Universidad de Valladolid

Escuela Técnica Superior de Ingenieros de Telecomunicación

Dpto. de Teoría de la Señal y Comunicaciones e Ingeniería Telemática

Tesis Doctoral

Pattern Recognition Applied to Airflow Recordings to Help in Sleep Apnea-Hypopnea Syndrome Diagnosis

Presentada por **D. Gonzalo César Gutiérrez Tobal** para optar al grado de doctor por la universidad de Valladolid

Dirigida por:
Dr. Roberto Hornero Sánchez y Dr. Daniel Álvarez González

Septiembre 2015
Valladolid, España

Título:	Pattern Recognition Applied to Airflow Recordings to Help in Sleep Apnea-Hypopnea Syndrome Diagnosis
Autor:	D. Gonzalo César Gutiérrez Tobal
Directores:	Dr. Roberto Hornero Sánchez y Dr. Daniel Álvarez González
Dpto.:	Teoría de la Señal y Comunicaciones e Ingeniería Telemática

Tribunal:

Presidente:	Dr. D.
Vocales:	Dr. D.
	Dr. D.
	Dr. D.
Secretario:	Dr. D.

acuerda otorgarle la calificación de

En Valladolid, a

A Pilar, por hacerme sonreir cada vez que pienso en ti.
Y a mis padres, Ana y César, y hermano, Pablo.
Menos mal que estabais ahí.

"...Sueña el rico en su riqueza,
que más cuidados le ofrece;
sueña el pobre que padece
su miseria y su pobreza;
sueña el que a medrar empieza,
sueña el que afana y pretende,
sueña el que agravia y ofende,
y en el mundo, en conclusión,
todos sueñan lo que son,
aunque ninguno lo entiende..."

La vida es sueño, Pedro Calderón de la Barca [23]

Espacio y tiempo que el cuerpo admite
como una fábula del pensamiento,
imaginada, acaso, en el momento,
en el que mi otro Yo, mejor orbite.

Sólo mi pensar, tardío y lento,
aplacará tu juvenil envite,
al sugerir una historia que transite,
al epílogo de un hermoso cuento.

Y hará que la aventura pergeñada,
en el ánimo de una testa lela,
aspire a ser impreso, sin portada.

Es la duda: si el cabo de mi vela
atravesará la noche cerrada,
y esbozar un apunte de novela.

Espacio y tiempo, César Francisco Gutiérrez de Manuel.

Agradecimientos

En primer lugar, me gustaría dar las gracias a mis directores, Dr. Roberto Hornero Sánchez y Dr. Daniel Álvarez González. Es dificil encontrar mentores más capacitados y con mayor dedicación, pero aún más dificil es imaginar personas que demuestren día a día mayor calidad humana. Su dirección, su disponibilidad y su paciencia han sido fundamentales para conlcuir con éxito esta Tesis Doctoral.

También me gustaría dar las gracias a los trabajadores de la unidad del sueño del Hospital Universitario Río Hortega de Valladolid, y muy especialmente al Jefe del servicio de Neumología Dr. Félix del Campo Matías. Sin sus conocimientos médicos y predisposición continua tampoco habría podido llevarse a cabo este estudio. Muchas gracias, Félix, por tu ayuda.

No puedo dejar de acordarme también de mis compañeros del Grupo de Ingeniería Biomédica (GIB). Todos y cada uno de ellos han estado siempre disponibles para ayudarme cuando lo he necesitado y, lo que es más importante, lo han hecho siempre con una sonrisa en la boca. Gracias Roberto, Jesús, María, Carlos, Dani, Víctor, Rebeca, Alejandro, Javier y Luis Fernando, por contribuir a generar un ambiente amable y cercano, capaz de hacerme levantar cada mañana feliz por ir a trabajar. Vosotros hacéis que sienta orgullo de pertenecer al GIB.

Finalmente, quiero dar las gracias a mi familia. Y no hay espacio suficiente en estas páginas para agradecer todo lo que les debo, ni palabras que hagan justicia al cariño, esfuerzo y amor derrochados durante todos estos años. Sencillamente, no puedo pensar en mejores padres ni en mejor hermano. Para terminar, infinitas gracias a tí, Pilar. Por compartir tu vida conmigo y por dejar que comparta la mia contigo. Por ayudarme, por quererme, por sonreirme. Gracias por hacerme feliz. Te quiero.

Abstract

The sleep apnea-hyponea syndrome (SAHS) is a disease characterized by episodes of complete absence (apneas) or significant reduction (hypopneas) of breathing during sleep. The apneic events recurrence leads to inadequate gas exchange which causes hypoxia and hypercapnia, resulting in oxygen saturation drops, periodic arousals, as well as sleep fragmentation. As a consequence, SAHS patients are not able to get restful sleep, which affects their quality of life. Hypersomnolence, decrease in the short-memory function, and depression are some of the daytime symptoms reported by affected people. Additionally, SAHS has been associated with major cardiovascular and metabolic illnesses such as heart failure, stroke, sudden death, and diabetes. Recently, it has been also associated with an increase in cancer incidence. These SAHS consequences make a fast diagnosis the key action to improve health and quality of life of patients.

SAHS is a very prevalent illness, affecting from 2% to 7% of adult population and up to 6% of children. It is also considered as an underdiagnosed disease, with a growing incidence due to the obesity epidemic present in developed countries. Overnight polysomnography (PSG), conducted in a specialized sleep unit, is the "gold standard" to diagnose SAHS. However, it is technically complex due to the high number of physiological signals to be recorded, costly due to the need for patient's hospitalization, as well as time-consuming due to the offline inspection of the recordings, which is required to reach diagnosis. This is obtained by computing the apnea-hypopnea index (AHI) after carefully reviewing of the recorded signals. Moreover, PSG test deprives patients of their natural sleep environment.

These drawbacks, the high prevalence of SAHS, as well as the limited availability of specialized facilities, have led to the search for new ways to simplify the diagnostic process. One common approach is the analysis of a reduced set of signals among those involved in full PSG. In this Doctoral Thesis it is posed the automatic analysis of single-channel airflow (AF) as a simple and reliable alternative to PSG. In addition, pattern recognition is proposed as the main approach to conduct an automatic SAHS diagnosis, including binary classification (presence or absence of SAHS) as well as determination of SAHS severity degree (multiclass classification and AHI estimation by means of regression). We hypothesize that *it is possible to reduce the complexity of the SAHS diagnostic process by means of an automatic pattern recognition analysis of AF.* Consequently, the general goal of this work is the comprehensive study and assessment of the diagnostic potential of the AF signal as a surrogate for full PSG in SAHS detection.

Our methodology is based on three main steps. First, a feature extraction stage is implemented to obtain information of SAHS from single-channel

AF. Physiological signals are known to behave both in deterministically and chaotically ways. For this reason, different methodologies were used to extract SAHS-related information, such as spectral and non-linear analyses. The purpose of this approaches was the optimum characterization of SAHS by means of obtaining complementary information. The second step is an automatic feature selection stage. The comprehensive analysis conducted in the previous stage may lead to the extraction of useless features for SAHS diagnosis or features sharing similar information than others. Thus, it has been implemented a feature selection stage to eliminate those non-relevant or redundant. Two different approaches have been used for this purpose: the well-known forward-selection backward-elimination method (SLR-FSBE) and the fast correlation-based filter (FCBF) algorithm. The former is a wrapper method since it is closely related to a specific classifier (logistic regression), whereas the latter is a filter since it is independent from subsequent analyses. Finally, the third stage is pattern recognition. In this Doctoral Thesis, it has been used to obtain an automatic SAHS diagnosis by the application of different classification and regression methods to data obtained and selected in previous stages. The main purposes of this step have been determining the presence or absence of SAHS (binary classification), classifying subjects into one out of the four SAHS severity degrees (multiclass classification), and the estimation of AHI (regression). This approach differs from the common approach followed in the state of the art, where the main studies focus on detecting each of the apneic events present in the recordings.

After applying our methodology to single-channel AF, the results showed that our proposal outperformed a classic event-detection algorithm applied to our databases. Thus, in the case of binary classification, an ensemble learning model based on AdaBoost, built with decision trees, reached 89.0% sensitivity (Se), 80.0% specificity (Sp), 86.5% accuracy (Acc), 0.950 area under the receiver-operating characteristics curve (AROC), and 0.672 Cohen's κ, in contrast to the classic event-detection algorithm which obtained 75.8% Se, 54.3% Sp, 64.0% Acc, 0.635 AROC, and 0.286 Cohen's κ. Regarding multiclass classification, another AdaBoost model, built with linear discriminant classifiers, obtained 86.5%, 81.0%, and 82.5% accuracies when evaluated in the AHI cutoffs which establish each of the four SAHS severity degrees (AHI = 5 events/hour, 15 e/h, and 30 e/h). The event-detection algorithm obtained lower statistics for each threshold, reaching 81.0%, 68.3% y 63.5%, respectively. Finally, when applying regression to estimate AHI, an artificial neural network model based on multi-layer perceptron (MLP) obtained an intra-class correlation coefficient (ICC) of 0.849, and 79.7%, 91.5%, 79.7%, and 88.1% diagnostic accuracies for AHI cutoffs = 5 e/h, 10 e/h, 15 e/h y 30 e/h, respectively, each of them associated with corresponding 0.903, 0.956, 0.904, and 0.973 AROC values. By contrast, the event detection algorithm reached 0.840 ICC, and accuracies of 79.7% (0.823 AROC), 78.0% (0.833 AROC), 66.1% (0.867 AROC) y 74.6% (0.982 AROC).

On the other hand, our methodology applied to at-home AF recordings from children showed higher performance than the oxygen desaturation index (ODI), which is commonly used in clinical practice. Additionally, the combination of spectral information from these recordings with ODI achieved 85.9% Se, 87.4% Sp, 86.3% Acc, 0.947 AROC and 0.720 κ.

Our proposal achieved high diagnostic performance comparing with PSG diagnosis, state-of-the-art studies focused on detecting apneic events in other

AF databases, as well as studies reporting similar approaches to ours applied in different PSG signals. Consequently, the main conclusion obtained from this Doctoral Thesis is that pattern recognition methods applied to single-channel AF are useful to improve the automation of the SAHS diagnosis process. Hence, it is also concluded that this process can be reliably simplified by means of the automatic analysis of AF.

Contents

Chapter 1

Introduction

The current Doctoral Thesis focuses on helping the sleep apnea-hypopnea syndrome (SAHS) diagnosis by means of biomedical signal processing methodologies. During the study, several feature extraction and selection procedures, as well as pattern recognition techniques, have been evaluated. This investigation has led to results which have been published, or accepted for publication, in journals indexed in the Journal Citation Reports (JCR) from Thomson Reuters Web of Science™. Specifically, up to four articles were published between December 2012 and March 2015. Additionally, a fifth article has been recently accepted for publication (August 2015). This scientific productivity has allowed writing this work as a compendium of publications.

The thematic consistency of the articles included in the thesis is justified in this introductory chapter. A brief introduction to biomedical engineering and signal processing can be also found. Moreover, there are two sections devoted to SAHS. Section 1.3 explains what SAHS is, its risks and severe consequences, and some related socio-economic issues. Section 1.4 focuses on the diagnosis of SAHS, i. e., the current standard test and its limitations. The latter, indeed, are the cause motivating the research problem. Finally, common alternatives to the standard diagnostic test are described as state of the art.

1.1. Compendium of publications: thematic consistency

SAHS is a highly prevalent disease which affects both health and life's quality of people [128]. In past years, it has become a major concern for the medical community due to its severe consequences and its association with other major illnesses [24, 81]. In spite of the effectiveness of the standard diagnostic test (the polysomnography, PSG), it is complex, costly, and time-consuming. Consequently, the search for simpler methods to diagnose SAHS has became a major goal.

This context is the common thread shared by the articles included in the compendium of publications. The approach followed to simplify the diagnostic test is also a constant. Since the main source of complexity in the PSG test is the need to record multiple physiological signals to diagnose SAHS (up to 32 channels), the common methodology conducted in the articles of the compendium has been the analysis of the information recorded from a reduced set of channels. In this regard, two out of the five articles focused on analyzing the information contained in airflow (AF) recordings acquired with a thermistor [64, 65], including the airflow-related signal respiratory rate variability (RRV). A third article aimed at studying the AF signal obtained from a nasal prong pressure sensor [62], whereas another one focused on the analysis of a cardiac signal, the heart rate variability (HRV) [63]. Finally, the fifth article involved AF (from thermistor) and oxygen saturation (SpO_2) recordings from children [61].

Another of the articles' connections, which is also one of the major novelties of the study, is the analytical strategy conducted on the PSG signals. Through the years, scientists have identified several SAHS-related events affecting different physiological systems and, therefore, being reflected in different biomedical signals. While most of the state-of-the-art works focus on an event-by-event study [14, 29, 95, 116], i. e., they aim at detecting each one of these characteristic events, our approach conducts the analysis of the signals from a general perspective, i. e., we perform a whole-signal characterization.

Finally, as a result of this common analytical approach, the five articles also

share the same methodological framework. Thereby, feature extraction and selection, as well as pattern recognition methodologies have been implemented during the study.

Titles, authors, and abstracts of the articles, as well as the scientific journals where these were published are shown below, sorted chronologically:

- **Linear and nonlinear analysis of airflow recordings to help in sleep apnoea-hypopnoea syndrome diagnosis.** [65]

  Gonzalo C. Gutiérrez-Tobal, Roberto Hornero, Daniel Álvarez, J. Víctor Marcos, Félix del Campo.*Physiological Measurement*, 2012, 33(7), 1261. Impact Factor: 1.496

  Abstract: This paper focuses on the analysis of single channel airflow signal (AF) to help in sleep apnoea-hypopnoea syndrome (SAHS) diagnosis. The respiratory rate variability (RRV) series is derived from AF by measuring time between consecutive breathings. A set of statistical, spectral and non-linear features are extracted from both signals. Then forward stepwise logistic regression procedure (FSLR) is used in order to perform feature selection and classification. Three logistic regression (LR) models are obtained by applying FSLR to features from AF, RRV and both signals simultaneously. The diagnostic performance of single features and LR models is assessed and compared in terms of sensitivity, specificity, accuracy and area under the receiver-operating characteristics curve (AROC). The highest accuracy (82.43%) and AROC (0.903) are reached by the LR model derived from the combination of AF and RRV features. This result suggests that AF and RRV provide useful information to detect SAHS.

- **Pattern recognition in airflow recordings to assist in the sleep apnoea-hypopnoea syndrome diagnosis.** [64]

  Gonzalo C. Gutiérrez-Tobal, Daniel Álvarez, J. Víctor Marcos, Félix del Campo, Roberto Hornero. *Medical & Biological Engineering & Computing*, 2013, 51(12), 1367-1380. Impact Factor: 1.500

  Abstract: This paper aims at detecting sleep apnoea-hypopnoea syndrome (SAHS) from single-channel airflow (AF) recordings. The study involves 148 subjects. Our proposal is based on estimating apnoea-hypopnoea index (AHI) after global analysis of AF, including the investigation of respiratory rate variability (RRV). We exhaustively characterize both AF and RRV by extracting spectral, non-linear and statistical features. Then, the fast correlation-based filter (FCBF) is used to select those relevant and non-redundant. Multiple linear regression (MLR), multi-layer perceptron (MLP) and radial basis functions (RBF) are fed with the features to estimate AHI. A conventional approach, based on scoring apnoeas and hypopnoeas, is also assessed for comparison purposes. An MLP model trained with AF and RRV selected features

achieved the highest agreement with the true AHI (intra-class correlation coefficient = 0.849). It also showed the highest diagnostic ability, reaching 92.5% sensitivity, 89.5% specificity and 91.5% accuracy. This suggests that AF and RRV can complement each other to estimate AHI and help in SAHS diagnosis.

- **Assessment of time and frequency domain entropies to detect sleep apnoea in heart rate variability recordings from men and women.** [63]

  Gonzalo C. Gutiérrez-Tobal, Daniel Álvarez, Javier Gomez-Pilar, Félix del Campo, Roberto Hornero. *Entropy*, 2015, 17(1), 123-141. Impact Factor: 1.502

  Abstract: Heart rate variability (HRV) provides useful information about heart dynamics both under healthy and pathological conditions. Entropy measurements have shown their utility to characterize these dynamics. In this paper, we assess the ability of spectral entropy (SE) and multiscale entropy (MsE) to characterize the sleep apnoea-hypopnea syndrome (SAHS) in HRV recordings from 188 subjects. Additionally, we evaluate eventual differences in these analyses depending on the gender. We found that SE measures in the very low frequency and low frequency bands showed ability to characterize SAHS regardless the gender; and that MsE features may be able to depict gender specificities. Also, SE and MsE showed complementarity to detect SAHS since features from both analyses were automatically selected by the logistic regression-based forward-selection backward-elimination algorithm. Finally, SAHS was modelled through logistic regression (LR) by using optimum sets of selected features. Modelling SAHS by genders reached significant higher performance than doing it in a jointly way. The highest diagnostic ability was reached by LR modelling of SAHS in women, achieving 80.9% sensibility, 89.3% specificity, 85.2% accuracy, and 0.951 area under the ROC curve. Our results show the usefulness of the SE and MsE analyses of HRV to detect SAHS, as well as suggest that, when using HRV, SAHS may be more accurately modelled differentiating by gender.

