)	27.5 (24.7, 30.7)	27.7 (24.8, 31.1)	28.4 (25.8, 32.1)	d,g,i
AHI (e/h)	15.0 (5.0, 30.0)	15.0 (5.0, 29.8)	12.6 (7.4, 21.5)	13.5 (6.8, 24.6)	26.2 (12.9, 46.6)	d,g,i,j

AHI: apnea–hypopnea index, BMI: body mass index, *p-value after Bonferroni's correction, [a]SHHS1$_{tr}$ *vs.* SHHS1$_v$, [b]SHHS1$_{tr}$ *vs.* SHHS1$_t$, [c]SHHS1$_{tr}$ *vs.* SHHS2, [d]SHHS1tr *vs.* RHUH, [e]SHHS1$_v$ *vs.* SHHS1$_t$, [f]SHHS1$_v$ *vs.* SHHS2, [g]SHHS1$_v$ *vs.* RHUH, [h]SHHS1$_t$ *vs.* SHHS2, [i]SHHS1$_t$ *vs.* RHUH, [j]SHHS2 *vs.* RHUH.

2.2. Information obtained from the SpO_2 signal

All SpO_2 signals were acquired at a sampling rate of 1 Hz. Artifacts were removed as reported in previous studies [17,36]. Up to 32 features were obtained to conduct a thorough characterization from different approaches, thus minimizing the limitations of previous studies (see Table 2) [17,18]. Three different analytical methodologies were conducted: classic oximetry-based clinical features, non-clinical features in time domain (including non-linear analysis), and non-clinical features in frequency domain. All these features were proposed because they can help characterize information derived from biomedical signals [14,17,24,36–49]. However, they have never been used concurrently. Moreover, obtaining a higher number of features than in previous works favors the performance of the LSBoost approach adopted in this study [29].

2.3. Least-squares boosting algorithm

LSBoost is an ensemble-learning boosting algorithm intended for estimation of a continuous target variable y [23]. Its output is obtained as a combination of estimations from several base learners (h). These are sequentially obtained so that each new learner is trained to fit the remaining residual error ($\tilde{y}^m$) after combining the outputs from all previous learners. [29] Formally, the algorithm can be described as follows [23,29]:

(i) Set $m = 0$ and initialize the estimated output $f^0(\mathbf{x})$.
(ii) Increase m by 1 and compute the residuals $\tilde{y}_i^m = y_i - f^{m-1}(\mathbf{x}_i), i = 1, 2, \ldots, N$
(iii) Fit the residuals using least squares loss function along with learner h and the predictors for each subject $\mathbf{x}_i$: $(\upsilon_m, \mathbf{a}_m) = \text{argmin}_{\mathbf{a},\upsilon} \sum_{i=1}^{N} \left[\tilde{y}_i^m - \upsilon h(\mathbf{x}_i; \mathbf{a})\right]^2$, with $\mathbf{a}$ being the set of parameters of h, and υ a regularization factor ranging $0 < \upsilon \leq 1$.
(iv) Update $f^m(\mathbf{x}) = f^{m-1}(\mathbf{x}) + \upsilon_m h(\mathbf{x}; \mathbf{a}_m)$.
(v) Iterate (ii) to (iv) until $m = M$, being M the maximum number of learners to be used.

In the current study, y was the actual AHI, $f^m(\mathbf{x})$ was the estimated AHI, $\mathbf{x}_i$ the set of features extracted from each i SpO_2 recording, and N the number of recordings in the training set. Both M and υ are tuning parameters [17]. We used our validation group ($SHHS1_v$) for this purpose. Regression trees of one parent and two children nodes (stumps) were used as the base learners h. In this way, every new $h(\mathbf{x}; \mathbf{a}_m)$ is a function of a single predictive variable [28,29], thus conducting a default feature selection procedure at each iteration. The relative importance $\hat{I}_j^2$ of a variable $\mathbf{x}_j$ can be estimated on the basis of its squared error empirical improvement along all the trees in which it has been involved [23,30]:

$$\hat{I}_j^2 = \frac{1}{M} \sum_{m=1}^{M} MSE_m(\mathbf{x}_j) \cdot w_m - (MSE_m^l(\mathbf{x}_j) \cdot w_m^l + MSE_m^r(\mathbf{x}_j) \cdot w_m^r), \quad (1)$$

where MSE_m is the mean squared error for the m stump associated to $\mathbf{x}_j$, w_m a weight accounting for the parent node probability, and $l - r$ denoting the corresponding parameters for the two child nodes. The relative $\hat{I}_j^2$ values of the features in this study were subsequently scaled to sum 100, with higher values denoting a higher influence in the ensemble output [51].

2.4. Statistical analyses

The agreement between the actual AHI and the AHI estimated by our regression model was evaluated by means of the intra-class correlation coefficient (ICC) and Bland–Altman plots [52,53]. Furthermore, as four severity categories based on the AHI are commonly used by clinicians in OSA context (no OSA: AHI $<$ 5 e/h; mild: 5 e/h $\leq$ AHI $<$ 15 e/h; moderate: 15 e/h $\leq$ AHI $<$ 30 e/h; and severe: 30 e/h $\leq$ AHI), Cohen's kappa, κ, and confusion matrices were used to measure agreement between the actual OSA-severity categories and those obtained from our AHI estimation [54]. The diagnostic usefulness of our model was also assessed in each of the common thresholds used to set the OSA-severity degrees: 5 e/h, 15 e/h, and 30 e/h. This was conducted in terms of sensitivity (Se, percentage of subjects with a true diagnosis above the corresponding threshold that are rightly classified by our model), specificity (Sp, percentage of subjects with a true diagnosis below the corresponding threshold that are rightly classified by our model), accuracy (Acc, total percentage of subjects rightly classified by our model), positive predictive value (PPV, percentage of subjects with a true diagnosis above the corresponding threshold among all those that our model classify above that threshold), negative predictive value (NPV, percentage of subjects with a true diagnosis below the corresponding threshold among all those that our test classify below that threshold), positive likelihood ratio (LR+, ratio of the true positive rate to the false positive rate), and negative likelihood ratio (LR-, ratio of the false negative rate to the true negative rate). The diagnostic metrics for these thresholds can be computed as follows:

$$Se = \frac{TP}{TP + FN} * 100, \quad (2)$$

$$Sp = \frac{TN}{TN + FP} * 100, \quad (3)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} * 100, \quad (4)$$

$$PPV = \frac{TP}{TP + FP} * 100, \quad (5)$$

$$NPV = \frac{TN}{TN + FN} * 100, \quad (6)$$

$$LR+ = \frac{Se}{1 - Sp}, \quad (7)$$

$$LR- = \frac{1 - Se}{Sp}, \quad (8)$$

where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives for each threshold, respectively. Decision curves were also used to further assess the clinical value of our proposal [55,56]. Accordingly, net benefit (NB) for each possible "threshold probability" was computed as follows [55,56]:

$$NB = \frac{TP}{TP + FP + TN + FN} - \frac{FP}{TP + FP + TN + FN} \left(\frac{pt}{1 - pt}\right), \quad (9)$$

where pt is the threshold probability considered at each case. Moreover, as our model is aimed at estimating a continuous variable, logistic regression was applied prior to the decision curve analysis to transform the estimated AHI into a probability [57]. Finally, a two-tailed p-value less than 0.01 was considered as achieving statistical significance. Matlab^{TM} 2018b and 2020b were used to conduct all the analyses of this study.

3. Results

3.1. Model training and validation

The original paper presenting the LSBoost method established m and υ as tuning hyperparameters [23]. It also offered the ranges among they were more likely to produce lower error rates. The main idea, confirmed in subsequent studies [28,29], was to choose a low υ value and vary m. Here, we chose to evaluate the

Table 2
Features extracted from each of the overnight SpO_2 recordings.

Features	Description
	Classic clinical features
ODI3	3% oxygen desaturation index. Number of 3% desaturation events per hour of sleep. [36]
CT90	Overnight cumulative time of blood oxygen saturation under 90%. [36]
	Statistics and Non-linear measures in Time Domain
Mt1-Mt4	First (mean), second (standard deviation), third (skewness), and fourth (kurtosis) statistical moments of the overnight signals in time domain. [14]
CTM	Central tendency measure to measure variability in a time series. [44,49]
LZC	Lempel–Ziv complexity to measure the complexity degree of a time series. [37,47]
SampEn	Sample entropy to measure irregularity in a time series. [38,45]
msEnt	Multiscale entropy analysis, which extends *SampEn* to different time scales. [39,46] Five features were extracted from up to 50 scales: – $msEnt_{max}$: the maximum *SampEn* value of the scales. – $msEnt_{scale}$: the scale at which the maximum *SampEn* value is reached. – $msEnt_{area}$: the area under the curve formed by the *SampEn* values of all the scales. – $msEnt_{slp1}$: the average slope of the lower scales (1 to 23). – $msEnt_{slp2}$: the average slope of the higher scales (24 to 50).
	Frequency Domain Analysis
Mf1-Mf4	First (mean), second (standard deviation), third (skewness), and fourth (kurtosis) statistical moments of the full spectrum. [14]
SpecEn	Spectral entropy to measure the full spectrum flatness. [17,40]
MF	Median frequency to estimate the distribution of the power of the full spectrum. [17]
ED	Euclidian distance to directly estimate the statistical distance between the full spectrum and a uniform distribution. [41,48]
WD	Wootter's distance to estimate the statistical distance between the full spectrum and a uniform distribution based on counting all the intermediate states between the distributions. [24,42]
MA^{*}_{BOI}	Maximum of the spectrum amplitude within the SpO_2 band of interest.
mA_{BOI}	Minimum of the spectrum amplitude within the SpO_2 band of interest.
$Mf1_{BOI}$ -$Mf4_{BOI}$	First, second, third, and fourth statistical moment of the spectral band of interest. [17]
$SpecEn_{BOI}$	Spectral entropy applied to the spectral band of interest. [17,40]
MF_{BOI}	Median frequency applied to the spectral band of interest. [17]
ED_{BOI}	Euclidian distance applied to the spectral band of interest. [41,48]
WD_{BOI}	Wootter's distance applied to the spectral band of interest. [24,42]

The OSA-related band of interest (BOI) in the SpO_2 signal comprises 0.014–0.033 Hz (events lasting 30–70 seconds) [14,50].

same low υ values than in the original paper, i.e., $\upsilon = 0.031$, 0.062, and 0.125 [23]. Also, it was shown that evaluating *m* values beyond 200 produced models with very similar behaviors [23]. Hence, we also chose to vary *m* from 1 to 200 in steps of 1. All the models resulting from the combination of the different values of υ and *m* were obtained from the training group and subsequently evaluated in the validation group to choose the optimum (υ, *m*) pair. Fig. 3 shows the results of hyperparameter tuning on $SHHS1_v$. According to Cohen's κ, the optimum values for υ and *m* were 0.125 and 199, respectively.

Each of the 199 regression stumps used a single feature to provide its contribution to the final ensemble output. Twenty five out of the 32 extracted features were used at least once. However, 9 of them (*ODI3*, *M4t*, $M1f_{BOI}$, MA_{BOI}, $msEnt_{scale}$, *LZC*, $msEnt_{area}$, *WD*, and *SampEn*) were selected 132 times and gathered 99.0% of the total relative importance $\hat{I}^2$, thus highlighting the relevance of all the analytical approaches conducted (clinical features, statistics and non-linear features, and frequency domain analysis), as well as the minimum effect on the final AHI estimation of 23 features. Moreover, *ODI3* alone accounted for 85.7% of $\hat{I}^2$ and was selected 43 times.

3.2. Agreement and diagnostic performance

Fig. 4 shows the Bland–Altman plots of actual and estimated AHIs for the three test sets: $SHHS1_t$, SHHS2, and RHUH. The smallest bias can be observed for $SHHS1_t$, whereas SHHS2 and RHUH show mild overestimation and underestimation of actual AHI, respectively. Furthermore, ICC is embedded in each corresponding plot, showing high agreement in the three groups. The referral dataset reached slightly higher ICC (0.924 RHUH) than the two non-referral ones (0.900 $SHHS1_t$; 0.889 SHHS2).

Fig. 5 shows the confusion matrices that compare the actual OSA severity degrees and the corresponding assignation using the estimated AHI. Cumulated accuracy in the four classes reached 70.02%, 62.30%, and 77.02% for $SHHS1_t$, SHHS2, and RHUH datasets, respectively. Moreover, their confusion matrices correspond to 0.561, 0.478, and 0.663 four-class κ. An increased proportion of subjects correctly diagnosed can be observed in the main diagonal of $SHHS1_t$ as OSA severity increases, with a good balance between overestimate and underestimate findings, as anticipated by the corresponding Bland–Altman plot. In contrast, the confusion matrix from SHHS2 shows a tendency for overestimation that decreases with the severity degree. Finally, the confusion matrix from RHUH shows small overestimation and underestimation for no OSA and moderate OSA, respectively, yet not obscuring the very high diagnostic ability reached in the referral database.

Table 3 displays the diagnostic statistics for each AHI threshold that defines OSA severity. They are directly derived from the confusion matrices in Fig. 5. Accuracies (Acc) are above 80% in all cases, and only 15 e/h in SHHS2 is below 85%. Results in $SHHS1_t$ are generally higher than in SHHS2 because of remarkable higher Sp values at the cost of mild lower Se. Actually, Sp values in SHHS2 are relatively low for 5 e/h and 15 e/h, in accordance with

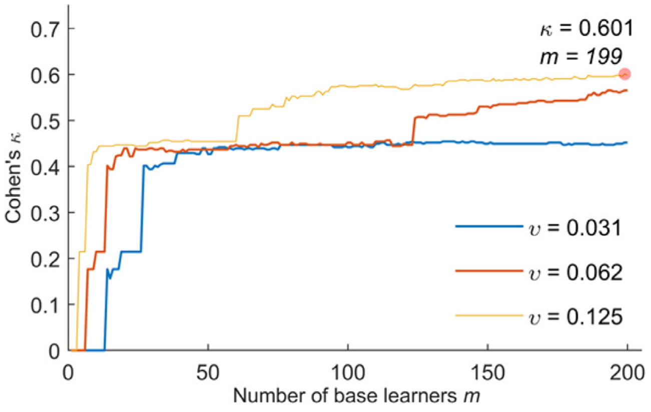

Fig. 3. Number of base learners (m) and regularization parameter (υ) tuned according to Cohen's κ in the validation set SHHS1v.

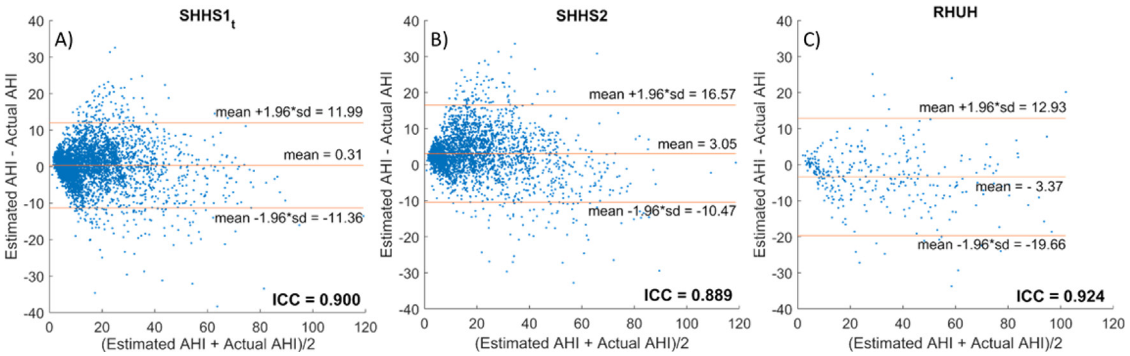

Fig. 4. Bland–Altman plots and intra-class correlation coefficient (ICC) of the test sets (A) SHHS1t, (B) SHHS2, and (C) RHUH. SHHS1t shows a small bias (mean = 0.31) and the tiniest difference in the 95% confidence interval (23.35). Higher overestimation (mean = 3.05) and 95% confidence interval (27.04) is reached in SHHS2. Actual AHI presents a small underestimation in RHUH (mean = −3.37), which also reaches the widest 95% confidence interval (32.59).

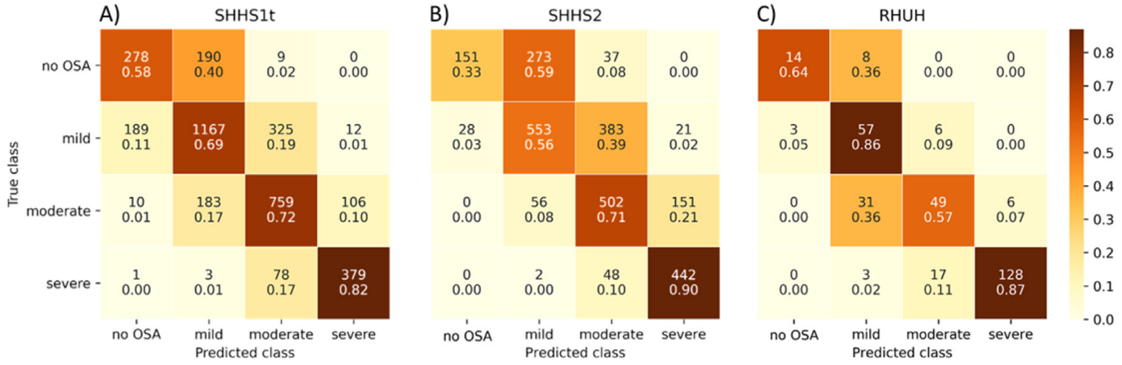

Fig. 5. Confusion matrices with the true OSA severity classes against the predicted ones for (A) SHHS1t, (B) SHHS2, and (C) RHUH. The main diagonal indicates the number and proportion of rightly assigned subjects for each severity degree, whereas the remaining cells are misclassified subjects. Darker colors are shown as more proportion of subjects of the same actual class are assigned to a cell.