- **Diagnosis of pediatric obstructive sleep apnea: preliminary findings using automatic analysis of airflow and oximetry recordings obtained at patients' home.** [61]

  Gonzalo C. Gutiérrez-Tobal, M. Luz Alonso-Álvarez, Daniel Álvarez, Félix del Campo, Joaquín Terán-Santos, Roberto Hornero. *Biomedical Signal Processing and Control*, 2015, 18, 401-407. Impact Factor: 1.419

  Abstract: The Obstructive Sleep Apnea Syndrome (OSAS) greatly affects both the health and the quality of life of children. Therefore, an early diagnosis is crucial to avoid their severe consequences. However, the standard diagnostic test (polysomnography, PSG) is time-demanding, complex, and costly. We aim at assessing a new methodology for the pediatric

OSAS diagnosis to reduce these drawbacks. Airflow (AF) and oxygen saturation (SpO2) at-home recordings from 50 children were automatically processed. Information from the spectrum of AF was evaluated, as well as combined with 3% oxygen desaturation index (ODI3) through a logistic regression model. A bootstrap methodology was conducted to validate the results. OSAS significantly increased the spectral content of AF at two abnormal frequency bands below (BW1) and above (BW2) the normal respiratory range. These novel bands are consistent with the occurrence of apneic events and the posterior respiratory overexertion, respectively. The spectral information from BW1 and BW2 showed complementarity both between them and with ODI3. A logistic regression model built with 3 AF spectral features (2 from BW1 and 1 from BW2) and ODI3 achieved (mean and 95% confidence interval): 85.9% sensitivity [64.5-98.7]; 87.4% specificity [70.2-98.6]; 86.3% accuracy [74.9-95.4]; 0.947 area under the receiver-operating characteristics curve [0.826-1]; 88.4% positive predictive value [72.3-98.5]; and 85.8% negative predictive value [65.8-98.5]. The combination of the spectral information from two novel AF bands with the ODI3 from SpO2 is useful for the diagnosis of OSAS in children.

- **Utility of AdaBoost to detect sleep apnea-hypopnea syndrome from single-channel airflow.** [62]

  Gonzalo C. Gutiérrez-Tobal, Daniel Álvarez, Félix del Campo, Roberto Hornero. *IEEE Transactions on Biomedical Engineering, In Press.* Accepted August 2015. Impact Factor: 2.347

  Abstract: Goal: The purpose of this study is to evaluate the usefulness of the boosting algorithm AdaBoost (AB) in the context of the sleep apnea-hypopnea syndrome (SAHS) diagnosis. Methods: We characterize SAHS in single-channel airflow (AF) signals from 317 subjects by the extraction of spectral and non-linear features. Relevancy and redundancy analyses are conducted through the fast correlation-based filter (FCBF) to derive the optimum set of features among them. These are used to feed classifiers based on linear discriminant analysis (LDA) and classification and regression trees (CART). LDA and CART models are sequentially obtained through AB, which combines their performances to reach higher diagnostic ability than each of them separately. Results: Our AB-LDA and AB-CART approaches showed high diagnostic performance when determining SAHS and its severity. The assessment of different apnea-hypopnea index cutoffs using an independent test set derived into high accuracy: 86.5% (5 events/h), 86.5% (10 events/h), 81.0% (15 events/h), and 83.3% (30 events/h). These results widely outperformed those from logistic regression and a conventional event-detection algorithm applied to the same database. Conclusion: Our results suggest that AB applied to data from single-channel AF can be useful to determine SAHS and its severity. Significance: SAHS detection might be simplified through the only use of single-channel AF data.

1.2. Context: biomedical engineering, biomedical signal processing, and pattern recognition

Biomedical Engineering is an interdisciplinary field that focuses on altering, controlling, or understanding biological systems by applying engineering principles [22]. It covers a wide range of industrial and academic activities, including both theoretical and experimental research. One of the greatest Biomedical Engineering benefits is the ability to identify issues and needs in healthcare systems, which can be solved using novel technologies and methodologies [22]. As a consequence, it is also seen as a mean to provide better services with ability to highly improve the life's quality of human beings. Additionally, this is the reason why Biomedical Engineering is involved in all aspects of the development of new medical technologies [22].

Biomedical signal processing is one of the activities included in the Biomedical Engineering field. In human body, different systems produce physiological signals which reflect their behavior. By studying these signals it has been possible to fully or partially explain and identify a wide range of pathological conditions [118]. Most of the time, however, the information contained in biomedical signals is not directly interpretable, and a processing stage is needed in order to provide meaning to the extracted data [118]. Consequently, biomedical signal processing has become an essential tool to extract the hidden clinical meaning from the obtained information. In addition, it has also become basic to develop new automatic diagnostic systems [118].

Pattern recognition, intimately related to machine learning, focuses on the automatic detection of regularities in data by means of computer algorithms [20]. These techniques have experimented great development in recent years. Their purpose is to use the knowledge obtained from the data to being able to automatically classify them into one out of several categories (classification task) or to automatically estimate one or several target continuous variables (regression) [20]. As a consequence, pattern recognition methodologies have been applied to solve a wide variety of problems, including development of new automatic methods to help in the diagnosis of different pathologies.

This doctoral thesis aims at helping in SAHS detection by reducing the complexity of the standard diagnostic test. For this purpose, several biomedical signals from PSG have been analyzed. Novel signal processing and pattern recognition techniques have been developed and assessed as well. Hence, all the above mentioned reflect the framework in which this study is encompassed.

1.3. Sleep Apnea-Hypopnea Syndrome (SAHS)

1.3.1. Definition, symptoms, and risk factors

SAHS is a highly prevalent disease which worsens both the health and the quality of life of affected people [81]. It is characterized by the recurrence of episodes of total absence of airflow (apnea) and/or significant airflow reduction (hypopnea) during sleep [93]. Apneic events are classified as obstructive, central, or mixed according to their origin, the former being the most common [92]. Whereas central events are caused by malfunction of the neural center that controls respiration, obstructive events are due to upper airway obstructions. The presence (obstructive) or absence (central) of respiratory effort is the key element to differentiate between them. Events starting as central and

ending as obstructive are classified as mixed [92]. The apnea-hypopnea index (AHI), i.e. the number of apneic events per hour of sleep time, is the clinical variable used to establish SAHS and its severity.

The occurrence of SAHS is associated with the presence of nocturnal and daytime symptoms. Thus, overnight apneic events cause inadequate gas exchange characterized by hypoxia and hypercapnia, which lead to drops in oxygen saturation, arousals, and sleep fragmentation. Loud snoring and gasping are also frequent among people affected by SAHS [93, 115]. Nocturnal symptoms cause restless sleep which, in turn, leads to daytime symptoms. Thus, daytime hypersomnolence, concentration and short-term memory difficulties, as well as depression, have been reported in SAHS patients [93]. In the case of children, cognitive and behavioral irregularities as well as atypical growth, are frequently present [59].

The major risk factors for SAHS are aging, male sex, and obesity [39, 49, 93]. Anatomical and clinical factors, such as deposition of fat around the pharynx and deterioration of the genioglossus negative pressure reflex, are suggested as the cause for an increased upper airway collapsibility with age, which leads to a higher number of apneic events. The larger amount of fat deposited around pharynx of men, as well as their longer pharyngeal way, have been also suggested as the cause for gender differences [39]. Moreover, hormonal reasons are argued to explain higher risk of SAHS for men and post- menopausal women [39, 93]. There exist other risk factors such as congestive heart failure, atrial fibrillation, type 2 diabetes, stroke, or even alcohol use [49]. However, it is recognized that obesity plays a key role in the development of SAHS and the increase of its severity [39, 49]. Fat deposition in abdominal and pharyngeal areas have been suggested as possible reasons for ventilatory control instabilities and increased upper airway collapsibility, respectively. Additionally, functional impairment in upper airway muscles has been related to obesity [39]. Each of these issues would lead to increase the number of apneic events. Furthermore, the clear association between obesity and SAHS has led to point the expanding epidemic of overweight as one of the major reasons for the high prevalence of SAHS, particularly in western countries [128].

1.3.2. Prevalence

Prevalence of SAHS is high. Several studies have focused on estimating it in different countries, ethnic groups, sexes, and age groups, including children [59, 100, 128]. Beyond the different exact figures reported for prevalence, all of them agree in pointing out that there exists a high number of people affected and, very often, not diagnosed [100, 128]. Thus, global prevalence in adults has been conservatively established in the range 2-7% [100, 128]. Men are known to be more affected than women. Thereby, adult male population present a prevalence in the range 4-14%, being 1-7% in the case of adult female population [100, 128]. When considering population older than age 65 years, prevalence increases beyond 50% for both sexes [128]. Moreover, it has been reported that up to 6% of children may be also affected [59].

The prevalence of SAHS in Spain is not lower than the global trend. Indeed, there exist studies reporting up to 14% men and 7% women affected by SAHS. This prevalence is higher than figures reported in studies from other countries [37]. In 2005, however, the Spanish Group for Sleep (Grupo Español del Sueño, GES) estimated that SAHS affected between 1.200.000 and 2.150.000 people in

our country, which would be equivalent to a prevalence ranging 2.7-4.9% [38]. The same report established that only 5-9% of these people had been diagnosed and treated.

1.3.3. Consequences and related illnesses

SAHS has been associated with a wide range of other major illnesses and pathological conditions. Indeed, some of them have already been mentioned as risk factors. Congestive heart failure (CHF) has been linked to recurrent obstructive and central apneic events. These cause hypoxia and high blood pressure during sleep, which combined with daytime hypertension increase the chances of such a cardiac failure. In addition, CHF may contribute to worsen SAHS by facilitating the upper airway collapsibility through periodic breathing, a characteristic pathological respiratory pattern to which these patients are predisposed [115]. A bidirectional relationship has been also indicated between SAHS and obesity [81]. As stated above, pharyngeal and abdominal fat contribute to the occurrence of apneas and hypopneas. Additionally, it has been also suggested that SAHS worsens obesity. Although the underlying mechanisms involved are not clear, there are evidence of SAHS patients gaining greater weight than no SAHS subjects. The occurrence of cardiac arrhythmias has been usually associated with SAHS as well [81, 115]. Thereby, mechanisms and signs such as high sympathetic activity, hypoxemia, or systemic inflammation, frequently seen in SAHS patients, can trigger atrial fibrillation. Hypertension, stroke, and sudden cardiac death are other major pathologies linked to SAHS at different degrees [81]. In recent years, it has been also established a relationship between SAHS and an increase in cancer incidence [24].

Beyond comorbidities, people suffering from SAHS are also exposed to a higher risk of having motor-vehicle collisions and occupational accidents. A significant proportion of traffic crashes, costs, and deaths has been linked to SAHS. In this regard, it has been estimated that SAHS treatment could save 70% of both costs and lifes [114]. Similarly, snorers with daytime hypersomnolence, two common signs associated with SAHS, present a significant higher risk of occupational accidents [80].

1.4. SAHS diagnosis

1.4.1. Polysomnography (PSG)

The high prevalence of SAHS, its severe consequences, as well as the effectiveness of the treatment, make the diagnosis process the key element to improve the health and quality of life of affected people. SAHS is diagnosed by means of overnight polysomnography test (PSG), which acts as "gold standard" [92, 93]. During PSG, multiple physiological signals from patients are monitored. Thus, up to 32 channels may be recorded, including electrocardiogram (ECG), electroencephalogram (EEG), respiratory effort, oxygen saturation of blood (SpO_2), and airflow (AF). Consequently, patients have to spend a whole night in a sleep unit, where specialists take care of them as well as supervise the course of PSG. After the test, the recordings need an offline inspection in order to compute the AHI, which determines the presence and severity of SAHS.

AHI is obtained as the average number of apneic events (apneas and hypopneas) per hour of sleep [70]. Consequently, the main objective of clinicians

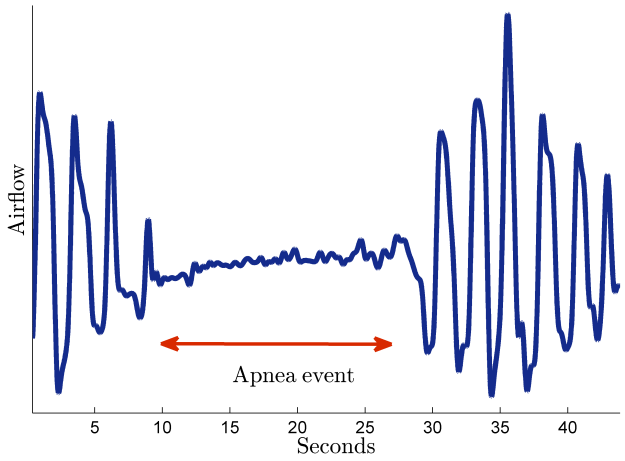

Figure 1.1: Example of an apnea event in AF

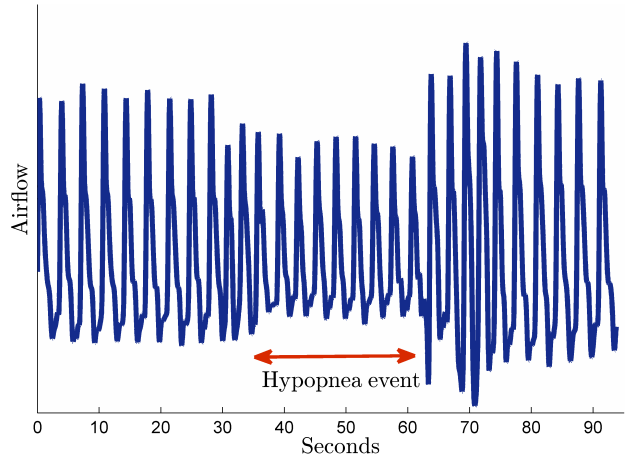

Figure 1.2: Example of an hypopnea event in AF

when inspecting the physiological signals is to detect and score each of these events. In the case of adults, the American Academy of Sleep Medicine (AASM) define an apnea as a minimum of 90% reduction in the airflow, acquired with an oronasal thermal sensor, and which lasts ten seconds or more. Similarly, an hypopnea is defined as a minimum of 30% reduction in the airflow, obtained with a nasal pressure sensor, lasting ten seconds or more and accompanied by a minimum of 3% drop in the oxygen saturation or/and an arousal [18]. Figures 1.1 and 1.2 display true examples of an apnea and a hypopnea event in AF, respectively. In the case of children, the 10-second criterion of both apneas and hypopneas is replaced by a minimum duration equivalent to two missed breathing cycles [18].

1.4.2. Limitations of the PSG

The high number of signals required to monitor patients during PSG leads to the need of complex acquisition equipments. This complexity, along with the qualified personnel needed overnight, makes PSG an expensive test [17]. PSG also implies that patients have to spend one night outside their usual sleep environment with many sensors attached in their bodies, which is cumbersome [35]. This may result in recording sleep patterns not representative of their usual sleep behavior [17], increasing the chances that a single patient requires more than one PSG. On the other hand, clinicians need to conduct an offline inspection of all the recordings in order to reach each diagnosis, which requires several hours. Summarizing, although the effectiveness of PSG is well-known, it is complex, costly, time-consuming, and may not represent accurately the actual sleep behavior of patients. Additionally, there exists a lack of availability of specialized laboratories to carry out the PSG test [46, 55]. This issue, along with the high prevalence of SAHS and the high number of not-diagnosed but affected people, leads to long waiting time and difficulties to access diagnosis and treatment [35, 55].

1.5. Alternatives to PSG: approaches and state of the art

PSG drawbacks have led to the search for different alternatives to diagnose SAHS [17, 35, 48, 55]. Most of the effort has been put into simplifying the diagnostic test. Reducing PSG complexity is the key factor to decrease cost, patient's cumbersome, and time delay, as well as gives the chance to develop home diagnosis portable devices [48, 55]. One direct way to simplify the diagnostic test is to analyze a reduced set of signals instead of the whole set used in PSG. According to the signals analyzed, the equipment used in sleep studies is categorized as follows [48]:

- **Type 1 or standard PSG.** This is the conventional PSG test, as described previously, to which the other types are compared.

- **Type 2 or comprehensive portable PSG.** These devices incorporate 7 seven channels at least, including EEG, electro-oculogram, chin electromyography, ECG or heart rate, AF, respiratory effort, and SpO_2.

- **Type 3 or modified portable sleep apnea testing.** Equipment included in this category incorporate a minimum of 4 channels: ventilation or AF (2 channels of respiratory movement or respiratory movement and AF), ECG or heart rate, and SpO_2.

- **Type 4 or continuous single-bioparameter or dual-bioparameter recording.** Most of devices of this type measure a single parameter or two parameters, frequently involving AF and/or SpO_2. However, all the equipment not meeting Type 3 criteria is classified as Type 4.

1.5.1. Signals commonly studied to simplify SAHS diagnosis

The AF signal is directly affected by the recurrence of apneas and hypopneas. Hence, AF carries crucial information about SAHS and, consequently, its study is a natural way of dealing with the problem of simplifying SAHS diagnosis. In past years, a lot of studies focused on evaluating AF as a reliable alternative to full PSG. One common approach has been the assessment of some specific type-4 portable device, comparing diagnosis from PSG with the corresponding one from the new appliance. Some of the proposals recently evaluated have been SleepStrip™ [116], SleepCheck [35], RUSleeping™ [58], ApneaLink™ [14, 28, 41, 91], and Flow Wizard [110, 112, 127]. Regardless they were conducted in sleep laboratories or at patient's' home, these studies compared the AHI from PSG with an apneic event index derived from detecting apneas and hypopneas in the alternative AF signal. Another usual approach focuses on developing new automatic methodologies with ability to properly detect apneic events. The objective is not to obtain a diagnosis but to accurately detect some events of interest like obstructive or central apneas, hypopneas, etc. In this regard, recent studies have applied signal processing and machine learning methodologies to the AF recordings obtained as part of the PSG protocol. Thus, Varady et al. [122] assessed four artificial neural networks (ANN) fed with 16-second epochs from AF and/or thoracic movement signals in order to classify each of them as normal breathing, apnea segment, or hypopnea segment. Álvarez-Estévez and Moret-Bonillo [12] applied a fuzzy algorithm to AF, SpO_2, and respiratory movements recordings to detect breathing events and classify them into apneas or hypopneas. Similarly, Koley and Dey [76] detected apneas and hypopneas by feeding a support vector machine (SVM) classifier with features extracted from AF. In the same way, Han et al. [67] used a simple detection algorithm based on the mean magnitude of the second derivatives of AF to detect only apnea events. A third typical approach completes the latter by evaluating the diagnostic ability of an event index derived from the automatic detection. This is the case of Nakano et al. [89]. They detected apneas and hypopneas in AF by evaluating the power spectral density of short-time windows. An event index was obtained on the basis of these detections, and its diagnostic ability was assessed using the AHI from PSG as reference. Rathnayake et al. [102] also assessed a surrogate of the AHI. They characterized apneic events by the use of recurrence plots and detected them with the help of a mixture discriminant analysis. All the above approaches rely on the definition of apnea and hypopnea, which is a frequent topic of discussion, as well as discards the information of the AF signal other than the apneic events themselves.

Overnight oxygen desaturations are very frequent in SAHS patients. Moreover, the SpO_2 signal, which measures the level of arterial blood oxygen saturation, is easily recorded by means of an oximeter placed on the finger. Consequently, SpO_2 has been also commonly assessed as a reliable surrogate for PSG. It has been analyzed following several approaches. Rodriguez et al [109]. conducted a visual inspection of SpO_2 recordings to distinguish SAHS patients. A more common approach, however, focuses on the evaluation of the diagnostic performance of some clinical index obtained from the signal. In this regard, Levy et al. [78] studied the yield of several cut-offs for the delta index. The oxygen desaturation indexes (ODIs), derived from several desaturation definitions (SpO_2 drops of 2%, 3%, and 4%), have been also assessed both in adults [7, 79, 82, 130] and children [25, 123]. Additionally, the analysis of

SpO_2 recordings through automatic signal processing and pattern recognition methodologies has also become a common approach. Thereby, Roche et al. [108] obtained a multivariate linear regression model to estimate AHI by combining clinical and oximetry features. Álvarez et al. [9] evaluated a binary logistic regression model, obtained from SpO_2 time and frequency domain features, to classify SAHS patients and no-SAHS subjects. Moreover, Garde et al. [53] used a linear discriminant to combine features extracted from SpO_2 recordings of pediatric subjects. Artificial neural networks have been also applied to SpO_2 data to estimate the presence of SAHS, as well as its severity [83, 87, 113]. Some of these studies reported high diagnostic ability, showing the usefulness of single-channel SpO_2 to help in SAHS diagnosis. However, recurrent desaturations during sleep are not exclusive of SAHS. Other medical conditions such as chronic obstructive pulmonary disease or asthma may present similar patterns in the SpO_2 signal.

HRV (or the RR time series), derived from ECG, is known to reflect the autonomic nervous system behavior [4]. As a consequence, it is a comprehensively studied signal which has been analyzed in relation to a wide range of illnesses, including SAHS. Particularly, it has been shown that apneic events are associated with a recurrent bradycardia/tachycardia pattern [21, 60], which is reflected in the HRV signal as a higher/lower amplitude pattern. Several analytical approaches have been used to study the effects of SAHS in HRV. Penzel et al. [94] compared spectral and detrended fluctuation analyses to quantify the changes in HRV caused by SAHS. De Chazal et al. [36] extracted spectral and time domain features to obtain linear and quadratic discriminant classifiers with ability to detect apneic segments. Roche et al. [107] combined wavelet analysis and classification and regression trees to detect SAHS patients. Al-Angari and Sahakian [5] used HRV features to train support vector machine models in order to detect apneic epochs. Moreover, Ravelo-García et al. [103] combined clinical data and symbolic dynamic markers to classify subjects into SAHS or non-SAHS. Finally, Shouldice et al. [117] evaluated a quadratic discriminant classifier feed with features from HRV of children. As in the case of SpO_2, the HRV signal is also modified by other cardiovascular diseases, which represents one of its limitations. Additionally, sex or age are known to change the HRV behavior too.

Signals involved in PSG other than AF, SpO_2, and HRV have been less frequently analyzed. Among them, the snoring sound recording [44, 74], the photoplethysmography from children [56, 57], and the respiratory effort [5], showed promising results.

Chapter 2

Hypothesis and objectives

As it has been previously shown, the simplification of the SAHS diagnosis process has become a major concern. Hence, the proposal developed in this Doctoral Thesis is focused on decreasing its complexity by reducing the number of signals to be analyzed, conducting a proper characterization of SAHS using the information provided by these signals, developing a methodology with ability to detect SAHS, and automating the whole diagnostic process. These actions have been implemented following the next scheme:

i) Signal acquisition.

ii) Pre-processing stage.

iii) Feature extraction.

iv) Feature selection.

v) Pattern recognition.

This proposal is substantiated by the hypotheses and objectives described below.

2.1. Hypothesis

As an introductory step in this section, a naive hypothesis can be formulated to indicate the starting point of the study: *the diagnostic process of SAHS can be simplified*. This high level statement, however, does not suffice to focus the investigation by its own. Several assumptions of lower level have been also assessed for this purpose. As previously explained, AF have been widely studied in the context of SAHS diagnosis due to its specific role in the definition of apneas and hypopneas. Thus, at the signal level, it has been assumed that *AF provides relevant and enough information to help in SAHS diagnosis*. On the other hand, it is obvious that the methodology used plays the key role in this investigation. Therefore, a number of related hypothesis have been also evaluated. Thereby, it has been hypothesized that *feature extraction methodologies coming from different approaches are able to characterize SAHS in AF*. Similarly, it has been assessed whether *these approaches provide complementary information in the study of SAHS*. Feature selection algorithms have been used for this purpose. After SAHS characterization, however, it is still necessary to transform the information obtained from AF into mathematical models with ability to help in diagnosis. In this regard, it has been also hypothesized that *the pattern recognition approach can be helpful to diagnose SASH* in an automated way.

These are the main assumptions which integrate the core of the present study, which can be summarized in the next global hypothesis:

"*It is possible to reduce the complexity of the SAHS diagnostic process by means of an automated pattern recognition analysis of AF*".

2.2. Objectives

The general goal of this work is the comprehensive study and assessment of the diagnostic potential of the AF signal as a surrogate for full PSG in SAHS detection. As it has been previously shown, a general methodological framework involving feature extraction, feature selection, and pattern recognition is proposed. The aim is to gain insight into how SAHS affect AF as well as use the obtained information to simplify the diagnostic process. In order to achieve the main objective, the following specific objectives arise:

I. To build an AF database with signals from adult and pediatric subjects suspected of suffering from SAHS. Clinical and demographic data, as well as diagnosis derived from PSG, have to be associated to the corresponding recordings.

II. To review the bibliography and state-of-the-art related to feature extraction, feature selection, and pattern recognition techniques, appropriate to be used along with biomedical signals and, particularly, AF signals in the context of SAHS.

III. To select and implement (through Matlab®) those signal processing methodologies which, according to the reviewed bibliography, are more suitable to help in SAHS diagnosis.

IV. To process the signals according to the feature extraction, feature selection, and pattern recognition methodologies previously implemented.

V. To conduct statistical analysis of results to evaluate the suitability of each methodology applied to the recordings, as well as to assess the overall performance of the proposal.

VI. To compare and discuss the results to extract appropriate conclusions. This objective includes the comparison with the state-of-the-art studies, the implementation of other classic methodologies in our databases, as well as comparing our methodology applied to other well-known and widely-studied signals such as HRV.

VII. To publish the obtained results and conclusions in indexed journals, as well as international and national congresses.

Chapter 3

Materials

3.1. Subject databases: demographic and clinical data

During this investigation, 4 different databases were analyzed. All of them were composed of recordings from subjects suspected of suffering from SAHS, including one children database. The first database consisted of 148 AF recordings acquired with a thermistor. A second one was integrated by 317 AF recordings obtained through a nasal-pressure sensor, whereas 188 HRV recordings were involved in a third database. The last database consisted of thermistor AF and SpO_2 recordings from 50 children.

Recordings from adult subjects were acquired "in-lab"' during their corresponding PSGs. These were conducted in the sleep unit of the Hospital Universitario Rio Hortega (HURH), Valladolid, Spain. Recordings from children were obtained at patient's home as part of the investigations of the unit of respiratory sleep disorders of the Hospital Universitario de Burgos (HUBU), Burgos, Spain. Physicians established a diagnosis for each subject according to their corresponding AHI. Both for adults and children, physicians followed the AASM rules for scoring apneas and hypopneas [18]. Common AHI cutoffs to determine SAHS and its severity in adults are 5, 10, 15, and 30 e/h [18, 35, 91]. AHI=10 e/h has been widely used as a cutoff to determine the presence or absence of SAHS [35, 74, 91]. Additionally, SAHS severity levels can be distinguished by defining: no-SAHS (5<AHI), mild-SAHS (5≤AHI<15), moderate-SAHS (15≤AHI<30), and severe-SAHS (AHI≥30) [101]. In pediatric subjects, AHI = 3 e/h is also a common cutoff to establish the presence of SAHS [6]. All the adult subjects, as well as the parents of the pediatric ones, gave their informed consent to participate in the studies. The Ethics Committees of both the HURH and HUBU accepted the corresponding protocols. Tables 3.1 to 3.4 show demographic and clinical data of the subjects involved in the four databases, including age, body mass index (BMI), and male percentage. Data are presented divided into no SAHS subjects (SAHS-negative) and SAHS subjects (SAHS-positive) according to the adult (AHI=10 e/h) and pediatric (AHI=3 e/h) AHI cutoffs.

3.2. Signals analyzed during the study

The signals recorded during PSG come from different body systems. Therefore, their nature can be electrical, mechanical, optical, etc. Some of the main signals obtained from PSG are ECG, EEG, respiratory effort, SpO_2, or AF. In this study, AF has the most important role. However, heart rate variability

Table 3.1: Demographic and clinical data for the **AF** database (signals of **adult subjects** obtained through a **thermistor**) divided into SAHS-negative and SAHS-positive groups (mean ± standard deviation). BMI: body mass index. AHI: apnea-hypopnea index.

	All	**SAHS-negative**	**SAHS-positive**
Subjects (n)	148	48	100
Males (n)	117(79.0%)	32(66.7%)	85(85.0%)
Age (years)	50.9 ± 11.7	48.8 ± 12.1	51.9 ± 11.4
BMI (Kg/m^2)	29.1 ± 4.6	27.6 ± 4.9	29.9 ± 4.7
AHI (e/h)	–	4.0 ± 2.4	32.9 ± 24.3

Table 3.2: Demographic and clinical data for the **AF** database (signals of **adult subjects** obtained through a **nasal prong**) divided into SAHS-negative and SAHS-positive groups (mean ± standard deviation). BMI: body mass index. AHI: apnea-hypopnea index.

	All	**SAHS-negative**	**SAHS-positive**
Subjects (n)	317	110	207
Males (n)	226(71.3%)	68(61.8%)	158(76.3%)
Age (years)	49.9 ± 12.0	47.6 ± 12.9	51.1 ± 11.4
BMI (Kg/m^2)	28.1 ± 5.2	26.5 ± 5.0	29.0 ± 5.1
AHI (e/h)	–	6.0 ± 2.6	39.9 ± 25.9

Table 3.3: Demographic and clinical data for the **AF** database (signals of **pediatric subjects** obtained through a **thermistor**) divided into SAHS-negative and SAHS-positive groups (mean ± standard deviation). BMI: body mass index. AHI: apnea-hypopnea index.

	All	**SAHS-negative**	**SAHS-positive**
Subjects (n)	50	24	26
Males (n)	27(54.0%)	11(45.8%)	16(61.5%)
Age (years)	5.3 ± 2.5	5.2 ± 2.4	5.4 ± 2.7
BMI (Kg/m^2)	16.5 ± 2.5	16.1 ± 1.7	16.9 ± 3.0
AHI (e/h)	–	1.3 ± 0.8	17.9 ± 15.4

Table 3.4: Demographic and clinical data for the **HRV** database (signals of **adult subjects**) divided into SAHS-negative and SAHS-positive groups (mean ± standard deviation). BMI: body mass index. AHI: apnea-hypopnea index.

	All	**SAHS-negative**	**SAHS-positive**
Subjects (n)	188	69	119
Males (n)	134(71.3%)	41(59.4%)	93(78.2%)
Age (years)	50.7 ± 12.0	47.3 ± 11.5	52.7 ± 12.3
BMI (Kg/m^2)	28.7 ± 4.7	28.0 ± 6.1	29.1 ± 3.7
AHI (e/h)	–	3.8 ± 2.4	33.0 ± 22.9

(HRV), and SpO_2, have been also analyzed, at least to some extent.

AF is a physiological signal mainly used to evaluate whether ventilation is properly established [120]. In the context of sleep disorders, it is common to use it to assess respiratory patterns as well as eventual nocturnal events [43]. Due to the periodic nature of respiration, a non-pathological segment of AF shows a regular behavior. As previously stated, apnea and hypopnea definitions directly involve the reduction of AF. Consequently, this is a crucial signal to detect these events [13]. Originally, AF was measured through a pneumotachograph. However, it was invasive and, consequently, uncomfortable for patients [43]. Nowadays, AF is measured during PSG in two complementary ways, by means of both a thermal sensor (thermocouple, thermistor) and a pressure sensor. Due to their corresponding limitations [13, 43], the AASM recommends the former to detect apneas and the latter to detect hypopneas [18]. In the present study, the AF recordings from adults acquired by means of thermistor were obtained at a sample rate of 10 Hz (PSG equipment: Alice 5, Respironics, Philips Healthcare, the Netherlands). Those acquired through

a nasal pressure sensor were sampled at a rate of 128 Hz (PSG equipment: E-series, Compumedics Limited, USA). In the case of the children database, the AF recordings were obtained at a sample rate of 100 Hz (polygraphy equipment: eXim Apnea, Bitmed, Sibel S.A., Spain).

HRV, or RR time series, is obtained by computing the time between consecutive R peaks of the characteristic QRS complex pattern from ECG [4]. In this study, HRV has been used to show the diagnostic ability derived from applying signal processing techniques similar to those conducted in AF. Thereby, results derived from these analyses can be used to be compared with those from AF and, therefore, to gain more insight into the diagnostic ability of the latter. Roughly speaking, HRV shows the time between consecutive heartbeats. The HRV signal is not involved in the definition of apnea or hypopnea. However, a bradycardia-tachycardia pattern in the heart rate has been observed as a consequence of apneic events [21, 60], which is reflected in HRV as recurrent amplitude increases and decreases. ECG recordings involved in this study were obtained at a sample rate of 200 Hz (PSG equipment: Alice 5, Respironics, Philips Healthcare, the Netherlands).

The SpO_2 signal shows the level of arterial blood oxygen saturation. It is also involved in the definition of hypopnea. The information is typically acquired through an optic sensor placed in the finger as part of the pulse oximetry test, which is included in the PSG protocol. The oximeter measures the intensity of light which is transmitted from one side of the finger to the other at two different wavelengths: red spectrum and near infra-red spectrum [133]. Thus, the more intensity is detected by the oximeter the less oxygen concentration in blood. A non-pathological SpO_2 signal is in the range 95-97% regardless the age, ethnicity, gender, or weight of the subject monitored [90]. Conversely, SAHS patients present recurrent drops in the oxygen saturation (desaturation), reaching levels below 40%. Previous studies of our research group have already shown the utility of a comprehensive analysis approach conducted in SpO_2 recordings from adult subjects [7, 8, 9, 84, 85]. Hence, in this study, SpO_2 (sample rate = 100 Hz., polygraphy equipment: eXim Apnea, Bitmed, Sibel S.A., Spain) has been only used in the case of children. Particularly, only the information provided by the 3% ODI has been used.