Table 3
Diagnostic performance on the clinical AHI thresholds used to demarcate OSA severity categories (5 e/h, 15 e/h, and 30 e/h).

	SHHS1$_t$			SHHS2			RHUH		
	5 e/h	15 e/h	30 e/h	5 e/h	15 e/h	30 e/h	5 e/h	15 e/h	30 e/h
Se (%)	93.77	87.03	82.21	98.70	95.17	89.83	99.00	85.47	86.49
Sp (%)	58.28	84.06	96.34	32.80	69.50	92.01	63.64	93.18	96.55
PPV (%)	93.89	79.26	76.26	87.44	72.16	71.99	97.38	97.09	95.52
NPV (%)	58.16	90.25	97.43	84.36	95.54	97.54	82.35	70.69	89.36
LR+	2.25	5.46	22.46	1.47	3.12	11.24	2.72	12.53	25.07
LR-	0.11	0.15	0.18	0.04	0.07	0.11	0.02	0.16	0.14
Acc (%)	89.18	85.28	94.58	87.23	81.14	91.61	96.58	87.58	91.93

Acc: accuracy, LR+/LR-: positive and negative likelihood ratio, PPV/NPV: positive and negative predictive value, Se/Sp: sensitivity and specificity.

the overestimated AHI scenario previously observed. The highest diagnostic performance, however, is obtained in the clinically referred high pre-test probability database RHUH, reaching the most favorable values for OSA diagnosis in most of the statistics.

3.3. Decision curve analysis

Fig. 6 A shows the decision curves corresponding to the $\mathrm{SHHS1}_t$ and SHHS2 groups, i.e., the non-referral subjects. Solid lines represent the net benefit curves of our LSBoost-based model. In a non-referral cohort, one would expect to use our proposal as a screening test to send patients to the standard diagnostic test, the PSG. Therefore, our proposal is compared to the option of sending all subjects to the PSG, (colored thin lines with triangles), and sending no one to PSG (gray horizontal thin line with triangles), since PSG is the next natural intervention for non-referral subjects. The curve shows higher net benefit for our model than for sending no one to the PSG for almost all probability thresholds. The model also shows higher net benefit than sending all subjects to PSG from the threshold probability = 0.39 onwards.

Fig. 6B represents the decision curves corresponding to the RHUH group, i.e., the referral subjects. As they have previous suspicious of OSA, this is a high OSA pre-test probability group. Therefore, the LSBoost-based model (colored solid lines) is compared to 'treat all' subjects for OSA and 'treat no one' for OSA (gray lines), since treatment is the next natural intervention for these subjects. As in the previous case, our model (purple solid line) reaches higher net benefit than 'treat no one', for all probability thresholds, and also higher net benefit than 'treat all' from 0.39 probability threshold onwards. In addition, we compare our model to the performance of the PSG (colored dashed lines). PSG produces no false positive results if no other drawback is considered. Consequently, its net benefit is maximum, as reflected by the dashed purple line at the top of the graphic. However, Eq. (9) can be reformulated to account for other issues, such as health risk or economic costs, by subtracting a penalty term called "test harm", ($NB_{\text{test-harm}} = NB$-"test arm") [55]. The test harm represents the negative implications caused by taking the test (i.e, conducting the intervention) and can be explained as the reciprocal of the number of tests that a clinician is willing to conduct in order to find one true positive subject provided that the test were perfectly accurate [56,58]. The pink and blue dashed lines represent the PSG performance if a clinician would do no more than 2 and 5 PSGs to identify one true positive, i.e., if the test harm is 1/2 and 1/5, respectively. To compare our model with PSG, we assumed that, in the absence of any other harm, a clinician would do more of our tests in the same proportion as the decreased cost of at-home oximetry with respect to standard PSG. Previous studies reported that standard PSG (810.8 $/test) is 3.87 times more expensive than at-home single-channel oximetry (209.2 $/test) [59], which is a sustained proportion over time [60]. Accordingly, the pink and blue solid lines represent the net benefit of our model when the test harm is $1/(2 * 3.87)$ and $1/(5 * 3.87)$. As observed, the orange rectangles represent the upper limits of the range in which our model achieves higher net benefit than PSG. This range is larger as less tests a clinician is willing to conduct to find a true positive.

Finally, the black rectangles represent the bottom limits of the range in which our model shows higher net benefit than the 'treat all' strategy. For each PSG 'test harm' different from 0, the range between the black and orange rectangles are the ranges in which our model achieves higher net benefit than all other options: 0.64–0.79 for a PSG test harm of 1/5 and 0.81–1.00 for a PSG test harm of 1/2.

3.4. Performance of other machine-learning approaches

Table 4 shows the performance of 3 additional machine-learning models that were obtained using the same training/validation/test strategy. Two of them are regression approaches (regression trees, RT; and regression support vector machines, SVMr), and another one a multi-class approach (multiclass adaptive boosting, AdaBoost.M2). These methods have shown its usefulness in several biomedical signal processing problems, including OSA automatic detection [17,18,24,61]. Consequently, they are used for comparison purposes. As observed, the other models outperformed our proposal in some discrete statistics (values in bold) when evaluating individual databases or AHI thresholds (SVMr in 27 out of 63 statistics, and AdaBoost.M2 and RT in 23). However, when considering the overall performance measurements (four-class accuracy and Cohen's κ) of our LSBoost model, it clearly outperformed the other methods in our three test sets ($\mathrm{SHHS1}_t$, SHHS2, and RHUH), except for the κ value of the RT model in SHHS2. In this unique case our LSBoost model reached $\kappa = 0.478$ and RT reached $\kappa = 0.479$. Moreover, in 23 out of 27 of the accuracies corresponding to each AHI threshold our proposal also reached the highest values. These figures highlight the superiority of our proposal comparing to the models from the other machine-learning approaches.

4. Discussion

In this study, we have accomplished substantial advances in SpO_2 characterization by extracting up to 32 overnight features from different analytical approaches. They serve to develop a novel OSA-specific LSBoost-based model that has the ability to accurately estimate AHI from single-channel oximetry data obtained in the patients' home. Furthermore, this model displays a high level of agreement with the actual PSG-derived AHI, and high diagnostic performance in both referral and non-referral cohorts. Our analytical approaches not only allowed us to develop a clinically useful tool, but also explain the model through the SpO_2 data used by each of the 199 base classifiers of the ensemble, thus decreasing the common 'black box' perception of automatic models.

4.1. Explaining the AHI estimation

ODI3 accounted for 85.7% of the relative importance, i.e., the contribution to the final overall AHI estimation. This is consistent with previous studies in which *ODI3* reflected the principal OSA-related informative component regarding SpO_2 [17,43]. However, *ODI3* alone underestimates AHI as not all apneic events lead to a desaturation [17,18,35]. Underestimation is not generally observed in our model, suggesting that the information contained in the remaining features is counteracting this effect. This would be supported by the similarity between the relative importance accounted by these features (14.3%) and the average amount of apneic events not accompanied by a 3% desaturation in SHHS dataset (11.5%) [18]. Most of this remaining relative importance is composed of the eight features mentioned in Results section, which are related to signal complexity (LZC, $msEnt_{\text{area}}$), irregularity ($mSEnt_{scale}$, $SampEn$), SpO_2 values distribution over time ($M4t$) and frequencies (WD), and the number and amplitude of recursive desaturations lasting 30 to 70 s irrespective of specific percentages of decrease ($M1f_{BOI}$, MA_{BOI}). Therefore, we posit that the overnight patterns characterized by these features reflect additional information about OSA, regardless of whether it is related to other events involved in AHI definition (e.g. arousals) or not (e.g. hypoxic burden). However, further research will be required to define the specific relationships between these parameters and other OSA-related effects.

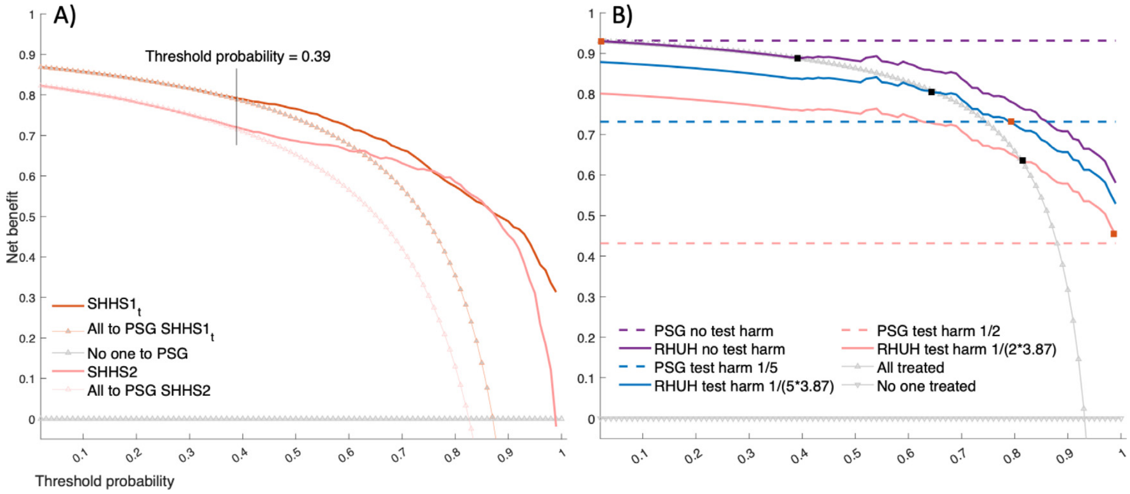

Fig. 6. Decision curves for the groups (A) SHHS1t and SHHS2, and (B) RHUH. A logistic regression transformation was conducted on the AHI to present it as the probability of having any OSA degree. In panel (A), the LSBoost-based model is compared to the strategy of "sending all to PSG" and "sending no one to PSG". In panel B the model is compared to "sending all to treatment", "sending no one to treatment", and the benefit of PSG under several "test harm" effects. Correspondence between threshold probability and the AHI thresholds 5 e/h, 15 e/h and 30 e/h are also shown.

Table 4
Performance of three machine-learning alternatives in our test sets.

	$SHHS1_t$			SHHS2			RHUH		
	5 e/h	15 e/h	30 e/h	5 e/h	15 e/h	30 e/h	5 e/h	15 e/h	30 e/h
	AdaBoost.M2								
Se (%)	87.32	**88.67**	**84.81**	96.11	**95.34**	**92.07**	90.00	85.47	85.81
Sp (%)	**70.23**	79.44	94.24	**48.80**	63.07	89.28	**100.00**	92.05	**97.13**
PPV (%)	**95.18**	75.12	67.76	**89.90**	68.20	66.23	**100.00**	96.62	**96.21**
NPV (%)	45.15	**90.93**	**97.75**	72.58	94.21	**98.01**	42.31	70.43	88.95
LR+	**2.93**	4.31	14.72	**1.88**	2.58	8.59	**Inf**	10.74	**29.86**
LR-	0.18	**0.14**	**0.16**	0.08	0.07	**0.09**	0.10	0.16	0.15
Acc (%)	85.12	83.25	93.06	**87.87**	77.71	89.80	90.68	87.27	91.93
Acc4 (%)	63.40			58.86			73.60		
Cohen's κ	0.489			0.443			0.622		
	SVMr								
Se (%)	**96.01**	**91.97**	**84.60**	**99.18**	**97.92**	**90.45**	**100.00**	**96.15**	**89.86**
Sp (%)	46.54	74.10	95.17	21.26	57.47	89.51	50.00	67.04	91.38
PPV (%)	92.36	71.31	71.43	85.66	65.66	66.32	96.46	88.58	89.86
NPV (%)	**63.43**	**92.95**	**97.74**	**84.48**	**97.08**	**97.62**	**100.00**	**86.76**	**91.38**
LR+	1.80	3.55	17.51	1.26	2.30	8.62	2.00	2.92	10.42
LR-	**0.09**	**0.11**	**0.16**	0.04	**0.04**	0.11	**0.00**	**0.06**	**0.11**
Acc (%)	**89.62**	81.46	93.85	85.60	75.82	89.69	96.58	**88.20**	90.68
Acc4 (%)	65.65			53.42			76.09		
Cohen's κ	0.501			0.364			0.639		
	RT								
Se (%)	87.02	83.61	**84.38**	95.88	93.76	**91.06**	90.00	84.62	85.14
Sp (%)	**70.65**	84.00	95.26	**49.46**	**69.78**	90.62	**100.00**	**94.32**	**97.13**
PPV (%)	**95.23**	78.54	71.77	**90.00**	72.04	68.92	**100.00**	**97.54**	**96.18**
NPV (%)	44.70	87.98	**97.71**	71.70	93.08	**97.80**	42.31	69.75	88.48
LR+	**2.96**	5.23	17.80	**1.90**	3.10	9.71	**Inf**	**14.89**	**29.63**
LR-	0.18	0.20	**0.16**	0.08	0.09	**0.10**	0.10	0.16	0.15
Acc (%)	84.90	83.84	93.90	**87.80**	80.66	90.71	90.68	87.27	91.61
Acc4 (%)	64.46			61.80			73.29		
Cohen's κ	0.498			**0.479**			0.618		

Acc: accuracy, Acc4: four-class accuracy, LR+/LR-: positive and negative likelihood ratio, PPV/NPV: positive and negative predictive value, Se/Sp: sensitivity and specificity. Bold values correspond to figures higher than our proposal.

4.2. A longitudinal perspective

A comparison of the results in $SHHS1_t$ and SHHS2 revealed that the latter shows a higher degree of OSA severity overestimation. As shown in Fig. 7, this behavior remains when the 2,647 exact same subjects from $SHHS1_t$ ($SHHS1_{t-fu}$) are compared with their follow-up test five years later (SHHS2). As age was the only characteristic in Table 1 that showed statistically significant differences between $SHHS1_t$ and SHHS2, we further analyzed our results to assess whether it is somehow influencing our predictions. For each $SHHS1_{t-fu}$ and SHHS2, the non-parametric Mann–Whitney *U* test was used to assess potential differences between the ages of subjects rightly or wrongly predicted within the same actual OSA group. Among the 12 comparisons for each of the datasets, age was only found significantly different (p-value < 0.01) in 2 for SHHS2 (no OSA vs. mild OSA and vs. moderate OSA within the actual class no OSA) and 2 more for SHHS1t-fu (no OSA vs. mild OSA within no OSA; and mild-OSA vs. no OSA

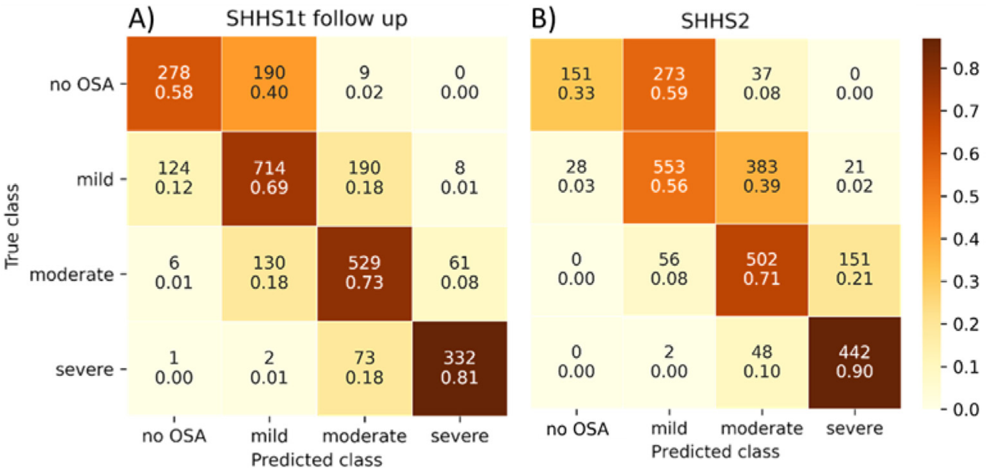

Fig. 7. Confusion matrices achieved in (A) the subjects from SHHS1t with a follow up PSG and (B) the same subjects 5 years later (SHHS2).

within mild OSA). Therefore, we conclude that age is not directly influencing the estimation of our model.

Night-to-night variability could underlie some of the different results found in SHHS1$_{t-fu}$ and SHHS2 [62]. However, since there is a clear tendency only towards overestimation, we hypothesize that the main reason may be that other morbidities developed during the 5-year span are influencing SpO$_2$ signals, thus increasing the estimated severity prediction in some subjects. This idea would be supported by the previously described rise in cardiovascular and respiratory diseases between SHHS1 and SHHS2, yet without significantly increasing AHI [63]. Although the diagnostic ability reached in the SHHS2 database is high, particularly for severe OSA, the future assessment of our model using cohorts with co-morbidities affecting oxygen saturation would potentially improve the accuracy of the application of our model. It would also provide more qualitative information to help clinicians in their therapeutic decisions.

4.3. Diagnostic ability and comparison with the state-of-the-art studies

Our OSA-specific LSBoost model reached high diagnostic ability in both non-referral and referral datasets. It additionally outperformed other regression and classification machine-learning methods evaluated on the same datasets. Previous studies also focused on the analysis of at-home SpO$_2$ to automate OSA detection (Table 5). Some of these models also reached high diagnostic ability, yet only assessed either one of the two cohort types.