Chapter 4

Methods

The general methodology conducted during the study (see Figure 4.1), starts with an initial stage proposed to minimize undesirable noise and artifacts in the signals (*pre-processing*). Then, the recordings from each subject under study are analyzed to gain insight into the effects that SAHS causes in them, i. e., to characterize SAHS in these signals (*feature extraction*). It is known that physiological signals tend to have both deterministic and stochastic behaviors [33]. Consequently, linear and non-linear analyses have been conducted in order to obtain as much complementary information as possible about SAHS. Once this exhaustive analysis is done, it might happen that irrelevant or redundant information is extracted. Consequently, a *feature selection* stage is implemented, aimed at optimizing the information to be used for the final diagnostic process. Two approaches have been followed to select this relevant and non-redundant information: a wrapper algorithm, which is dependent on the subsequent pattern recognition technique to be applied; and a filter algorithm, which is independent of subsequent analyses. After this stage, each subject is characterized by a vector $\mathbf{x}$ whose components are the corresponding values of the selected features. These vectors can be viewed as patterns which gather the main information obtained from each subject. Therefore, they are the inputs in the final *pattern recognition* stage. Pattern recognition techniques are divided into two main groups: classification and regression methods. The former are aimed at classifying data into two or more categories, which in the current study has meant classifying subjects into SAHS-positive or SAHS-negative (binary classification), as well as into one of the four severity degrees of SAHS (multi-classification). On the other hand, regression is aimed at estimating one or several continuous variables. In this study, we have used regression techniques to estimate the AHI of every subject. During the whole methodological process, several statistical issues need to be considered to properly measure the diagnostic ability of the proposal, as well as to ensure the result's validation. This is explained in the *statistical analysis* section.

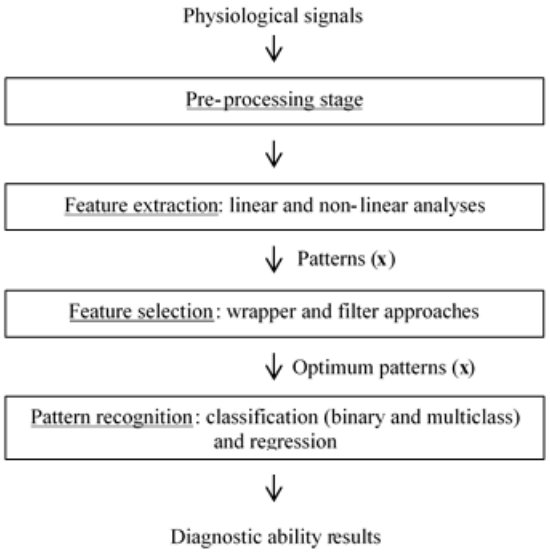

Figure 4.1: Scheme of the general methodology conducted in the study

4.1. Pre-processing

4.1.1. AF signal

Although the AF signal is not particularly noisy, some actions were taken in order to ensure the quality of the recordings analyzed. First, an anti-aliasing filter was applied during the acquisition process to satisfy the Nyquist-Shannon theorem. Then, a visual inspection of the signals was conducted to avoid occasional recordings without enough data due to prolonged malfunction of the sensor. In this regard, only signals with data corresponding to 3 or more hours of sleep were considered for the study. Eventually, less than 3% of the AF recordings were discarded. Finally, a Butterworth infinite impulse response low-pass filter (cutoff = 1.2 Hz) was also used in order to reduce noise for prospective analyses in time domain.

The acquisition of the respiratory rate variability signal (RRV), directly derived from AF, was also part of the pre-processing stage. Similarly to the well-known HRV, which is derived from ECG, the RRV signal is computed by measuring the time between consecutive breaths [34]. In this regard, a peak detection algorithm was implemented to locate inspiratory onsets in AF [77]. Thus, the first derivative of AF was examined to find time intervals in which the original signal grew. Then the AF maximums at each interval were located. Finally, consecutive locations were used as references to measure the time from one breath to the next [77]. Due to this computation process, the RRV signal lose the constant sample rate of AF, which is needed for subsequent analyses in the frequency domain. Therefore, prior the the spectral analysis, a cubic spline interpolation was applied to the RRV series in order to resample the recordings to a constant sample rate.

4.1.2. HRV signal

As in the case of AF, an anti-aliasing filter as well as a visual inspection of the signals were applied during and after the acquisition process of the ECG signal, respectively. The main pre-processing actions were related to the computation of the HRV signal. Each sample in the HRV signal is the time between two consecutive R peaks, which are located in the characteristic QRS complex of the ECG [15]. Hence, to derive HRV, a QRS-complex detection algorithm was firstly applied [16]. It was reported to reach high sensitivity (99.94%) and positive predictive value (99.93%), even in the presence of muscular noise and baseline artifacts (99.88% sensitivity and 99.73% positive predictive value, respectively) [16]. This algorithm is based on Hilbert transform and consists of two stages. Initially, the first differential of the ECG signal is computed (dECG). This is carried out to avoid baseline shifts and motion artifacts. Then, the Hilbert transform is applied to dECG ($h(n)$ = H[dECG]). Due to the properties of Hilbert transform, points around peaks in $h(n)$ are regions of high probability of containing actual QRS peaks [16]. Since the P and T waves are low comparing with the R waves in $h(n)$ [16], an adaptive threshold is used to establish those regions truly corresponding to R waves. In the second stage of the algorithm, these regions are used to look for the actual peaks in the original ECG. After QRS-complex detection, the difference between R-R peaks is computed.

In order to deal with arrhythmia-related artifacts, we excluded those R-R intervals not fitting: i) 0.33 seconds $<$ R-R interval $<$ 1.5 seconds and ii)

difference to the previous R-R interval > 0.66 seconds [94]. As in the case of the RRV signal, HRV was resampled to a constant sample rate before carrying out spectral analyses [94].

4.1.3. SpO_2 signal

SpO_2 is not a noisy signal. However, artifacts due to subject movements might arise. Hence, in addition to an anti-aliasing filter and a visual inspection, an artifact removal was applied. In this regard, SpO_2 values equal to zero as well as differences between consecutive SpO_2 samples $\geq$4% were considered artifacts [82]. Removed samples were substituted by interpolated data.

4.2. Feature extraction

As mentioned above, linear and non-linear analyses were conducted in order to characterize SAHS in the physiological signals. In this regard, linear analysis was conducted both in time and frequency domain, whereas non-linear features were obtained from time series.

4.2.1. Frequency domain: spectral analysis

Spectral analysis is a classic approach to investigate physiological signals [22]. Particularly, the recurrent behavior of the apneic events during sleep justified the use of this frequency analysis during the study. Thus, the power spectral density (PSD) of each recording was computed to look for the effects that SAHS causes in the physiological signals under study. PSD was estimated using the nonparametric Welch's method, which is suitable for non-stationary signals [125].

Spectral bands of interest

When analyzing overnight physiological recordings, like AF's and HRV's from PSG, normal patterns are still predominant even in the presence of SAHS. Hence, in order to define the effects of SAHS in the spectrum of the signals, it is useful to find the particular frequency range or ranges in which these are observed, i. e., to establish bands of interest. In this regard, PSDs from SAHS-positive and SAHS-negative subjects were compared, frequency by frequency, by the use of the proper statistical hypothesis test for each case. Then, the band or bands of interest were defined according to the p-values reached in the comparison of PSD amplitudes at each frequency. That is, bands of interest were defined by those frequencies in which the highest statistical differences (i. e. lowest p-values) were found between the PSDs of healthy subjects and SAHS subjects. Figure 4.2 illustrates this methodology in the case of the AF database from children, where BW1 and BW2 refer to each of the 2 bands of interest found for that case.

Spectral features

Once spectral bands were defined, different features were computed in the frequency domain. Thus, during the study, up to 11 spectral features were

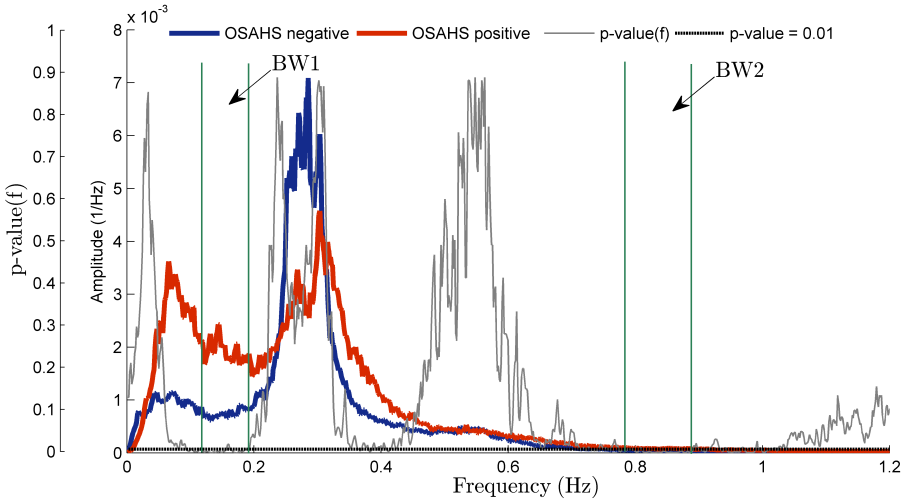

Figure 4.2: Illustration of the methodology conducted to obtain spectral bands of interest [61].

obtained from the PSDs of the signals under study. A brief explanation of each of them is shown below.

- **First to fourth statistical moments in frequency domain**, which are the well-known mean (M_{f1}), standard deviation (M_{f2}), skewness (M_{f3}), and kurtosis (M_{f4}), obtained from the bands of interest or the whole PSDs as appropriate. They quantify the central tendency, dispersion, asymmetry, and peakedness of data, respectively.

- **Maximum and minimum PSD amplitude**, computed as the highest (MA) and the lowest (mA) PSD values in a certain frequency band. If the PSD is normalized (PSD_n), these features estimate the maximum and minimum occurrence of events within the band considered.

- **Power features**. By definition, the power in a spectral band (P_B) can be estimated from PSD by computing the area under the curve in that band. When considering PSD_n, the more power, the higher occurrence of events in the band considered. Sometimes, it can be also useful to estimate the power ratio between two bands. This is the case of the PSD of HRV, where there are defined clear high frequency (HF) and low frequency (LF) bands related to respiratory rhythms and sympathetic activity, respectively [94]. The ratio between P_{LF} and P_{HF} ($P_{LF/HF}$) is a commonly analyzed feature when studying HRV [4].

- **Median frequency (MF)**. MF is defined as the frequency component which separates the spectrum of certain band into two parts, holding 50% of the power each of them [99]. Thus, the lower the MF value, the more comprised is the spectrum into small frequencies. MF can be obtained as follows [99]:

$$\frac{1}{2} \cdot \sum_{f=f_1}^{f_2} PSD(f) = \sum_{f=f_1}^{MF} PSD(f), \tag{4.1}$$

where f_1 and f_2 are the frequency limits of the band considered, with $f_1 < f_2$.

- **Spectral entropy ($SpecEn$).** $SpecEn$ quantifies the flatness of the PSD content, which indirectly measures the irregularity of the associated time series [71, 99]. Thereby, high values of $SpecEn$ ($SpecEn \to 1$) are related to a flat PSD (similar to white noise) and, consequently, it is associated with more irregularity in time domain. By contrast, low $SpecEn$ values ($SpecEn \to 0$) imply a spectrum condensed into a narrow frequency band, which is related to less irregularity in time domain (like in a sum of sinusoids) [99]. Since PSD needs to be considered as a probability density function, it has to be normalized in order to sum 1. Then $SpecEn$ can be computed from the following expression [99]:

$$SpecEn = \sum_{f=f_1}^{f_2} PSD_n(f) \cdot \log PSD_n(f), \tag{4.2}$$

  which is the application of Shannon's entropy to the normalized values of the PSD, between the f_1 and f_2 frequency limits.

- **Wootter's distance (WD).** WD is a disequilibrium measurement which assigns values close to 1 to those distributions with higher statistical distance to the uniform distribution. By contrast, values close to 0 are achieved as this distance becomes smaller [86]. WD also requires PSD to be normalized. It can be estimated as follows:

$$WD = \arccos\left\{ \sum_{f=f_1}^{f_2} \sqrt{PSD_n(f)} \cdot \sqrt{1/N} \right\}, \tag{4.3}$$

  where N is the number of the PSD_n points.

4.2.2. Time domain

Nine features from time domain have been also used during the study in order to complement the spectral analysis. These are described below.

Common statistics: first to fourth statistical moments

As in the case of the frequency domain, first to fourth statistical moments were also obtained in time domain ($M_{t1} - M_{t4}$). These estimated central tendency, dispersion, asymmetry, and peakedness from time series data instead of from the PSD.

Non-linear features

Non-linear methods were used to obtain information from the stochastic components of the biomedical signals under study. In the past, these methods have shown its usefulness as a complementary approach to spectral analyses, as well as to obtain helpful information where these are not possible or useless. During the study, 5 non-linear features have been extracted from the recordings

associated to the subject database. A brief description of each of them is shown below:

- **Central tendency measure (CTM).** CTM quantifies the degree of variability or chaos in a given time series x [31]. It is based on the plots of the first-order differences, $x(n+2)-x(n+1)$ vs. $x(n+1)-x(n)$, where $x(n)$ represents the n value of the time series [3]. CTM is computed by obtaining the proportion of points of the plot which fall within a radius ρ around the origin [31]:

$$CTM = \frac{1}{N-2}\sum_{n=1}^{n-2}\delta(n), \tag{4.4}$$

where

$$\delta(n) = \begin{cases} 1 \text{ if } \{(x(n+2)-x(n+1))^2 + (x(n+1)-x(n))^2\}^{1/2} \leq \rho \\ 0 \text{ otherwise,} \end{cases} \tag{4.5}$$

with N being the size of the time series. CTM ranges between 0 and 1, reaching values closer to 1 when a given series is less variable (values more concentrated around center) and closer to 0 when it has more variability (values more dispersed). Radius ρ has to be selected experimentally, depending on the character of the data [31].

- **Lempel-Ziv complexity (LZC).** LZC estimates the complexity of a given finite sequence of symbols $P = s(1), s(2), ..., s(n)$ [132], with larger values of LZC corresponding to a higher level of complexity. The first step of the algorithm is to transform a time-series x into a symbol sequence. Usually, a binary transformation is carried out (symbols 0-1), taking the median of x as threshold (Th) [3]:

$$s(n) = \begin{cases} 0 \text{ if } x(n) < Th \\ 1 \text{ if } x(n) \geq Th \end{cases} \tag{4.6}$$

where $x(n)$ and $s(n)$ are the nth values of x and P, respectively.

Once the sequence P is obtained, it is scanned from left to right, and a complexity counter $c(n)$ is increased every time a new subsequence of consecutive characters is encountered [131]. The following algorithm is used to obtain $c(n)$ [131]:

1. Let S and Q denote two subsequences of P, and SQ be the concatenation of S and Q. Let sequence $SQ\pi$ be the sequence SQ with the last character removed. Let $v(SQ\pi)$ denote the vocabulary of all different subsequences of $SQ\pi$. Initialize $c(n) = 1$, $S = s(1)$, $Q = s(2)$ and, consequently, $SQ\pi = s(1)$.
2. In general, $S = s(1), s(2), ..., s(r)$, $Q = s(r+1)$, then $SQ\pi = s(1), s(2), ..., s(r)$. If Q belongs to $v(SQ\pi)$, then Q is not a new sequence but a subsequence of $SQ\pi$.
3. Redefine Q to be $Q = s(r+1), s(r+2)$ and check whether Q belongs to $v(SQ\pi)$.

4. Repeat step 3 until Q does not belong to $v(SQ\pi)$. Since $Q = s(r+1), s(r+2), ..., s(r+i)$ is not a subsequence of $SQ\pi = s(1), s(2), ..., s(r+i-1)$ increase $c(n)$ by 1.
5. Thereafter, S and Q are redefined as $S = s(1), s(2), ..., s(r+i)$ and $Q = s(r+i+1)$.

These steps are repeated until Q is the last character. Then, $c(n)$ is the number of different subsequences in P. In order to make $c(n)$ independent from the length of the time series, it has to be normalized as follows:

$$C(n) = \frac{c(n)}{b(n)}, \tag{4.7}$$

where $b(n)$ is defined as:

$$b(n) = \lim_{n \to +\infty} c(n) \equiv \frac{n}{log_\alpha(n)}, \tag{4.8}$$

and $\alpha = 2$ since the number of symbols is 2.