Five studies have focused on non-referral cohorts. All but one used the SHHS dataset, at least partially. In contrast, Chung et al. directly evaluated different thresholds of 4% ODI to determine OSA severity in a sample of 475 surgical patients [64]. Accuracies were high at the cost of unbalanced Se/Sp pairs, which differed for more than 20 percentage points in all cases. Schlotthauer et al. and Rolón et al. (2018 and 2020) used a subset of the SHHS2 database to respectively evaluate estimations of *ODI3* and AHI in the 15 e/h severity threshold [19–21]. Schlotthauer et al. used empirical mode decomposition as the main analytical tool, whereas Rolón et al. applied discrepancy measures (2018) and structured dictionary learning (2020). The three studies reported high Acc with balanced Se/Sp. Finally, Deviaene et al. used the whole SHHS dataset to develop a random forest model focused on detecting 3% desaturations caused by apneic events [14]. They subsequently estimated AHI by counting these events and applied a robust regression methodology to correct biases. Their results reached lower κ than our LSBoost model in SHHS1, mainly due to our high Acc and more balanced Se/Sp in 15 e/h and 30 e/h. In contrast, they reported higher κ in SHHS2, where our model suffers from the aforementioned overestimation. However, their methodology to select the training set led them to include 223 subjects from SHHS1 that have a corresponding recording in SHHS2 [14]. No bias was found towards correct predictions in these subjects for the 15 e/h threshold, but this was not evaluated for 5 e/h and 30 e/h [14].

Five additional studies have focused on clinical referral databases. Olson et al. Rofail et al. and Gumb et al. used univariate approaches evaluating delta index, *ODI3*, and *ODI4*, respectively [22,62,65]. Rofail et al. reached high Acc and balanced Se/Sp for 30 e/h, but the remaining figures were moderate in the three studies. Gutiérrez-Tobal et al. used AdaBoost.M2 to directly classify subjects into four OSA severity degrees without AHI estimation prior to the class assignment [12]. Their results were noticeably less performant than in the current study probably due to the need for splitting their smaller sample into training and testing, a less deep characterization of SpO$_2$, and the use of a reference in-hospital AHI obtained in a different night than SpO$_2$ [12]. This latter drawback was corrected by Álvarez et al. resulting in a high four class κ after estimating AHI with support vector machines [61]. However, the performance in 5 e/h and 15 e/h suffered from unbalanced Se/Sp.

In summary, the single LSBoost model proposed performed similarly to all other methods that exhibited the highest diagnostic ability among non-referral cohorts, while clearly outperforming all the proposed approaches focused on referral databases. Additionally, our model also outperformed the two studies that used the SHHS dataset with other signals [66,67], thus suggesting the superiority of SpO$_2$ when following a single-channel approach to simplify OSA diagnosis. Uddin et al. however, followed a two-channel approach by the jointly use of airflow and SpO$_2$ [68]. They involved 988 subjects from the SHHS1 dataset to develop and test a new ad-hoc detection algorithm. Their method, when evaluated in the 15 e/h AHI threshold, reached clearly higher performance than the results of our model in our SHHS1 test group. Moreover, very similar figures were reached when comparing the results from 5 e/h and 30 e/h AHI thresholds, with their proposal increasing the complexity of the test due to the extra airflow channel.

4.4. Clinical usefulness of the proposal

Our OSA-specific model offers high diagnostic capability regardless of whether the strategy used focuses on primary care

Table 5
State-of-the-art studies focused on analyzing SpO_2 recordings acquired at home or using SHHS database with other signals.

Study	# Subjects	Purpose and main predictor	AHI (e/h)	Se (%)	Sp (%)	Acc (%)	Four class κ
Non-referral							
Chung et al. (2012) [64]	475	*ODI4* direct assessment (univariate)	5 15 30	96.3 70.0 76.0	67.3 92.5 97.2	87.0 84.0 93.7	nd$^+$
Schlotthauer et al. (2014) [19]	996 (SHHS2)	*ODI3* estimation and evaluation using Empirical Mode Decomposition (univariate)	15	83.8	85.5	nd	nd
Rolón et al. (2018) [20]	954 (SHHS2)	AHI estimation using Discrepancy Measures and Extreme Learning Machine	15	81.9	87.3	84.6	nd
Deviaene et al. (2019) [18]	5793 (SHHS1)	AHI estimation using Random Forest to detect apneic events and Robust Regression to correct bias	5 15 30	83.5 75.6 77.3	88.0 95.8 97.7	84.3 87.0 94.3	0.547
	2651 (SHHS2)		5 15 30	94.4 88.8 87.8	67.5 87.7 94.4	89.7 88.2 93.2	**0.612**
Rolón et al. (2020) [21]	954 (SHHS2)	AHI estimation after using structured dictionary learning and Multi-layer perceptron to detect apneic events	15	89.1	86.7	87.9	nd
This study	5793 (SHHS1)	AHI estimation using LSBoost	5 15 30	93.8 87.0 82.2	56.3 84.1 96.3	89.2 85.3 94.6	**0.561**
	2647 (SHHS2)		5 15 30	98.7 95.2 89.8	32.8 69.5 92.0	87.2 81.1 91.6	0.478
Referral							
Olson et al. (1999) [62]	793	Classification of sleep apnea using Delta index (univariate)	5 15 30	82.7 88.5 92.6	54.2 39.6 34.1	nd 67.1* nd	nd
Rofail et al. (2010) [65]	72	*ODI3* direct assessment (univariate)	5 30	63.0 90.0	83.0 88.0	69.5* 88.9*	nd
Gumb et al. (2018) [22]	178	*ODI4* direct assessment (univariate)	5	88.0	74.5	nd	nd
Gutiérrez-Tobal et al. (2019) [17]	320	OSA severity classification using AdaBoost.M2	5 15 30	96.6 92.5 88.9	50.0 73.5 65.5	92.9 87.4 78.7	0.479
Álvarez et al. (2020) [61]	239	AHI estimation using Support Vector Machines	5 15 30	97.8 97.3 89.4	16.7 54.6 95.9	92.7 87.5 92.7	0.610
This study	322 (RHUH)	AHI estimation using LSBoost	5 15 30	99.0 85.5 86.5	63.6 93.2 96.6	96.6 87.6 91.9	**0.663**
SHHS evaluation using other signals (non-referral)							
Van Steenkiste et al. (2019) [66] Abd. Respiration	2100 (SHHS1)	AHI estimation after using Long-short Term Memory networks to detect apneic events	5 15 30	99.3 91.0 67.2	5.9 37.6 81.5	97.8 78.8 76.0	0.329
Olsen et al. (2020) [67] ECG	9704 (SHHS and others)	AHI estimation after using Gated Recurrent Units networks to detect apneic events	5 30	99.1 69.1	32.0 95.5	95.7 89.2	nd
Uddin et al. (2021) [68] Airflow and SpO_2	988(SHHS1)	AHI estimation after using a new ad-hoc automatic algorithm to detect apneic events	5 15 30	98.9 94.7 87.8	60.0 88.5 98.2	90.7 91.0 96.7	nd

$^+$nd: not enough data to estimate; *estimated from reported data.

services (low pre-test probability) or specialized sleep facilities (high pre-test probability). In a context in which approximately 80% of moderate and severe patients remain undiagnosed [69], a reasonable purpose in a primary care setting is to conduct a protocol to screen as many hidden OSA positive subjects as possible. The high Se (>93%) and PPV (>87%) reached by our model in 5 e/h for both SHHS1t and SHHS2 groups suggest the suitability of our model for this task. Moreover, the decision curves showed that, when used to establish the non-referral subjects that should undergo PSG, our model produces higher net benefit than not conducting any protocol for almost any probability threshold a clinician would consider. This is an important result as sending no one to PSG is the current strategy for non-referral subjects in most healthcare systems [70]. It also showed that our model reaches higher net benefit than sending all the subjects to PSG for any probability thresholds above 0.39. However, this result is more trivial as sending all non-referral subjects to PSG is a unapproachable strategy due to the limited availability of sleep facilities and the high prevalence of the disease [11]. Notice that, when using our model to screen patients in non-referral populations, the costs associated to the diagnostic tests would not be reduced but increased by the costs of at-home oximetry. Further studies should address whether these costs may be compensated by a potential less utilization of healthcare systems by patients who would not develop morbidities associated to OSA.

On the other hand, in a specialized sleep unit, a reasonable purpose for our model would be to use it as a surrogate for the PSG, as determined by the high OSA probability context. The appropriateness of our proposal for this goal is first supported by the high values reached in both Se and Sp in our RHUH group, especially for the AHI thresholds 15 e/h and 30 e/h (Se > 85% and Sp >93%). Moreover, in the decision curve analysis, to surrogate PSG is equivalent to recommend medical treatment for the referral subjects that the model considers they are suffering from the disease. In this way, above a probability threshold of 0.39, our model achieves higher net benefit than a strategy based on 'treat all' subjects. As expected, it also reached higher net benefit than 'treat no one' for all thresholds, however this would be an impractical approach in this high pretest probability scenario. Moreover, we introduced in the analysis the relative cost of the standard PSG and the at-home oximetry test used to obtain the data for our proposal. Importantly, the LSBoost-based model showed higher net benefit than PSG in the most realistic scenarios, that is, when a clinician would be willing to conduct few PSGs to confirm that a high pretest probability subject is finally positive. No other 'test harms' were considered as they are difficult to measure objectively. However, we hypothesize that adding aspects such as patient discomfort during the test, need to move to specialized sleep units, accelerated patient access to diagnosis and treatment, and in-lab lack of representativeness of the habitual sleep environment, would increase the relative net benefit of our proposal comparing to PSG. Contrary, unnecessary treatment for false positive subjects would act in the opposite sense. In this regard, it is noteworthy to mention that only mild or moderate side effects are usually reported in the most widespread treatment approaches [71]. Nevertheless, future ad-hoc studies should focus on providing a more comprehensive view on all these aspects.