- **Approximate entropy (*ApEn*).** *ApEn* is an irregularity measurement in time series which was originally developed to be applied over short and noisy data sets [97]. It discriminates series for which clear feature recognition is difficult by the assessment of both dominant and subordinate patterns [98]. *ApEn* has two user-specified parameters: a length m and a tolerance window r. Given N points of a time series $x(n) = x(1), x(2), ..., x(N)$, it can be computed following the next steps [68, 98]:

  1. Form $N-m+1$ vectors $X(1), ..., X(N-m+1)$, each of them defined as $X(i) = [x(i), x(i+1), ..., x(i+m-1)]$, $i = 1, ..., N-m+1$, i. e., each *ith*-vector represents m consecutive values commencing with the *ith* data point.
  2. Define the distance between $X(i)$ and $X(j)$, $d[X(i), X(j)]$, as the maximum absolute difference between their respective scalar components.
  3. For a given $X(i)$, count the number $(N^m(i))$ of j $(j = 1, ..., N-m+1)$ for which $d[X(i), X(j)] \leq r$. Then, for $i = 1, ..., N-m+1$:

  $$C_r^m(i) = \frac{N^m(i)}{N-m+1}, \tag{4.9}$$

  where $C_r^m(i)$ measures, within a tolerance r, the frequency of patterns similar to a given one with a window length m.
  4. Compute the natural logarithm of each $C_r^m(i)$ and average it over i:

  $$\phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_r^m(i). \tag{4.10}$$

  5. Increase m to $m+1$ and repeat steps 1st to 4th in order to find $C_r^{m+1}(i)$ and ϕ^{m+1}

6. Finally, *ApEn* is defined by

$$ApEn(m, r, N) = \phi^m(r) - \phi^{m+1}(r). \tag{4.11}$$

According to this algorithm, *ApEn* measures the logarithmic likelihood that patterns which are close (within r) for m contiguous observations remain close (within the same r) for $m+1$ contiguous observations. Thus, *ApEn* assigns larger values to more irregular data [98]. Both m and r are user-specified parameters. However, a range of values have been proposed to provide *ApEn* with good statistical reproducibility: $m = 1, 2$ and $r = 0.1, 0.15, 0.20, 0.25$ times the standard deviation of the original series [68, 98].

- **Sample entropy (*SampEn*)**. Richman and Moorman developed *SampEn* to reduce the bias caused by self-matching in the estimation of *ApEn* [105]. As in the case of *ApEn*, a run length m and a tolerance window r must be specified to compute *SampEn*. Thus, time-series are divided into consecutive vectors of length m and it is assessed whether the maximum absolute distance between the corresponding components of each pair of vectors is less than or equal to the tolerance r, i.e., if the vectors match each other within r. If so, the vectors are considered as similar. The same process is repeated for vectors of length $m+1$. Then, it is computed the conditional probability of similar vectors of length m remaining similar when the length is $m + 1$. The final *SampEn* value is obtained as the negative logarithm of such conditional probability [2, 105]. Thus, higher values of *SampEn* indicate less self-similarity in the times-series and, consequently, more irregularity [2]. Given N points of a time series $x(n) = x(1), x(2), ..., x(N)$, *SampEn* can be computed following the next steps:

  1. Form $N-m+1$ vectors $X(1), ..., X(N-m+1)$, each of them defined as $X(i) = [x(i), x(i+1), ..., x(i+m-1)]$, $i = 1, ..., N-m+1$, i. e., each ith-vector represents m consecutive values commencing with the ith data point.
  2. Define the distance between $X(i)$ and $X(j)$, $d[X(i), X(j)]$, as the maximum absolute difference between their respective scalar components.
  3. For a given $X(i)$, count the number (B_i) of j $(j = 1, ..., N-m$, $j \neq i)$ for which $d[X(i), X(j)] \leq r$. Then, for $i = 1, ..., N-m$:

$$B_i^m(r) = \frac{B_i}{N-m+1}. \tag{4.12}$$

  4. Compute $B^m(r)$ as:

$$B^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} B_i^m(r). \tag{4.13}$$

  5. Increase m to $m+1$ and calculate A_i as the number of $X_{m+1}(i)$ within r of $X_{m+1}(j)$, where $j = 1, ..., N-m$, $j \neq i$. $A_i^m(r)$ is

$$A_i^m(r) = \frac{A_i}{N-m+1}. \tag{4.14}$$

6. $A^m(r)$ is computed as

$$A^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} A_i^m(r). \tag{4.15}$$

Thus, $B^m(r)$ and $A^m(r)$ are the probability that two sequences will match for m and $m+1$ points, respectively. Finaly, $SampEn$ is estimated by:

$$SampEn(m, r, N) = -\ln \frac{A^m(r)}{B^m(r)}. \tag{4.16}$$

- **Multi-scale entropy (MsE).** MsE was originally developed by Costa et al. on the basis of $ApEn$ or $SampEn$ [32]. As previously shown, both of them measure irregularity in time series. MsE procedure, however, computes entropy for different time-scales of time series, which makes it a complexity measure rather than an irregularity measure [33, 42]. In the current study, only $SampEn$ has been involved in MsE. Thereby, estimating $SampEn$ for an original time series or recording is equivalent to $SampEn$ at scale 1. Scale 2 is obtained by averaging the original time series every 2 samples without overlapping; scale 3 when the average is every 3 samples, and so on. The scaled time series, y^τ, can be computed as follows [32]:

$$y_j^\tau = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, 1 \leq j \leq N/\tau, \tag{4.17}$$

where τ is the corresponding scale and y_j is each of the elements of the new time series. Then, each of the scaled time series are subsequently characterized by the corresponding $SampEn$ value. The tendency of the $SampEn$ curve throughout the scales shows the complexity level in time series.

4.3. Feature selection

Two automated selection algorithms were used in order to obtain optimum sets of features among those extracted in the previous stage. Thus, we implemented the forward-selection backward-elimination algorithm on the basis of logistic regression. Moreover, the fast correlation-based filter, which is independent of the pattern recognition technique subsequently used, was also applied to the extracted features.

4.3.1. Stepwise logistic regression: the forward-selection backward-elimination algorithm (SLR-FSBE)

Given a problem context, stepwise logistic regression provides a fast way to determine significant associations among some variables under study [69]. In this case, these variables are features extracted from physiological signals, and the context is SAHS diagnosis. The associations among variables are defined

according to a fixed decision rule, the logistic function, as well as the statistical significance of the coefficients (or weights) of each variable. Since logistic regression assumes errors to follow a binomial distribution, the likelihood ratio chi-square test is used to measure that statistical significance [69]. The dependence of this feature selection algorithm on a specific prediction model (logistic regression, LR) classifies stepwise logistic regression as a *wrapper method* [66].

Forward-selection backward-elimination (FSBE) is a common approach to implement stepwise logistic regression (SLR-FSBE) [69]. This strategy is efficient in terms of computation as well as robust against overfitting [66]. The SLR-FSBE algorithm can be implemented following the next steps [69]:

1. **Step 0.** Given a set of N variables to be evaluated, the starting point of this algorithm is a LR model only composed of the intercept, i. e., the constant term. Then, N univariate LR models, one for each variable, are fit to compare them with this LR intercept model by means of the p-value of the likelihood ratio chi-square test. Thus, for each independent variable x_i, a LR model and a p-value $p_i^{(0)}$ is computed. The most significant variable x_{e1} is the one with the lowest p-value:

$$p_{e1}^{(0)} = min(p_i^{(0)}). \tag{4.18}$$

   x_{e1} is considered significant enough iff $p_{e1}^{(0)} < \alpha_I$ and, in such case, it is included in the model. Otherwise the algorithm stops after selecting no variables. The election of α_I should be guided by the context [69]. However, common values are in the range 0.05 (which is restrictive to include a low number of variables) and 0.25 (which let more variables be part of the model) [69].

2. **Step 1.** The starting point of this step is a LR model with the intercept and x_{e1}. In order to assess whether the remaining $N-1$ variables are significant, $N-1$ LR models are fit with the intercept, x_{e1}, and x_i, with $i = 1, 2, ..., N$ and $i \neq e1$. These models are compared with the model from the previous step by computing the p-value of the likelihood ratio chi-square test, i. e., $p_i^{(1)}$. As in the previous step, the most significant variable is computed as:

$$p_{e2}^{(1)} = min(p_i^{(1)}). \tag{4.19}$$

   x_{e2} is included in the model iff $p_{e2}^{(1)} < \alpha_I$ and the algorithm proceeds with the next step. Otherwise, the algorithm stops.

3. **Step 2.** The two preceding steps are part of the forward-selection stages of the algorithm. In this step it is also integrated the backward-elimination. The starting point is a LR model with the intercept, x_{e1}, and x_{e2}. The objective is to evaluate whether x_{e1} is still significant once the variable x_{e2} is included in the model. Thus, they are fitted as much models as variables included in previous steps. For each one, only one of the variables is excluded. All the models are compared with the starting one by means of the p-value of the likelihood ratio chi-square test, i. e., $p_{-ei}^{(2)}, i = 1, 2$. A variable x_{r2} is a candidate to be removed from the model if:

$$p_{-r2}^{(2)} = max(p_{-ei}^{(2)}), i = 1, 2. \tag{4.20}$$

x_{r2} is removed from the model iff $p^{(2)}_{-r2} > \alpha_R$, with $\alpha_R > \alpha_I$. This new threshold is usually in the range 0.2-0.9 [69].

After the backward-elimination process, it is evaluated the requirement for a new variable to be included in the model. Thus, $N-2$ LR models are fitted containing the intercept, the variables x_{e1}, and x_{e2} selected in previous steps, and a new variable x_i, with $i = 1, 2, ..., N$ and $i \neq e1, e2$. Then, each model is compared with the model obtained after the backward-elimination process. As in previous steps, a new variable x_i is a candidate to enter the model, x_{e3}, if its associated p-value is the lowest among all variables. Iff $p^{(2)}_{e3} < \alpha_I$, x_{e3} is finally included in the model and the algorithm proceeds with the next step. Otherwise, the algorithm stops.

4. **Subsequent steps.** At each step, the algorithm carries out a backward-elimination procedure followed by a forward-selection one.

5. **End of the algorithm.** The algorithm ends when the N variables of the original set has been included in the model or when none of the candidate variables satisfies the condition to enter the model and none of the variables included satisfies the condition to be removed from the model.

4.3.2. Fast correlation-based filter (FCBF)

The fast correlation-based filter (FCBF) is an automated selection algorithm which is independent of the posterior pattern recognition methods applied to the features, i. e., it is considered a *filter method* [66]. It has shown its utility in previous studies involving biomedical applications [1]. FCBF relies on relevance and redundancy analyses of the variables under study [129]. Thus, the purpose is to discard those features x_i which share more information with the others than with a dependent variable of interest, y. For the current study, y is a vector whose components are the AHI values of the subjects.

FCBF is based on symmetric uncertainty (SU), which is a normalized quantification of the information gain (IG) between two variables [129]. The algorithm consists of two steps. In the first one, a relevance analysis of the features x_i is conducted. Thus, SU between each feature x_i and y is computed as follows [129]:

$$SU(x_i, y) = \left[\frac{IG(x_i \,|\, y)}{H(x_i) + H(y)} \right], i = 1, 2, ..., N, \tag{4.21}$$

where $IG(x_i \,|\, y) = H(x_i) - H(x_i \,|\, y)$, H is the well-known Shannon's entropy, and N is the number of variables considered. SU is constrained to 0-1. A 0 value indicates that the two variables are independent, whereas $SU = 1$ indicates that knowing one feature it is possible to completely predict the other [129]. Thus, the higher the value of SU, the more information shares the corresponding feature with y and, consequently, the more relevant is. Then, a ranking of variables is carried out on the basis of their $SU(x_i, y)$ values, i.e., they are sorted from most relevant to least relevant. The second step is a redundancy analysis in which SU between each pair of features ($SU(x_i, x_j)$) is sequentially estimated beginning from the first-ranked ones. If $SU(x_i, x_j) \geq SU(x_i, y)$, with x_i being more highly ranked than x_j, the feature x_j is discarded due to redundancy and is not considered in next comparisons [129]. The optimum set

of variables is composed of those not discarded after all comparisons between variables have been done.

4.4. Pattern recognition

Pattern recognition concerns the identification of underlying behaviors in data through the use of automatic algorithms. These behaviors, or patterns, can then be used to define data into a class (classification) or to derive continuous variables (regression) from them [20].

4.4.1. Classification

Linear discriminant analysis

Linear discriminant analysis (LDA) is a supervised classifier which assigns a vector $\mathbf{x}_i$ (with $i = 1, 2, ..., S$ and S the number of instances), into one out of K classes, C_j ($j = 1, 2, ..., K$). It relies on the assumption that the conditional density function of each class, $P(\mathbf{x}_i \,|\, C_j)$, follows a multivariate normal distribution (normality) with identical covariance matrices, $\boldsymbol{\Sigma}$, for all classes (homocedasticity) [20]. A discriminant score $y_j(\mathbf{x})$ is computed for each class following [51, 84]:

$$y_j(\mathbf{x}) = \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \mathbf{x}\, \boldsymbol{\mu}_j + \ln P(C_j), \tag{4.22}$$

where $\boldsymbol{\mu}_j$ is the mean vector for class C_j and $P(C_j)$ its corresponding prior probability, i.e., the initial proportion of vectors $\mathbf{x}_i$ belonging to class C_j. The classification task is carried out through the decision rule, "assign a new vector $\mathbf{x}_i$ to the class C_j if $y_j(\mathbf{x}_i) = \max_{j=1,2,...,K}(y_j(\mathbf{x}_i))$".

Logistic regression

LR has become a standard when classifying data into one out of two classes, i. e., it is a binary classifier. It is also a supervised learning algorithm. Thus, given S observations or instances $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, 2, ..., S$, $\mathbf{x}_i$ denotes the vector (or pattern) which characterizes the instance i, whereas $\mathbf{y}_i$ is the value of a binary outcome associated with the same instance [69]. In this regard, LR estimates the posterior probability that a given instance $\mathbf{x}_i$ belongs to certain class C_j ($j = 1, 2$) that is, $P(C_j \,|\, \mathbf{x}_i)$. This is carried out through the logistic function [69]:

$$P(C_j \,|\, \mathbf{x}_i) = \frac{e^{\beta_0 + \mathbf{x}_i \boldsymbol{\beta}}}{1 + e^{\beta_0 + \mathbf{x}_i \boldsymbol{\beta}}}, \tag{4.23}$$

where β_0 and $\boldsymbol{\beta} = \beta_1, \beta_2, ..., \beta_N$, are the coefficients of the LR model, and N is the number of independent variables which compose each $\mathbf{x}_i$ vector. $\beta_0, \beta_1, \beta_2, ..., \beta_N$ are obtained through the maximum likelihood estimator [69]. Then, an instance $\mathbf{x}_i$ is assigned to the class C_j with larger posterior probability, $P(C_j \,|\, \mathbf{x}_i)$.

Classification and regression trees

The classification and regression trees algorithm (CART) is a strategy to implement decision trees. It is a non-parametric learning method which relies on a recursive binary partitioning of the input space to take decisions on some data [20]. As its name suggests, CART let these decisions be approached as classification or regression problems. However, in the current study, only the classification approach has been adopted.

Given an input space $\mathbf{X}$ composed of N variables, $\mathbf{x}_i = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, CART starts by dividing it according to one of its variables. Then, the two subsets formed after this division can be also split according to another variable, and so on. This general description poses two issues to be solved: *i)* what strategy to implement in order to grow the tree and *ii)* when to stop adding nodes to the tree [20].

The first issue includes selecting the space that can be split (at each step there will be several candidate regions), as well as the specific variable and threshold used to do it. The three of them are optimized jointly through an exhaustive greedy strategy, adding one node at a time [20]. In the case of classification tasks, the measure of performance to carry out these selections is the *Gini* index, which can be computed as follows [20]:

$$G_\tau(T) = \sum_{j=1}^{K} p_{\tau j}(1 - p_{\tau j})\,. \tag{4.24}$$

where $\tau = 1, 2, ..., T$, with T as the number of leaf nodes at each step (nodes without children nodes), and $p_{\tau j}$ is the proportion of original data points (or instances), associated with a region R_τ, which are assigned to class j, $j = 1, 2, ...K$.

For the second issue, one common approach is to grow a large tree until a specific number of leaf nodes is reached. Then this tree is pruned back. The criterion for the pruning process is given by [20]:

$$C(T) = \sum_{\tau=1}^{T} Q_\tau(T) + \lambda T\,, \tag{4.25}$$

where $Q_\tau(T)$ is the misclassification error associated with the corresponding region R_τ, and λ is a regularization parameter which is a trade-off between this error and the complexity of the model (the number of leaf nodes, T).

Ensemble learning: adaptive boosting

Ensemble learning refers to combine different models fitted from the available data in order to achieve better performance than each one separately [126]. Boosting is one of the most powerful strategies to develop ensemble learning algorithms [20], being known for achieving good generalization ability when testing on new data [126].

Boosting procedures are iterative algorithms designed to combine models that complement one another [126]. Such a combination is conducted on the basis of weighted votes assigned to *base* classifiers of the same type, which are fitted at each iteration [20, 126]. *AdaBoost*, for adaptive boosting, is a widely used boosting algorithm which can be used along with any classifier [50, 126]. However, if *AdaBoost* is applied to complex classifiers, the prediction ability

on new data may be significantly decreased [126], i.e., its generalization ability may be lost. Thus, simpler procedures known as *weak* classifiers are preferable [126].

At each m iteration, *AdaBoost* assigns a weight, w_i^m to every instance $\mathbf{x}_i$ ($i = 1, 2, ..., S$, being S the number of instances). Thus, the m_{th} weak classifier is trained using the corresponding weighted instances. Then, its performance is assessed through an error ϵ_m. This error is used to determine the weighted vote α_m for this m_{th} classifier [126]. Those classifiers with smaller ϵ_m contribute more to the final decision (higher α_m). At the end of each iteration, the weights of the misclassified instances are updated (w_i^{m+1}) [126]. Finally, the weights of all instances are normalized in order to maintain the original distribution [50].

Two versions of *AdaBoost* have been implemented in this study: *AdaBoost.M1*, for binary classification and *AdaBoost.M2* for multiclass classification. Both of them rely on reweighing those instances which have been misclassified in the previous iteration. Thus, the weak classifier trained during the next iteration gives more importance to these instances [20], being more likely to classify them rightly [126]. The main difference between *AdaBoost.M1* and *AdaBoost.M2* is how the error ϵ_m is defined. For *AdaBoost.M1*, ϵ_m is the sum of the weights of the misclassified instances in a given iteration m, divided by the sum of the total weights of all instances at that iteration:

$$\epsilon_m = \frac{\sum_{i=1}^{S} w_i^m(miss.)}{\sum_{i=1}^{S} w_i^m} \,. \tag{4.26}$$

By contrast, a weighted pseudo-loss is defined in the case of *AdaBoost.M2*, for which ϵ_m is computed as follows [50]:

$$\epsilon_m = \frac{1}{2}\sum_{i=1}^{S}\sum_{c\neq c_{true}} w_{k,c}^m(1 - h_m(\mathbf{x}_k, c_{true}) + h_m(\mathbf{x}_k, c)) \,, \tag{4.27}$$

where c is a categorical variable representing the multiple classes, c_{true} refers to the actual class of $\mathbf{x}_k$, and h_m is the confidence of the prediction of the weak learner for an instance $\mathbf{x}_k$ and a class from c.

Both *AdaBoost.M1* and *AdaBoost.M2* carry out the final classification task by returning the class with the highest sum of the votes from all classifiers, taking into account the weight α_m of their corresponding predictions [50]:

$$\alpha_m = \ln \beta_m \,, \tag{4.28}$$

where β_m is defined as $(1 - \epsilon_m)/\epsilon_m$. Additionally, the shrinkage regularization technique has been proposed to minimize overfitting [52]. It is based on adding a learning rate υ to the iterative process by redefining β_m as $(\beta_m)^\upsilon$, where υ ranges 0-1 and has to be experimentally estimated.

Finally, two criteria were used to stop the *AdaBoost.M1* algorithm: *i)* ϵ_m does not belong to the interval (0, 0.5) [126] or *ii)* the number of weak learners is not higher than 400 (to minimize the overfitting chances). In the case of *AdaBoost.M2*, only the second criterion was applied since the first one is considered too restrictive for multiclass approaches [50].

4.4.2. Regression

Multiple linear regression

Multiple linear regression (MLR) is a traditional method to predict a target variable $\mathbf{t}$, through an output variable $\mathbf{y}$, estimated from multivariate patterns of size N, $\mathbf{x}_i = x_{i1}, x_{i2}, ..., x_{iN}$, $i = 1, 2, ..., S$, with S being the total number of instances considered. A linear relationship between $\mathbf{y}$ and $\mathbf{x}_i$ is assumed [72]:

$$y(\mathbf{x}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_N x_{iN} \,, \tag{4.29}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_N)^T$ are the regression coefficients for each input variable $\mathbf{x}_i$, as well as the intercept (β_0). $\boldsymbol{\beta}$ is computed by means of the sum-of-squares error (E_D) minimization [19]:

$$E_D = \frac{1}{2} \sum_{i=1}^{S} \left[y(\mathbf{x}_i, \boldsymbol{\beta}) - t_i \right]^2 \,, \tag{4.30}$$

where t_i corresponds to the actual value of the predicted variable $\mathbf{y}$ for the instance $\mathbf{x}_i$.

Multi-layer perceptron

The multi-layer perceptron (MLP) is an artificial neural network inspired by the human brain. Its architecture is arranged in several interconnected layers (input, hidden layers, and output), which are composed of simple units known as perceptrons [19]. Each unit is characterized by an activation function $g()$ as well as by its connections to units from other layers. These connections are associated with adaptive weights (w_{ij}).

The output layer provides the response, y. In a regression task, the purpose is to estimate a target continuous variable, t. Hence, a single output unit with a linear activation function was used [88]. Additionally, a single hidden layer, composed of units with non-linear activation functions, was implemented. This configuration is known to be able to provide a universal function approximation [19]. Thus, y can be expressed as follows:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{N_H} \left[w_j \, g \left(\sum_{i=1}^{I} w_{ij} x_i + b_j \right) + b_0 \right] , \tag{4.31}$$

where $\mathbf{w}$ is a vector with all the adaptive parameters (weights and bias), w_j is the weight connecting hidden units hu_j with the output unit, b_0 is the bias associated with the output unit, w_{ij} is the weight connecting the input unit iu_i with hidden unit hu_j, and b_j is its associated bias. N_H, the number of units in the hidden layer, is a design parameter whereas I is the total number of units in the input layer, which in this case is also the number of features used to train the network. Weights are optimized by the sum of squares error function minimization. The scaled conjugate gradient method was used for this purpose [19].

Weight decay regularization was used to minimize overfitting and achieve good generalization. Thus, a penalty term (Ω) was added to the sum-of-squares error function in order to favor small weights [19]:

$$E_D = \frac{1}{2}\sum_{i=1}^{S}\left[y(\mathbf{x}_i, \mathbf{w}) - t_i\right]^2 + \upsilon \sum_i w_i^2\,, \tag{4.32}$$

where S is the total number of instances considered and υ is the regularization parameter, which has to be configured.

Radial basis function

The radial basis function (RBF) is another artificial neural network approach. It is composed of one hidden and one output layer. The output y is computed from the responses provided by the basis functions $\psi()$ from the hidden layer nodes. These functions only depend on the radial distance (typically the Euclidean distance) between the input vector $\mathbf{x}$ and a set of suitable centers c_j [19]. Since the problem is a regression task, a single output neuron with a linear activation function was used to implement the output layer. Thus, y is given by the following expression [19]:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{N_B} w_j\, \psi_j\left(\|\mathbf{x} - \mathbf{c}_j\|\right) + b, \tag{4.33}$$

where N_B is the number of basis functions (or centers), $\mathbf{c}_j$ is the center of the function ψ_j, w_j is the weight connecting ψ_j and the output neuron, and b is the bias parameter for this neuron.

A Gaussian function is commonly used for ψ [19]:

$$\psi(\mathbf{x}) = exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_j\|^2}{2\sigma_j^2}\right), \tag{4.34}$$

where σ_j is the standard deviation (or width) of each function. Thus, the numbers of centers (N_B) and their locations c_j as well as the widths of radial basis functions σ_j and the weights w_j are parameters to be optimized. N_B and σ_j are usually experimentally determined. The K-means algorithm is commonly used to optimize the location of the centers, and w_j are estimated through the sum-of-squares error minimization [19].

4.5. Conventional approach algorithm

A conventional approach algorithm, based on detecting and scoring apneic events, was also implemented for comparison purposes. Thus, a peak detection algorithm was used to locate inspiratory onsets and endings in AF time series [77]. The difference between AF values in consecutive onsets and endings locations determined the amplitude of every inspiration. According to the rules of the AASM, the algorithm scored those respiratory events which meet with *i)* a drop of 30% or more from the AF pre-event baseline and *ii)* the drop lasts 10 seconds or more [18]. The baseline was computed as the mean amplitude of the s previous inspirations [67]. Hence, s was a design parameter to be fitted in a training set. Once all events are scored, the total amount of them is divided by the recording time to obtain an AHI estimation.

4.6. Statistical analysis

4.6.1. Diagnostic ability statistics

There exist common statistics used to measure the diagnostic ability of a given model or test. Their definitions rely on the number of subjects rightly and wrongly classified. In the case of binary classification they are obtained from the corresponding confusion matrix, which compares the results of the test under study and a reference test in terms of the presence or absence of a disease. The elements of this matrix are:

a) *True positives (TP).* Number of patients (according to the reference test) which have been rightly classified by the test evaluated.

b) *False negatives (FN).* Number of patients (according to the reference test) which have been wrongly classified by the test evaluated.

c) *True negatives (TN).* Number of subjects without the disease (according to the reference test) which have been rightly classified by the test evaluated.

d) *False positives (FP).* Number of subjects without the disease (according to the reference test) which have been wrongly classified by the test evaluated.

According to these elements, the next statistics can be defined [47]:

– **Sensitivity (Se)**. Proportion of patients rightly classified, that is:

$$Se = \frac{TP}{TP + FN} \times 100. \tag{4.35}$$

– **Specificity (Sp)**. Proportion of subjects without the disease rightly classified, that is:

$$Sp = \frac{TN}{TN + FP} \times 100. \tag{4.36}$$

– **Accuracy (Acc)**. Proportion of overall subjects rightly classified. This definition is also valid for multiclass tasks. However, for binary classification it can be defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100. \tag{4.37}$$

– **Predictive values**. Given a certain class, predictive values are the proportion of subjects rightly classified among all the subjects that the test under study has assigned to that class. Thus, predictive values can be also used for multiclass problems. However, positive and negative predictive values (PPV and NPV, respectively), for binary classification, are the most common:

$$PPV = \frac{TP}{TP + FP} \times 100. \tag{4.38}$$

$$NPV = \frac{TN}{TN + FN} \times 100. \tag{4.39}$$

– **Likelihood ratios**. Positive and negative likelihood ratios ($LR+$ and $LR-$) are also common measures in binary classification problems which estimates the performance of a test in a dimensionless way. They are defined as follows:

$$LR+ = \frac{Se}{1 - Sp}, \tag{4.40}$$

$$LR- = \frac{1 - Se}{Sp}. \tag{4.41}$$

Hence, the higher the $LR+$ the higher the proportion of patients rightly classified with respect to the proportion of healthy subjects wrongly classified (desired values close to $+\infty$). Similarly, the lower the $LR-$ the lower the proportion of patients wrongly classified with respect to the healthy subjects rightly classified (desired values close to 0).

Receiver-operating characteristics (ROC) analysis

The receiver-operating characteristics analysis (ROC) measures the overall performance of a test under evaluation. It has been particularly useful as an assessment tool in clinical practice, in part due to its independence of the imbalance of the classes in a sample [134], i. e., its independence of the prevalence of the target disease. It is based on a plot which represents a Se vs. $1 - Sp$ curve, where Se and Sp result from evaluating a range of decision thresholds for the same test. Useful information can be derived from such analysis. One approach may focus on finding a suitable threshold that acts as a trade-off between FP and FN [134]. On the other hand, since the ROC curves provide a comprehensive insight of classification ability, they can be used to compare the performances of different tests in a wider sense. In this regard, a perfect discriminative test should pass through the point $Se = 1$ (or 100%), $1 - Sp = 0$. Therefore, the closer the plot to the upper left corner the higher the overall performance of a given test [134]. Figure 4.3 displays an example of ROC curve, where the closest point to (1,0) has been highlighted. The threshold associated with this point is often chosen as the optimum one when the purpose is the trade-off previously mentioned.

One common approach to avoid visual comparisons of ROC plots is to estimate the area under the curve ($AROC$). It is a way to quantify in a single number the overall performance of a test. It may range between 0 and 1. However, the less discriminative power is achieved when $AROC = 0.5$. For $AROC$ values lower than 0.5 it is enough to change the positiveness of the test in order to get $AROC$ values higher than 0.5. A perfect discriminative performance is reached when $AROC = 1$ (or $AROC = 0$). $AROC$ is interpreted as follows: given $AROC = 0.9$, one positive instance, randomly selected, would have a larger value in the test under evaluation 90% of the time comparing with a randomly chosen negative instance.

According to this interpretation, one way to measure the performance of a test through $AROC$ (considering $AROC$ values ≥ 0.5) is as follows [3]:

a) $AROC$ ranging 0.9 to 1: excellent discriminative ability.

b) $AROC$ ranging 0.8 to 0.89: good discriminative ability.

c) $AROC$ ranging 0.7 to 79: fair discriminative ability.

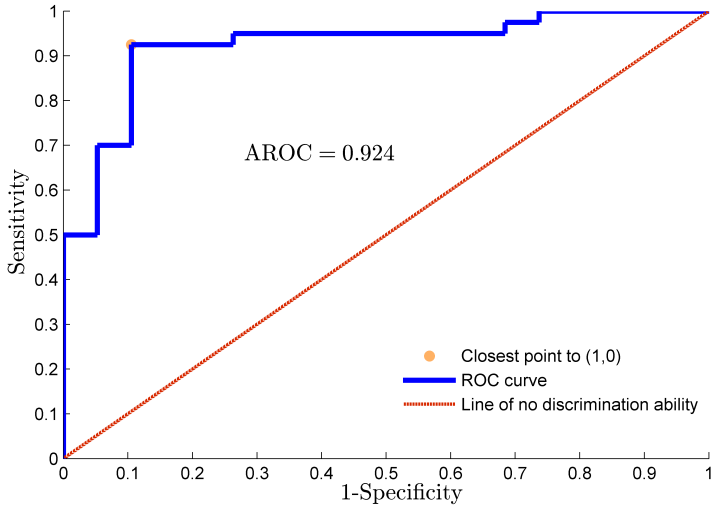

Figure 4.3: Example of a ROC curve.

d) *AROC* ranging 0.6 to 0.69: poor discriminative ability.

e) *AROC* ranging 0.5 to 0.59: bad discriminative ability.

4.6.2. Measures of agreement

The agreement between the diagnostic standard and the alternatives proposed can be measured both for classification and regression approaches.

Cohen's kappa

The Cohen's kappa index, κ, is a measure of agreement between predicted and observed classes (two or more classes) which does not consider the agreement that occurs by chance [30, 126]. It can be computed as follows [30]:

$$\kappa = \frac{p_o - p_c}{1 - p_c}, \tag{4.42}$$

where p_o is the proportion of data (or instances) in which the observed and predicted classes agree and p_c is the proportion of data (or instances) for which agreement is expected by chance [30]. The maximum value for κ is +1, meaning that there exists a perfect agreement between the observed and predicted classes. The lower limit for κ ranges between 0 and -1 depending on the marginal distributions of the classes [30]. $\kappa = 0$ means that the agreement is due entirely to chance whereas $\kappa = -1$ corresponds to total disagreement.

Intra-class correlation coefficient

The intra-class correlation coefficient (*ICC*) can be used to measure agreement between predicted and observed continuous variables, i. e., when considering regression approaches. In contrast to other popular measures, like Pearson's

correlation coefficient, ICC takes systematic error into account when assessing agreement.

There exist several versions of ICC according to its specific purpose as well as the underlying model assumed (one-way ANOVA, two-way ANOVA with and without interactions, etc). For this case, assessing agreement between variables, when no ANOVA assumptions are required and there are no replicated measurements, the next estimation is recommended [27]:

$$ICC = \frac{MS_I - MS_E}{MS_I - (J-1)MS_E + J(MS_T - MS_E)/N}, \tag{4.43}$$

where J is the number of observers, N is the number of instances considered, MS_I is the instances mean square, MS_E is the error mean square, and MS_T is the observers mean square. Theoretically, ICC ranges between 0 and +1 [124], with values close to +1 indicating a high degree of agreement between observers whereas values close to 0 indicate no agreement at all.

4.6.3. Validation

Several methods have been used in order to validate the results obtained during the study. They were chosen according to the size of the sample used at each case as well as the number of degrees of freedom needed to be adjusted for each specific problem. Thus, for the smallest samples, bootstrapping was preferred since it is known to provide good estimation of statistics when few data is available [40]. For medium-size samples, with no model free parameters to be adjusted, leave-one-out cross-validation (loo-cv) was used. By contrast, when parameter adjustment was required, a combination of Hold-out (training and test sets) with loo-cv or bootstrapping was applied, i. e., first the sample was divided into training and test sets, then loo-cv or bootstrapping was applied to the training set in order to tune the free parameters.

Hold-out

The natural way to properly estimate the performance of a given model or methodology is to divide the entire data sample into a training set, for model fitting, and an independent test set, for estimating a reliable performance. This is called the hold-out method [20, 126]. This approach imply excluding a significant amount of data from the model fitting process, which may derive into less generalizable models if the training set is not enough representative of the problem [126]. Hence, hold-out is only recommended if enough data is available. However, as previously mentioned, hold-out can take part of combined validation methodologies with the purpose of keeping some independent data, unseen for the rest of the process.