4.5. Limitations and future work

We have already discussed the need for future evaluation of the performance of our model in the presence of co-morbidities, and the need for further analysis of the extracted SpO_2 information to find associations with other common OSA events. Another limitation is the sample size of the referred database. Although it is one of the largest among the studies focused on SpO_2 recordings, it is remarkably smaller than the SHHS dataset. Therefore, future assessment of larger referral databases would help match the statistical power of our results on referral and non-referral cohorts. In addition, our validation strategy led to a smaller proportion of subjects for training than for testing. It was the result of trying to avoid biases in the model AHI estimation. However, using more training instances could derive into even more accurate models. There is also a need for assessing recordings from younger subjects as only 25% of those included in RHUH dataset are less than 46 years old and none from the SHHS dataset are less than 40 years of age. However, OSA is known to increase its prevalence with age, with only around 1.2% of subjects below 44 years old presenting AHI $\geq$ 5 e/h [72]. Similarly, although the SHHS dataset provides race data, it is only detailed for white and black subjects. Moreover, only white subjects were included in the RHUH dataset as a result of the natural demographics in the area of Valladolid (Spain). Therefore, a comprehensive assessment of our proposal involving subjects from other ethnicities is still pending. In this sense, despite significant differences in the proportion of black subjects included in our SHHS1 training set comparing with our SHHS1 test set, our model rightly classified a similar proportion of black and all other races subjects (65.5% *vs.* 70.3%), showing non-significant *p*-values (> 0.01 in Fisher's exact test). The current use of the AHI thresholds to predict OSA-related adverse outcomes or mortality is also under discussion [73,74]. Although it is accepted as the main diagnostic option to establish OSA and its different severity categories [73], it is also as true that there is an active search for improving the AHI capability to predict the related risks [73,74]. In this regard, one strength of our regression model is that changes in AHI thresholds could be easily adopted in our AHI estimations. Future research, however, would be needed to assess to what extent our estimated AHI is sensitive to OSA-related negative consequences different from respiratory events. In addition, a comprehensive cost-effectiveness study would be also useful to complement our findings. Finally, the investigation of new automated diagnostic techniques is another future goal. Deep learning algorithms could be interesting alternatives at the cost of reduced model interpretability.

5. Conclusions

Our LSBoost-based model exhibits very high performance in automatic OSA diagnosis when compared with the extant literature focused on referral or non-referral cohorts. The model also informed that some of the characteristics of the SpO_2 signal can counteract the tendency of the *ODI3* to underestimate OSA severity. According to our findings, we conclude that our approach can be used to derive clinically valuable protocols to screen OSA patients attending primary care services or avoid full PSGs in specialized sleep facilities, thus demonstrating the usefulness of the SpO_2 signal obtained at home.

Funding

This work was supported by 'Ministerio de Ciencia, Innovación y Universidades' and 'European Regional Development Fund (FEDER)' under projects DPI2017-84280-R and RTC-2017-6516-1, by Sociedad Española de Neumología y Cirugía Torácica (SEPAR) under project 649/2018, Sociedad Española de Sueño (SES) under project "Beca de Investigación SES 2019", by 'European Commission' and 'FEDER' under projects 'Análisis y correlación entre el genoma completo y la actividad cerebral para la ayuda en el diagnóstico de la enfermedad de Alzheimer' and 'Análisis y correlación entre la epigenética y la actividad cerebral para evaluar el riesgo de migraña crónica y episódica en mujeres' ('Cooperation Programme Interreg V-A Spain-Portugal POCTEP 2014–2020'), and by 'CIBER en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)' through 'Instituto de Salud Carlos III' co-funded with FEDER funds. D. Álvarez is supported by a "Ramón y Cajal" grant (RYC2019-028566-I) from the 'Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación' co-funded by the European Social Fund. Vaquerizo-Villar was in receipt of a "Ayuda para contratos predoctorales para la Formación de Profesorado Universitario (FPU)" grant from the "Ministerio de Educación, Cultura y Deporte" (FPU16/02938). Funders had no role in the design of the study, nor in the collection, analysis, or interpretation of data, nor in manuscript preparation. LKG and DG are supported in part by National Institutes of Health grants HL130984 and HL140548 and a University of Missouri Tier 2 grant.

CRediT authorship contribution statement

Gonzalo C. Gutiérrez-Tobal: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Daniel Álvarez:** Software, Validation, Formal analysis, Writing - review & editing. **Fernando Vaquerizo-Villar:** Software, Formal analysis, Methodology, Writing - review & editing. **Andrea Crespo:** Data curation, Resources, Writing - review & editing. **Leila Kheirandish-Gozal:** Data curation, Resources, Writing - review & editing, Funding acquisition. **David Gozal:** Conceptualization, Resources, Writing - review & editing, Supervision, Funding acquisition. **Félix del Campo:** Conceptualization, Resources, Writing - review & editing, Supervision, Funding acquisition. **Roberto Hornero:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank the original researchers from The Sleep Heart Health Study (SHHS), which was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, RFP 75N92019R002).

References

[1] R.B. Berry, R. Brooks, C. Gamaldo, S.M. Harding, R.M. Lloyd, S.F. Quan, M.T. Troester, B.V. Vaughn, AASM scoring manual updates for 2017 (version 2.4), J. Clin. Sleep Med. 13 (2017) 665–666.

[2] L.J. Epstein, D. Kristo, P.J. Strollo, N. Friedman, A. Malhotra, S.P. Patil, K. Ramar, R. Rogers, R.J. Schwab, E.M. Weaver, M.D. Weinstein, Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults, J. Cinical Sleep Med. 5 (2009) 263–276.

[3] F. Lopez-Jimenez, F.H. Sert Kuniyoshi, A. Gami, V.K. Somers, Obstructive sleep apnea: Implications for cardiac and vascular disease, Chest 133 (2008) 793–804, http://dx.doi.org/10.1378/chest.07-0800.

[4] S. Bailly, M. Destors, Y. Grillet, P. Richard, B. Stach, I. Vivodtzev, J.F. Timsit, P. Lévy, R. Tamisier, J.L. Pépin, Obstructive sleep apnea: A cluster analysis at time of diagnosis, PLoS One (2016) http://dx.doi.org/10.1371/journal.pone.0157318.

[5] D. Lacedonia, G.E. Carpagnano, G. Patricelli, M. Carone, C. Gallo, I. Caccavo, R. Sabato, A. Depalo, M. Aliani, A. Capozzolo, M.P. Foschino Barbaro, Prevalence of comorbidities in patients with obstructive sleep apnea syndrome, overlap syndrome and obesity hypoventilation syndrome, Clin. Respir. J. (2018) http://dx.doi.org/10.1111/crj.12754.

[6] T. Saaresranta, J. Hedner, M.R. Bonsignore, R.L. Riha, W.T. McNicholas, T. Penzel, U. Anttalainen, J.A. Kvamme, M. Pretl, P. Sliwinski, J. Verbraecken, L. Grote, F. Barbé, B. Basoglu, P. Bielicki, Z. Dorkova, P. Escourrou, I. Fietze, C. Esquinas, L. Hayes, M. Kumor, S. Kurki, L. Lavie, P. Levy, C. Lombardi, O. Marrone, J.F. Masa, J.M. Montserrat, G. Parati, A. Pataka, J.L. Pépin, R. Plywaczewski, D. Rodenstein, G. Roisman, S. Ryan, R. Schulz, R. Tkacova, R. Staats, P. Steiropoulos, G. Varoneckas, A. Vitols, H. Vrints, J. Zielinski, Clinical phenotypes and comorbidity in European sleep apnoea patients, PLoS One (2016) http://dx.doi.org/10.1371/journal.pone.0163439.

[7] D.E. Jonas, H.R. Amick, C. Feltner, R. PalmieriWeber, M. Arvanitis, A. Stine, L. Lux, R.P. Harris, Screening for obstructive sleep apnea in adults evidence report and systematic review for the US preventive services task force, JAMA - J. Am. Med. Assoc. (2017) http://dx.doi.org/10.1001/jama.2016.19635.

[8] M.D. Ghegan, P.C. Angelos, A.C. Stonebraker, M.B. Gillespie, Laboratory versus portable sleep studies: A meta-analysis, Laryngoscope 116 (2006) 859–864, http://dx.doi.org/10.1097/01.mlg.0000214866.32050.2e.