Leave-one-out cross-validation

Loo-cv is another common way of performance validation. It is based on excluding from the training process one instance at a time [20, 126]. Thus, if the sample size is N, N models are trained using $N-1$ instances. Then, these are tested on the corresponding excluded instance, which leads to the final performance estimation. One advantage of loo-cv is that each of the $N-1$ models are trained using the greatest possible amount of data. This may result

in more generalizable models [126]. By contrast, it may be computationally costly for large data sets, as well as produce pessimistic performance estimation for some specific cases [126].

Bootstrapping

The bootstrap 0.632 algorithm was also used as performance estimator when the data set was small. As previously mentioned, this procedure is known to be particularly useful in such cases [40, 126]. It is based on the *sampling with replacement* method. Thus, given a set $\mathbf{x}$ of N instances, $\mathbf{x} = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$, B new sets (bootstrap sets) $\mathbf{x}_b$ ($b = 1, 2, ..., B$) of size N are formed by resampling with replacement from the original one [126]. A uniform probability is used to randomly select the instances from $\mathbf{x}$ for each new $\mathbf{x}_b$. Hence, instances can be chosen several times for a particular $\mathbf{x}_b$. These will act as a training group and, most probably, will contain repeated instances from $\mathbf{x}$. Consequently, for each new resampling process, a number of instances from the original $\mathbf{x}$ will not be selected. These instances will act as the test group. Thus, B new sets of size N are formed, acting as training groups, and the instances not included in each case act as the corresponding test groups. Following bootstrap 0.632, a statistic S obtained from a test set would be a downward estimation of the true one [40]. Hence, both the training and the test groups are used to compute S by weighting their corresponding estimations as follows [126]:

$$S = 0.368 S_{training} + 0.632 S_{test}\,, \tag{4.44}$$

where $S_{training}$ is the statistic computed from the training set of a given $\mathbf{x}_b$ whereas S_{test} is the corresponding value from the test set. Finally, the B estimations of S are averaged to show a global performance.

4.6.4. Dealing with data imbalance: SMOTE

The high prevalence of SAHS leads to prioritize diagnosis of at-risk population [46]. Consequently, data from SAHS patients is much more available than data from no SAHS subjects. If the imbalance is too pronounced, it affects the learning process of some pattern recognition algorithms, which bias its performances towards the majority class. When considering classification into the four SAHS severity degrees (multi classification task), the imbalance is particularly marked for the group with the lowest AHI (AHI $<$ 5 e/h). Thus, the synthetic minority oversampling technique (SMOTE) was implemented to compensate for this imbalance [26].

SMOTE creates new synthetic instances on the basis of the available minority class real ones [26]. According to the number of new instances (or vectors $\mathbf{x}_i$) required for the compensation of the classes, the algorithm selects the K-nearest neighbors of each of the real ones [26]. Thus, if doubling the minority class instances is needed, K should be 1, and so on. Then, the difference between each vector $\mathbf{x}_i$ and its K-nearest neighbors is computed. These differences, multiplied by a random number in the range 0 to 1, are subsequently added to the original vector again to form new synthetic ones, whose components range between the vector considered and its corresponding K-nearest neighbors [26].

4.6.5. Statistical hypothesis tests

Statistical hypothesis testing was used to assess data normality (extracted features) as well as differences among the groups under study (SAHS-negative/ SAHS-positive, SAHS-severity degrees, etc). Thereby, Lilliefors test was applied to the extracted features in order to evaluate normality. These data did not pass the test (data not normal) and, as a consequence, non-parametric statistical significance tests were used to assess differences in the above mentioned groups. Mann-Whitney U test was used for comparisons between two classes (SAHS-negative/SAHS-positive), whereas its multiclass extension, the Kruskal-Wallis test, was used for comparisons among the four SAHS-severity groups (no-SAHS, mild, moderate, and severe).

Chapter 5

Results

This chapter summarizes the most relevant results displayed in the compendium of publications. They have been split according to the different pattern recognition approaches followed during the study, i. e., binary classification (presence or absence of SAHS), multiclass classification (prediction of SAHS severity categories), and regression (estimation of the AHI). Multiclass classification and regression was not possible in the case of children AF and adult HRV databases due to their small size. The former was only composed of 50 children whereas the number of women (54) limited the study in the second case. All results presented were obtained after conducting one of the validation methodologies described in Chapter 4.

5.1. Binary classification

5.1.1. Adults

Feature extraction: bands of interest and separability of classes

As explained in Chapter 4, several frequency and time domain features were obtained from the recordings involved in the study. Most of the spectral features were extracted from bands of special interest, which were established according to SAHS specificities (AF and RRV signals) or due to their relationships to other main body systems. The latter is the case of the HRV signal, whose spectral information has been widely associated with the behavior of the autonomic nervous system. Thus, there exist fixed frequency bands, well-established in the literature. The bands of interest used in the current study were:

- **AF** signal from **thermistor** (Figure 5.1): $0.022 - 0.059$ Hz (statistically obtained).
- **AF** signal from **nasal pressure** (Figure 5.2): $0.025 - 0.050$ Hz (derived from apneic event typical duration).
- **RRV** signal from **thermistor** (Figure 5.3): $0.09 - 0.13$ Hz (statistically obtained).
- **HRV** signal (Figure 5.4):
  - Very low frequencies (VLF) $\equiv 0{-}0.04$ Hz (associated with autonomic nervous system behavior).
  - Low frequencies (LF) $\equiv 0.04 - 0.15$ Hz (associated with sympathetic activity).
  - High frequencies (HF) $\equiv 0.15 - 0.4$ Hz (associated with parasympathetic activity).

Figures 5.1 and 5.2 show the averaged PSDs, with the corresponding bands of interest, of AF and RRV from thermistor for the SAHS-negative and SAHS-positive groups, i. e., differences are taken into account according to a binary classification task. Similarly, Figure 5.4 shows the well-established HRV bands of interest of SAHS-negative and SAHS positive groups separated by gender. Finally, Figure 5.3 displays the band of interest for nasal-pressure AF, taking into account the four severity degrees of SAHS.

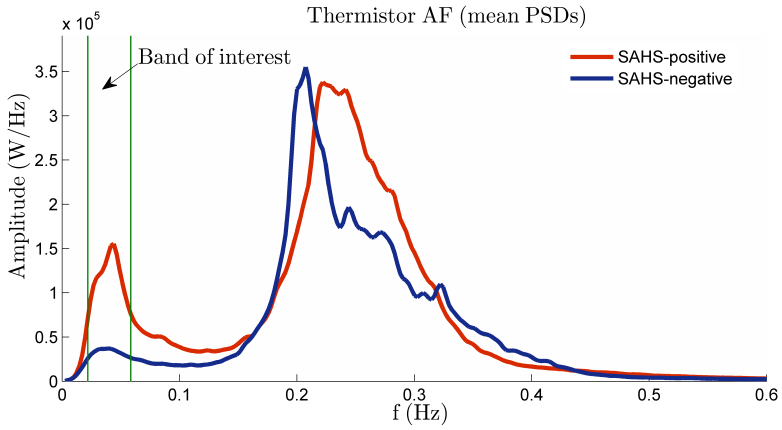

Figure 5.1: PSDs and band of interest in thermistor AF [65].

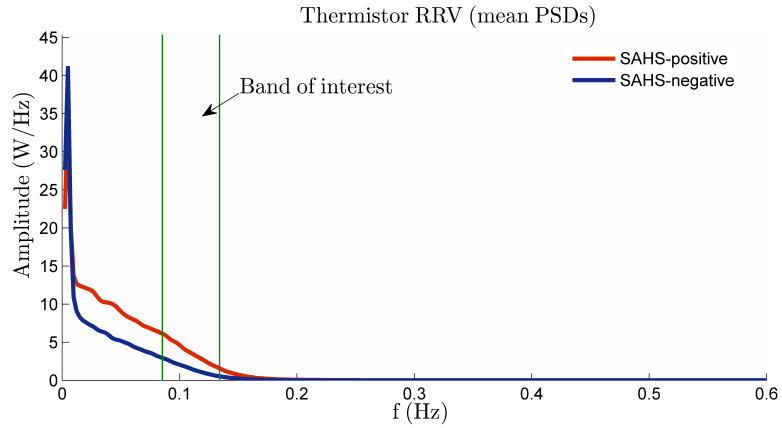

Figure 5.2: PSDs and band of interest in thermistor RRV [65].

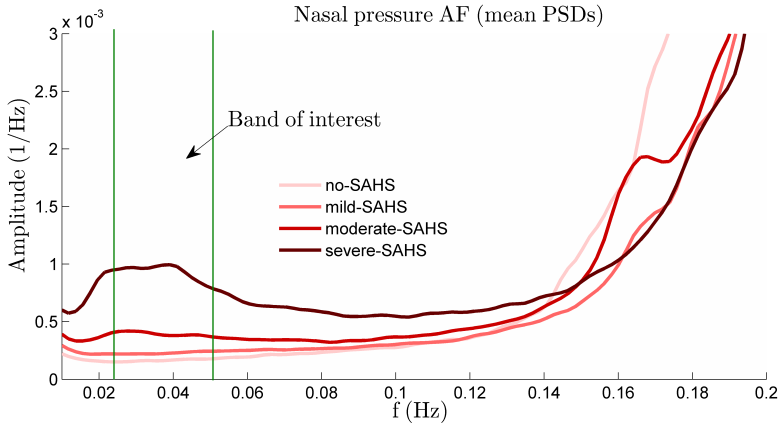

Figure 5.3: PSDs and band of interest in nasal-pressure AF [62].

Table 5.1 displays statistical moments in time domain, non-linear measures, and spectral features, extracted from AF recordings obtained through a ther-

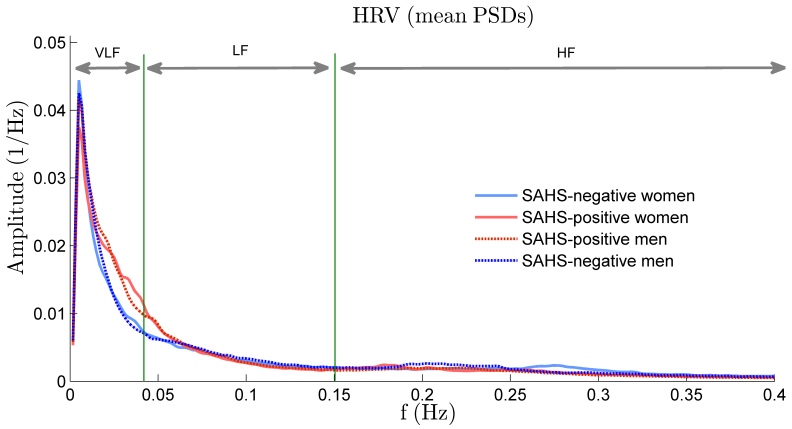

Figure 5.4: PSDs and bands of interest in HRV [63].

mistor, for SAHS-negative and SAHS-positive subjects (mean ± standard deviation). p-values between the two groups are also shown (significance level p-value < 0.01). It can be observed that none of the time-domain features showed statistically significant differences and only one of the spectral features from the whole spectrum did it (M_{f3}). By contrast, 5 out of 6 spectral features obtained from the band of interest showed p-value < 0.01. These results highlight that, for the AF signal, most of the information about SAHS is comprised within the spectral band of interest.

Table 5.1: Features extracted from the **AF** signal obtained through a **thermistor** sensor for the SAHS-negative and the SAHS-positive groups (mean ± standard deviation). p-values between the two groups are also shown for each feature (significance level p-value < 0.01). ns: not significant (p-value ≥ 0.01). X_B refers to features extracted from the band of interest.

Features	**SAHS-negative**	**SAHS-positive**	p-**value**
M_{t1}	0.04 ± 0.11	0.06 ± 0.21	ns
M_{t2}	179.9 ± 91.4	190.4 ± 80.3	ns
M_{t3}	0.28 ± 0.31	0.28 ± 0.24	ns
M_{t4}	11.74 ± 21.40	8.25 ± 15.83	ns
CTM	0.628 ± 0.185	0.635 ± 0.184	ns
LZC	0.283 ± 0.027	0.279 ± 0.029	ns
$ApEn$	0.435 ± 0.074	0.412 ± 0.073	ns
M_{f1}	$4.9 \cdot 10^4 \pm 6.4 \cdot 10^4$	$8.4 \cdot 10^3 \pm 8.0 \cdot 10^3$	ns
M_{f2}	0.06 ± 0.21	$4.7 \cdot 10^4 \pm 5.1 \cdot 10^4$	ns
M_{f3}	8.94 ± 2.19	8.02 ± 1.77	< 0.01
M_{f4}	96.98 ± 46.25	80.08 ± 35.18	ns
MA	$68.5{\cdot}10^4 \pm 96.7{\cdot}10^4$	$59.9{\cdot}10^4 \pm 68.2{\cdot}10^4$	ns
WD	0.808 ± 0.21	0.798 ± 0.022	ns
M_{f1B}	$3.4 \cdot 10^4 \pm 2.3 \cdot 10^4$	$9.9 \cdot 10_4 \pm 12.9 \cdot 10^4$	< 0.01
M_{f2B}	$7.10 \cdot 10^3 \pm 0.8 \cdot 10^3$	$39.6 \cdot 10_3 \pm 10.5 \cdot 10^3$	< 0.01
M_{f3B}	-0.451 ± 0.634	0.042 ± 0.689	< 0.01
M_{f4B}	2.675 ± 1.026	2.402 ± 0.905	ns
MA_B	$4.4 \cdot 10^4 \pm 3.7 \cdot 10^4$	$16.8{\cdot}10^4 \pm 30.7{\cdot}10^4$	< 0.01
WD_B	0.063 ± 0.0372	0.109 ± 0.0608	< 0.01

Table 5.2: Features extracted from the **RRV** signal obtained through a **thermistor** sensor for the SAHS-negative and the SAHS-positive groups (mean ± standard deviation). p-values between the two groups are also shown for each feature (significance level p-value < 0.01). ns: not significant (p-value ≥ 0.01). X_B refers to features extracted from the band of interest.

Features	**SAHS-negative**	**SAHS-positive**	p-**value**
M_{t1}	3.64 ± 0.50	3.66 ± 0.51	ns
M_{t2}	0.85 ± 0.28	1.04 ± 0.33	< 0.01
M_{t3}	0.03 ± 1.11	0.81 ± 1.39	< 0.01
M_{t4}	9.9 ± 8.0	12.9 ± 12.7	ns
CTM	0.998 ± 0.002	0.989 ± 0.017	< 0.01
LZC	0.975 ± 0.037	0.992 ± 0.035	< 0.01
$ApEn$	1.44 ± 0.075	1.46 ± 0.072	ns
M_{f1}	0.15 ± 0.10	0.22 ± 0.14	< 0.01
M_{f2}	1.4 ± 1.04	1.69 ± 1.03	ns
M_{f3}	17.51 ± 4.51	12.58 ± 4.28	< 0.01
M_{f4}	418.02 ± 184.65	232.60 ± 158.77	< 0.01
MA	39.60 ± 32.36	37.55 ± 23.52	ns
WD	0.908 ± 0.008	0.904 ± 0.009	< 0.01
M_{f1B}	1.39 ± 0.91	2.96 ± 2.53	< 0.01
M_{f2B}	0.56 ± 0.42	0.98 ± 0.86	< 0.01
M_{f3B}	0.23 ± 0.38	0.16 ± 0.39	ns
M_{f4B}	1.99 ± 0.40	2.0 ± 0.43	ns
MA_B	2.34 ± 1.48	4.61 ± 3.72	< 0.01
WD_B	0.16 ± 0.008	0.14 ± 0.07	ns

In the case of the features extracted from RRV (derived from thermistor AF), Table 5.2 shows a different behavior. Thus 4 out of 7 time-domain features reached statistically significant differences (2 statistical moments and 2 non-linear features). Additionally, 4 out of 6 spectral features obtained from the whole spectrum, as well as 3 out of 6 spectral features from the band of interest, also showed p-value < 0.01. In contrast to thermistor AF, RRV time-domain (common statistics and non-linear features) and spectral features outside the spectral band of interest, were also able to summarize useful information about SAHS.

Table 5.3 shows mean ± standard deviation and p-values of the features extracted from nasal-pressure AF of SAHS-negative and SAHS-positive subjects. Similarly to thermistor AF, the features from the band of interest (7 out of 9) showed statistical significant differences between both groups. By contrast, CTM was the only nonlinear feature which reached p-value < 0.01.