[9] F.R. de Almeida, N.T. Ayas, R. Otsuka, H. Ueda, P. Hamilton, F.C. Ryan, A.A. Lowe, Nasal pressure recordings to detect obstructive sleep apnea, Sleep Breath 10 (2006) 62–69, http://dx.doi.org/10.1007/s11325-005-0042-x.

[10] A. Pinho, N. Pombo, B.M.C. Silva, K. Bousson, N. Garcia, Towards an accurate sleep apnea detection based on ECG signal: The quintessential of a wise feature selection, Appl. Soft Comput. J. (2019) http://dx.doi.org/10.1016/j.asoc.2019.105568.

[11] A.V. Benjafield, N.T. Ayas, P.R. Eastwood, R. Heinzer, M.S.M. Ip, M.J. Morrell, C.M. Nunez, S.R. Patel, T. Penzel, J.L.D. Pépin, P.E. Peppard, S. Sinha, S. Tufik, K. Valentine, A. Malhotra, Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis, Lancet Respir. Med. (2019) http://dx.doi.org/10.1016/S2213-2600(19)30198-5.

[12] J. Bennett, W. Kinnear, Sleep on the cheap: The role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome, Thorax 54 (1999) 958, http://dx.doi.org/10.1136/thx.54.11.958.

[13] F. del Campo, A. Crespo, A. Cerezo-Hernández, G.C. Gutiérrez-Tobal, R. Hornero, D. Álvarez, Oximetry use in obstructive sleep apnea, Expert Rev. Respir. Med. (2018) http://dx.doi.org/10.1080/17476348.2018.1495563.

[14] D. Álvarez, R. Hornero, J. Víctor Marcos, F. Delcampo, Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis, IEEE Trans. Biomed. Eng. 57 (2010) 2816–2824, http://dx.doi.org/10.1109/TBME.2010.2056924.

[15] J.V. Marcos, R. Hornero, D. Álvarez, M. Aboy, F. Del Campo, Automated prediction of the apnea-hypopnea index from nocturnal oximetry recordings, IEEE Trans. Biomed. Eng. 59 (2012) 141–149, http://dx.doi.org/10.1109/TBME.2011.2167971.

[16] D.S. Morillo, N. Gross, A. León, L.F. Crespo, Automated frequency domain analysis of oxygen saturation as a screening tool for SAHS, Med. Eng. Phys. 34 (2012) 946–953, http://dx.doi.org/10.1016/j.medengphy.2011.10.015.

[17] G.C. Gutiérrez-Tobal, D. Álvarez, A. Crespo, F. del Campo, R. Hornero, Evaluation of machine-learning approaches to estimate sleep apnea severity from at-home oximetry recordings, IEEE J. Biomed. Heal. Inform. 23 (2019) 882–892.

[18] M. Deviaene, D. Testelmans, B. Buyse, P. Borzée, S. Van Huffel, C. Varon, Automatic screening of sleep apnea patients based on the SpO 2 signal, IEEE J. Biomed. Heal. Inform. (2019) http://dx.doi.org/10.1109/JBHI.2018.2817368.

[19] G. Schlotthauer, L.E. Di Persia, L.D. Larrateguy, D.H. Milone, Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry, Med. Eng. Phys. (2014) http://dx.doi.org/10.1016/j.medengphy.2014.05.008.

[20] R.E. Rolón, L.D. Larrateguy, L.E. Di Persia, R.D. Spies, H.L. Rufiner, Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea–hypopnea detection, Biomed. Signal Process. Control. (2017) http://dx.doi.org/10.1016/j.bspc.2016.12.013.

[21] R.E. Rolon, I.E. Gareis, L.D. Larrateguy, L.E. Di Persia, R.D. Spies, H.L. Rufiner, Automatic scoring of apnea and hypopnea events using blood oxygen saturation signals, Biomed. Signal Process. Control. (2020) http://dx.doi.org/10.1016/j.bspc.2020.102062.

[22] T. Gumb, A. Twumasi, S. Alimokhtari, A. Perez, K. Black, D.M. Rapoport, J. Sunderram, I. Ayappa, Comparison of two home sleep testing devices with different strategies for diagnosis of OSA, Sleep Breath (2018) http://dx.doi.org/10.1007/s11325-017-1547-9.

[23] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Statist. (2001) http://dx.doi.org/10.2307/2699986.

[24] G.C. Gutierrez-Tobal, D. Alvarez, F. Del Campo, R. Hornero, Utility of AdaBoost to detect sleep apnea-hypopnea syndrome from single-channel airflow, IEEE Trans. Biomed. Eng. 63 (2016) 636–646, http://dx.doi.org/10.1109/TBME.2015.2467188.

[25] Y. Wang, D. Wang, N. Geng, Y. Wang, Y. Yin, Y. Jin, Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection, Appl. Soft Comput. J. (2019) http://dx.doi.org/10.1016/j.asoc.2019.01.015.

[26] C. Fan, B. Hou, J. Zheng, L. Xiao, L. Yi, A surrogate-assisted particle swarm optimization using ensemble learning for expensive problems with small sample datasets, Appl. Soft Comput. J. (2020) http://dx.doi.org/10.1016/j.asoc.2020.106242.

[27] J. Solé-Casals, C. Munteanu, O.C. Martín, F. Barbé, C. Queipo, J. Amilibia, J. Durán-Cantolla, Detection of severe obstructive sleep apnea through voice analysis, Appl. Soft Comput. J. (2014) http://dx.doi.org/10.1016/j.asoc.2014.06.017.

[28] P. Bühlmann, B. Yu, Boosting with the L 2 loss, J. Amer. Statist. Assoc. (2003) http://dx.doi.org/10.1198/016214503000125.

[29] P. Bühlmann, T. Hothorn, Boosting algorithms: Regularization, prediction and model fitting, Stat. Sci. (2007) http://dx.doi.org/10.1214/07-STS242.

[30] J.H. Friedman, J.J. Meulman, Multiple additive regression trees with application in epidemiology, Stat. Med. (2003) http://dx.doi.org/10.1002/sim.1501.

[31] Z. Xu, G.C. Gutiérrez-Tobal, Y. Wu, L. Kheirandish-Gozal, X. Ni, R. Hornero, D. Gozal, Cloud algorithm-driven oximetry-based diagnosis of obstructive sleep apnoea in symptomatic habitually snoring children, Eur. Respir. J. (2019) http://dx.doi.org/10.1183/13993003.01788-2018.

[32] G.Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, S. Redline, The national sleep research resource: Towards a sleep data commons, J. Am. Med. Inform. Assoc. (2018) http://dx.doi.org/10.1093/jamia/ocy064.

[33] S.F. Quan, B.V. Howard, C. Iber, J.P. Kiley, F.J. Nieto, G.T. O'Connor, D.M. Rapoport, S. Redline, J. Robbins, J.M. Samet, et al., The sleep heart health study: Design, rationale, and methods, Sleep 20 (1997) 1077–1085.

[34] S. Redline, M.H. Sanders, B.K. Lind, S.F. Quan, C. Iber, D.J. Gottlieb, W.H. Bonekat, D.M. Rapoport, P.L. Smith, J.P. Kiley, Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep heart health research group, Sleep (1998).

[35] R.B. Berry, R. Budhiraja, D.J. Gottlieb, D. Gozal, C. Iber, V.K. Kapur, C.L. Marcus, R. Mehra, S. Parthasarathy, S.F. Quan, 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. 8 (2012) (2007) 597–619.

[36] U.J. Magalang, J. Dmochowski, S. Veeramachaneni, A. Draw, M.J. Mador, A. El-Solh, B.J.B. Grant, Prediction of the apnea-hypopnea index from overnight pulse oximetry, Chest 124 (2003) 1694–1701, http://dx.doi.org/10.1378/chest.124.5.1694.

[37] D. Álvarez, R. Hornero, D. Abásolo, F. Del Campo, C. Zamarrón, Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection, Physiol. Meas. (2006) http://dx.doi.org/10.1088/0967-3334/27/4/006.

[38] A. Garde, P. Dehkordi, W. Karlen, D. Wensley, J.M. Ansermino, G.A. Dumont, Development of a screening tool for sleep disordered breathing in children using the phone oximeterTM, PLoS One 9 (2014) http://dx.doi.org/10.1371/journal.pone.0112959.

[39] A. Crespo, D. Álvarez, G.C. Gutiérrez-Tobal, F. Vaquerizo-Villar, V. Barroso-García, M.L. Alonso-Álvarez, J. Terán-Santos, R. Hornero, F. del Campo, Multiscale entropy analysis of unattended oximetric recordings to assist in the screening of paediatric sleep apnoea at home, Entropy 19 (2017) 284, http://dx.doi.org/10.3390/e19060284.

[40] T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda, M. Hirano, Quantification of EEG irregularity by use of the entropy of the power spectrum, Electroencephalogr. Clin. Neurophysiol. 79 (1991) 204–210.

[41] M.T. Martin, A. Plastino, O.A. Rosso, Statistical complexity and disequilibrium, Phys. Lett. Sect. A Gen. At. Solid State Phys. 311 (2003) 126–132, http://dx.doi.org/10.1016/S0375-9601(03)00491-2.

[42] W.K. Wootters, Statistical distance and Hilbert space, Phys. Rev. D (1981) http://dx.doi.org/10.1103/PhysRevD.23.357.

[43] R. Hornero, L. Kheirandish-Gozal, G.C. Gutiérrez-Tobal, M.F. Philby, M.L. Alonso-Álvarez, D. Alvarez, E.A. Dayyat, Z. Xu, Y.S. Huang, M.T. Kakazu, A.M. Li, A. Van Eyck, P.E. Brockmann, Z. Ehsan, N. Simakajornboon, A.G. Kaditis, F. Vaquerizo-Villar, A.C. Sedano, O.S. Capdevila, M. Von Lukowicz, J. Terán-Santos, F. Del Campo, C.F. Poets, R. Ferreira, K. Bertran, Y. Zhang, J. Schuen, S. Verhulst, D. Gozal, Nocturnal oximetry-based evaluation of habitually snoring children, Am. J. Respir. Crit. Care Med. 196 (2017) 1591–1598, http://dx.doi.org/10.1164/rccm.201705-0930OC.

[44] M.E. Cohen, D.L. Hudson, New chaotic methods for biomedical signal analysis, IEEE EMBS Int. Conf. Inf. Technol. Appl. Biomed. (2000) 123–128, http://dx.doi.org/10.1109/ITAB.2000.892363.

[45] J.S. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol. Heart Circ. Physiol. 278 (2000) H2039–H2049.

[46] M. Costa, A.L. Goldberger, C.K. Peng, Multiscale entropy analysis of complex physiologic time series, Phys. Rev. Lett. (2002) http://dx.doi.org/10.1103/PhysRevLett.89.068102.

[47] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Trans. Inf. Theory 22 (1976) 75–81, http://dx.doi.org/10.1109/TIT.1976.1055501.

[48] R. Bruña, J. Poza, C. Gómez, M. García, A. Fernández, R. Hornero, Analysis of spontaneous MEG activity in mild cognitive impairment and Alzheimer's disease using spectral entropies and statistical complexity measures, J. Neural Eng. 9 (2012) 36007, http://dx.doi.org/10.1088/1741-2560/9/3/036007.

[49] D. Álvarez, R. Hornero, M. García, F. del Campo, C. Zamarrón, 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. 41 (2007) 13–24, http://dx.doi.org/10.1016/j.artmed.2007.06.002.

[50] C. Zamarrón Sanz, P.V. Romero, J.R. Rodriguez, F. Gude, Oximetry spectral analysis in the diagnosis of obstructive sleep apnoea, Clin. Sci. (1999) http://dx.doi.org/10.1042/CS19980367.

[51] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Anim. Ecol. (2008) http://dx.doi.org/10.1111/j.1365-2656.2008.01390.x.

[52] C.C. Chen, H.X. Barnhart, Comparison of ICC and CCC for assessing agreement for data without and with replications, Comput. Statist. Data Anal. 53 (2008) 554–564, http://dx.doi.org/10.1016/j.csda.2008.09.026.

[53] J.M. Bland, D.G. Altman, Statistical methods in medical research measuring agreement in method comparison studies, Stat. Methods Med. Res. 8 (1999) 135–160, http://dx.doi.org/10.1177/096228029900800204.

[54] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, USA, 2011.

[55] A.J. Vickers, E.B. Elkin, Decision curve analysis: A novel method for evaluating prediction models, Med. Decis. Mak. (2006) http://dx.doi.org/10.1177/0272989X06295361.

[56] A.J. Vickers, B. van Calster, E.W. Steyerberg, A simple, step-by-step guide to interpreting decision curve analysis, Diagnostic Progn. Res. (2019) http://dx.doi.org/10.1186/s41512-019-0064-7.

[57] A.J. Vickers, A.M. Cronin, E.B. Elkin, M. Gonen, Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers, BMC Med. Inform. Decis. Mak. (2008) http://dx.doi.org/10.1186/1472-6947-8-53.

[58] A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests, BMJ (2016) http://dx.doi.org/10.1136/bmj.i6.

[59] J.B. Pietzsch, A. Garner, L.E. Cipriano, J.H. Linehan, An integrated health-economic analysis of diagnostic and therapeutic strategies in the treatment of moderate-to-severe obstructive sleep apnea, Sleep (2011) http://dx.doi.org/10.5665/SLEEP.1030.

[60] L.J. Epstein, G.R. Dorlac, Cost-effectiveness analysis of nocturnal oximetry as a method of screening for sleep apnea-hypopnea syndrome, Chest (1998) http://dx.doi.org/10.1378/chest.113.1.97.

[61] D. Álvarez, A. Cerezo-Hernández, A. Crespo, G.C. Gutiérrez-Tobal, F. Vaquerizo-Villar, V. Barroso-García, F. Moreno, C.A. Arroyo, T. Ruiz, R. Hornero, F. del Campo, A machine learning-based test for adult sleep apnoea screening at home using oximetry and airflow, Sci. Rep. (2020) http://dx.doi.org/10.1038/s41598-020-62223-4.

[62] L.G. Olson, A. Ambrogetti, S.G. Gyulay, Prediction of sleep-disordered breathing by unattended overnight oximetry, J. Sleep Res. (1999) http://dx.doi.org/10.1046/j.1365-2869.1999.00134.x.

[63] G.E. Silva, M.W. An, J.L. Goodwin, E. Shahar, S. Redline, H. Resnick, C.M. Baldwin, S.F. Quan, Longitudinal evaluation of sleep-disordered breathing and sleep symptoms with change in quality of life: The sleep heart health study (SHHS), Sleep (2009) http://dx.doi.org/10.1093/sleep/32.8.1049.

[64] F. Chung, P. Liao, H. Elsaid, S. Islam, C.M. Shapiro, Y. Sun, Oxygen desaturation index from nocturnal oximetry: A sensitive and specific tool to detect sleep-disordered breathing in surgical patients, Anesth. Analg. (2012) http://dx.doi.org/10.1213/ANE.0b013e318248f4f5.

[65] L.M. Rofail, K.K.H. Wong, G. Unger, G.B. Marks, R.R. Grunstein, Comparison between a single-channel nasal airflow device and oximetry for the diagnosis of obstructive sleep apnea, Sleep (2010) http://dx.doi.org/10.1093/sleep/33.8.1106.

[66] T. Van Steenkiste, W. Groenendaal, Di. Deschrijver, T. Dhaene, Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks, IEEE J. Biomed. Heal. Inform. (2019) http://dx.doi.org/10.1109/JBHI.2018.2886064.

[67] M. Olsen, E. Mignot, P.J. Jennum, H.B.D. Sorensen, Robust ECG-based algorithm for Sleep Disordered Breathing detection in large population-based cohorts using an automatic, data-driven approach, Sleep (2019) http://dx.doi.org/10.1093/sleep/zsz276.

[68] M.B. Uddin, C.M. Chow, S.H. Ling, S.W. Su, A novel algorithm for automatic diagnosis of sleep apnea from airflow and oximetry signals, Physiol. Meas. (2021) http://dx.doi.org/10.1088/1361-6579/abd238.

[69] V. Kapur, D.K. Blough, R.E. Sandblom, R. Hert, J.B. De Maine, S.D. Sullivan, B.M. Psaty, The medical cost of undiagnosed sleep apnea, Sleep (1999) http://dx.doi.org/10.1093/sleep/22.6.749.

[70] J.N. Miller, A.M. Berger, Screening and assessment for obstructive sleep apnea in primary care, Sleep Med. Rev. (2016) http://dx.doi.org/10.1016/j.smrv.2015.09.005.

[71] M. Schwartz, L. Acosta, Y.L. Hung, M. Padilla, R. Enciso, Effects of CPAP and mandibular advancement device treatment in obstructive sleep apnea patients: A systematic review and meta-analysis, Sleep Breath (2018) http://dx.doi.org/10.1007/s11325-017-1590-6.

[72] E.O. Bixler, A.N. Vgontzas, T. Ten Have, K. Tyson, A. Kales, Effects of age on sleep apnea in men. I. Prevalence and severity, Am. J. Respir. Crit. Care Med. (1998) http://dx.doi.org/10.1164/ajrccm.157.1.9706079.

[73] T. Penzel, C. Schöbel, I. Fietze, Revise respiratory event criteria or revise severity thresholds for sleep apnea definition? J. Clin. Sleep Med. 11 (2015) 1357–1359.

[74] H. Korkalainen, J. Töyräs, S. Nikkonen, T. Leppänen, Mortality-risk-based apnea–hypopnea index thresholds for diagnostics of obstructive sleep apnea, J. Sleep Res. (2019) http://dx.doi.org/10.1111/jsr.12855.