Finally, Table 5.4 shows the values of $SpecEn$ and $Power$ (mean ± standard deviation), obtained from the common HRV spectral bands of interest (VLF, LF, HF, and the whole band VLF-HF) and separated by gender. As can be observed, all the classic power parameters, but the power in the HF band of men (P^m_{HF}), did not show differences between SAHS-negative and SAHS-positive subjects. By contrast, $SpecEn$ in VLF and LF showed statistically significant differences both in women and men. Twenty five scales of $SampEn$ (Multi-scale entropy) were also obtained from the HRV recordings. Figure 5.5 displays the mean value of each scale for women and men SAHS-negative and SAHS-positive groups. Significant p-values (< 0.01) are also shown. Only scale 13th reached statistically significant differences between SAHS-negative and

Table 5.3: Features extracted from the **AF** signal obtained through a **nasal prong** sensor for the SAHS-negative and the SAHS-positive groups (mean ± standard deviation). p-values between the two groups are also shown for each feature (significance level p-value < 0.01). ns: not significant (p-value ≥ 0.01). X_B refers to features extracted from the band of interest.

Features	**SAHS-negative**	**SAHS-positive**	p-**value**
CTM	0.999 ± 0.001	0.997 ± 0.002	< 0.01
LZC	0.0567 ± 0.008	0.0572 ± 0.006	ns
$SampEn$	0.061 ± 0.014	0.060 ± 0.015	ns
M_{f1B}	$1.981 \cdot 10^{-4} \pm 0.989 \cdot 10^{-4}$	$6.348 \cdot 10^{-4} \pm 5.818 \cdot 10^{-4}$	< 0.01
M_{f2B}	$0.259 \cdot 10^{-4} \pm 0.169 \cdot 10^{-4}$	$1.522 \cdot 10^{-4} \pm 2.147 \cdot 10^{-4}$	< 0.01
M_{f3B}	0.194 ± 0.503	0.309 ± 0.662	ns
M_{f4B}	2.201 ± 0.552	2.453 ± 0.900	ns
MA_B	$2.405 \cdot 10^{-4} \pm 1.176 \cdot 10^{-4}$	$8.992 \cdot 10^{-4} \pm 9.081 \cdot 10^{-4}$	< 0.01
mA_B	$1.599 \cdot 10^{-4} \pm 0.806 \cdot 10^{-4}$	$4.387 \cdot 10^{-4} \pm 3.59 \cdot 10^{-4}$	< 0.01
WD_B	0.049 ± 0.023	0.073 ± 0.050	< 0.01
MF_B	0.0376 ± 0.001	0.0367 ± 0.002	< 0.01
$SpecEn_B$	0.996 ± 0.003	0.990 ± 0.015	< 0.01

SAHS-positive men. By contrast, 15 out of the 25 scales in women reached p-values < 0.01.

Table 5.4: Features extracted from the **HRV** signal for the SAHS-negative and the SAHS-positive groups (mean ± standard deviation). p-values between the two groups are also shown for each feature (significance level p-value < 0.01). ns: not significant (p-value ≥ 0.01). VLF: very low frequency; LF: low frequency; HF: high frequency. w: women; m: men.

Features	**SAHS-negative**	**SAHS-positive**	p-**value**
$SpecEn^w_{VLF}$	0.959 ± 0.020	0.971 ± 0.011	< 0.01
$SpecEn^w_{LF}$	0.984 ± 0.011	0.959 ± 0.028	< 0.01
$SpecEn^w_{HF}$	0.979 ± 0.021	0.970 ± 0.022	ns
$SpecEn^w_{VLF-HF}$	0.899 ± 0.060	0.863 ± 0.051	ns
P^w_{VLF}	0.425 ± 0.153	0.489 ± 0.170	ns
P^w_{LF}	0.236 ± 0.052	0.241 ± 0.068	ns
P^w_{HF}	0.234 ± 0.087	0.199 ± 0.118	ns
$P^w_{LF/HF}$	1.164 ± 0.514	1.639 ± 1.065	ns
$SpecEn^m_{VLF}$	0.958 ± 0.020	0.966 ± 0.018	< 0.01
$SpecEn^m_{LF}$	0.983 ± 0.012	0.960 ± 0.035	< 0.01
$SpecEn^m_{HF}$	0.983 ± 0.015	0.976 ± 0.023	ns
$SpecEn^m_{VLF-HF}$	0.900 ± 0.053	0.873 ± 0.061	ns
P^m_{VLF}	0.437 ± 0.167	0.503 ± 0.168	ns
P^m_{LF}	0.250 ± 0.051	0.250 ± 0.068	ns
P^m_{HF}	0.228 ± 0.102	0.183 ± 0.108	< 0.01
$P^m_{LF/HF}$	1.407 ± 0.874	1.898 ± 1.247	*ns*

Feature selection: optimum feature subsets

The feature selection stage was conducted through the SLR-FSBE algorithm for thermistor AF and RRV, as well as HRV. In the case of nasal-pressure AF, a filter method (FCBF) was preferred since data were used to feed differ-

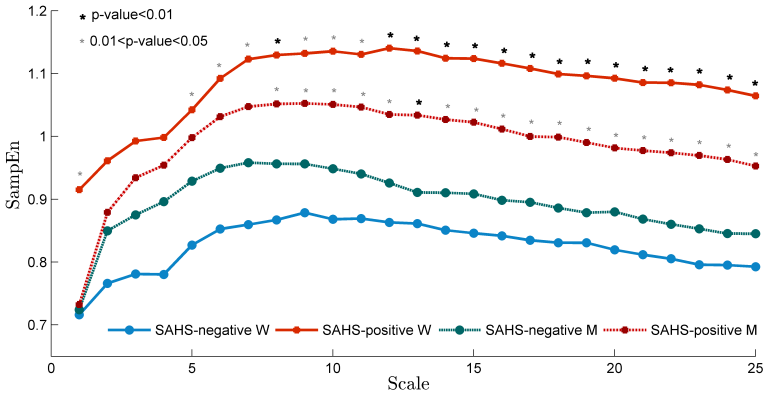

Figure 5.5: Mean value of each *SampEn* scale for women and men divided into SAHS groups.

ent classifiers following both a binary and a multiclass approach in the same study. Thus, Table 5.5 displays the optimum set of features selected by the corresponding automatic algorithms for each signal. It can be observed that spectral and non-linear features were selected in the cases of RRV, nasal prong AF, and HRV, highlighting the complementarity of these approaches to characterize SAHS. Conversely, only spectral features were selected in the case of thermistor AF. This different behavior agrees with the results showed in Table 5.1, where none of the time-domain features exhibited statistical significant differences.

Results in Table 5.5 also show that thermistor AF and RRV signals contain complementary information since the SLR-FSBE algorithm automatically selected features from both of them when applied to a joint set of features (M_{f3} from RRV, and MA and M_{f1B} from thermistor AF). Additionally, $SpecEn_{VLF}$, $SpecEn_{LF}$, and $SampEn_2$ were common in the three optimum sets obtained for the HRV features (from women, men, and women and men joint, respectively). However, different tendencies for women and men were observed in the remaining MsE scales selected. Thus, for female subjects only scales below the 8th were selected, whereas in the case of men the remaining features were selected from scales above the 9th.

Pattern recognition: binary classification

The optimum sets of features obtained for each signal were used to feed different classifiers in order to test their diagnostic performances. Table 5.6 shows the statistics which measure the diagnostic ability of each of these classifiers when detecting the presence or absence of SAHS (binary classification). As previously explained, Acc, $AROC$, and κ are statistics which show a global diagnostic behavior. By contrast, Se, PPV, and $LR+$ are focused on the performance when classifying SAHS-positive subjects, whereas Sp, NPV, and $LR-$ serve a similar function for SAHS-negative ones.

All the models trained with features from AF (including RRV) reached higher diagnostic ability in terms of Acc and κ than the LR model trained with non gender-segregated features from HRV ($LR_{w,m}$). Additionally, most

Table 5.5: Features automatically selected for the binary classification task

Signal	Method	#Features	Features
$AF(Th)$	SLR-FSBE	4	WD_B, M_{f1B}, MA, M_{f2B}
$RRV(Th)$	SLR-FSBE	2	M_{f3}, CTM
$AF + RRV(Th)$	SLR-FSBE	3	M_{f3}^{RRV}, MA^{AF}, M_{f1B}^{AF}
$AF(NPP)$	FCBF	10	M_{f1B}, MA_B, mA_B, M_{f2B}, $SpecEn_B$, MF_B, M_{f4B}, CTM, LZC, $SampEn$
HRV^w	SLR-FSBE	5	$SpecEn_{VLF}$, $SpecEn_{LF}$, $SampEn_1$, $SampEn_2$, $SampEn_7$
HRV^m	SLR-FSBE	12	$SpecEn_{VLF}$, $SpecEn_{LF}$, $SpecEn_{VLF-HF}$, $SampEn_2$, $SampEn_{10}$, $SampEn_{13}$, $SampEn_{16}$, $SampEn_{17}$,$SampEn_{20-23}$
$HRV^{w,m}$	SLR-FSBE	15	$SpecEn_{VLF}$, $SpecEn_{LF}$, $SpecEn_{VLF-HF}$, $SampEn_2$, $SampEn_7$, $SampEn_9$, $SampEn_{11}$, $SampEn_{13}$, $SampEn_{14}$, $SampEn_{17}$,$SampEn_{19-23}$

Table 5.6: Diagnostic performance of different binary classifiers and the classic event detection algorithm in adult databases. Se: sensitivity (%); Sp: specificity (%); Acc: accuracy; PPV: positive predictive value (%); NPV: negative predictive value (%); $LR+$: positive likelihood ratio; $LR-$: negative likelihood ratio; AROC: area under ROC curve; κ: Cohen's kappa coefficient. Th: thermistor; NP: nasal prong. w: women; m: men; w,m: women and men joined.

Classifier	Signal	Se	Sp	Acc	PPV	NPV	$LR+$	$LR-$	AROC	κ
LR	AF(Th)	84.0	70.8	79.7	85.7	68.0	2.88	0.23	0.889	0.543
LR	RRV(Th)	84.0	58.3	75.7	80.8	63.6	2.10	0.27	0.850	0.433
LR	AF,RRV(Th)	88.0	70.8	82.4	86.3	73.9	3.01	0.17	0.903	0.595
LR	AF(NP)	83.5	80.0	82.5	91.6	65.1	4.17	0.21	0.915	0.593
LDA	AF(NP)	72.5	74.3	73.0	88.0	51.0	2.82	0.37	0.835	0.410
$CART$	AF(NP)	85.7	68.6	81.0	87.6	64.9	2.73	0.21	0.830	0.593
$AB - LDA$	AF(NP)	86.8	77.1	84.1	90.8	69.2	3.79	0.17	0.855	0.618
$AB - CART$	AF(NP)	89.0	80.0	86.5	92.0	73.7	4.45	0.14	0.950	0.672
$Event - det.$	AF(NP)	75.8	54.3	69.0	81.2	46.3	1.66	0.45	0.635	0.286
LR_w	HRV	80.8	89.3	85.2	87.5	83.3	7.60	0.22	0.951	0.703
LR_m	HRV	87.1	56.1	77.6	81.8	65.7	1.98	0.23	0.895	0.450
$LR_{w,m}$	HRV	79.8	59.4	72.3	77.2	63.1	1.97	0.34	0.885	0.397

of the remaining class-specific statistics from AF-related signals also improved those from the HRV model. This suggests that, in the case of SAHS, AF has higher global diagnostic potential than a widely studied signal such as HRV when gender is not taken into account. Moreover, $AB-CART$ reached the highest global statistics in terms of *Acc*, *AROC*, and κ (86.5%, 0.993, and 0.672, respectively). Similarly, $AB-CART$ obtained the highest Se (89.0%), Sp (80.0%), PPV (92.0%), and $LR+$ (4.45), as well as the lowest $LR-$ (0.14). These results suggest this classifier as the one with the highest diagnostic ability. By contrast, the lowest diagnostic ability was shown by the conventional event-detection algorithm.

On the other hand, very similar results were obtained when comparing LR models obtained from thermistor and nasal prong information: 82.4% vs. 82.5% *Acc*, 0.903 vs. 0.917 *AROC*, and 0.595 vs. 0.593 κ, suggesting a similar diagnostic potential. However, RRV information was needed to improve the LR model trained with features from thermistor. It is also noteworthy that both AB models ($AB-LDA$ and $AB-CART$) widely outperformed the corresponding single LDA and $CART$ models, which highlights the usefulness of the ensemble learning approach.

As previously explained, one of the limitations of the HRV signal is gender specificities. Hence, we also considered models trained with HRV features from women and men, separately (LR_w and LR_m). In this regard, both LR_w and LR_m showed higher diagnostic ability in terms of *Acc*, *AROC*, and κ than the joint model. The results reached by LR_w were particularly high since it achieved the highest overall κ (0.703), Sp (89.3%), NPV (82.3%), and $LR+$ (7.6). These indicated higher discriminative power than $AB-CART$ in the case of SAHS-negative subjects (SAHS-negative women in the case of LR_w).

5.1.2. Children: an at-home study

Children are particularly sensitive to changes in the sleep environment. Additionally, they do not tolerate well all the body sensors needed to record the PSG signals. Consequently, an at-home approach was preferred to study the diagnostic ability of AF when considering pediatric SAHS. AF was obtained by means of a 6-channel polygraph with ability to record thermistor AF, thoracic movements, body sensor, snoring sounds, and heart rate and oxygen saturation from oximetry.

Feature extraction: bands of interest and separability of classes

In contrast to adults, two spectral bands of interest were statistically determined in the case of thermistor AF: $0.119-0.192$ Hz (BW_1) and $0.784-0.890$ Hz (BW_2) (see Figure 4.2). Table 5.7 shows the values of the extracted features from these bands for the SAHS-negative ($AHI < 3$ e/h) and the SAHS-positive groups ($AHI \geq 3$ e/h) (mean $\pm$ standard deviation). It also shows the corresponding p-values. MA, mA, and M_{f1}, both in BW1 and BW2, were significantly higher in SAHS-positive than in SAHS-negative children (p-value <0.01). By contrast, no differences in M_{f2}, M_{f3} and M_{f4} were found.

Table 5.7: Features extracted from the **AF** signal obtained through a **thermistor** sensor for the SAHS-negative and the SAHS-positive children (mean ± standard deviation). p-values between the two groups are also shown for each feature (significance level p-value < 0.01). ns: not significant (p-value ≥ 0.01).

Features	**SAHS-negative**	**SAHS-positive**	p-**value**
BW_1			
$MA\,(10^{-3})$	1.90 ± 2.00	4.10 ± 5.70	< 0.01
$mA\,(10^{-4})$	5.80 ± 2.90	13.00 ± 8.00	< 0.01
$M_{f1}\,(10^{-3})$	1.10 ± 0.80	2.20 ± 1.70	< 0.01
$M_{f2}\,(10^{-4})$	3.20 ± 4.30	7.50 ± 16.60	ns
$M_{f3}\,(10^{-1})$	6.50 ± 0.50	6.50 ± 0.50	ns
$M_{f4}\,(10^{0})$	3.10 ± 1.00	3.00 ± 1.00	ns
BW_2			
$MA\,(10^{-3})$	0.14 ± 0.21	0.21 ± 0.23	< 0.01
$mA\,(10^{-4})$	0.38 ± 0.28	0.87 ± 0.78	< 0.01
$M_{f1}\,(10^{-3})$	0.07 ± 0.07	0.14 ± 0.13	< 0.01
$M_{f2}\,(10^{-4})$	0.25 ± 0.48	0.28 ± 0.31	ns
$M_{f3}\,(10^{-1})$	5.7 ± 4.6	4.7 ± 4.8	ns
$M_{f4}\,(10^{0})$	2.8 ± 0.9	3.1 ± 1.2	ns

Table 5.8: Features automatically selected from the **AF** signal and $ODI3$

Signals	#**Features**	**Features**
AF Spectral features	3	mA^{BW_1}, $M_{f3}^{BW_2}$,$M_{f4}^{BW_2}$
AF Spectral features + $ODI3$	4	$ODI3$, mA^{BW_1}, $M_{f4}^{BW_1}$, $M_{f3}^{BW_2}$

Feature selection: optimum feature subsets from children recordings

The SLR-FSBE algorithm was used twice in the case of children database. First, it was only applied to the 12 spectral features obtained from BW_1 and BW_2. Then, the clinical variable $ODI3$ (from SpO_2) was also included in the selection process. Table 5.8 displays the features selected in each case. It can be observed that features from BW_1 and BW_2 are selected in both cases, highlighting their information about SAHS as complementary. Additionally, features from both bands are also selected along with $ODI3$, suggesting that they complement the information of this clinical parameter as well.

Pattern recognition: binary classification in children

Table 5.9 shows the diagnostic ability of the $ODI3$ clinical parameter as well as the LR models trained with the optimum features selected in the previous step. Both LR_{AF} and $LR_{AF+ODI3}$ outperformed $ODI3$ in terms of Acc, $AROC$, and κ. Particularly high was the diagnostic ability of the $LR_{AF+ODI3}$, which showed the highest performance at each statistic. These results reveal the usefulness of data obtained at home from only 2-channels (thermistor AF and SpO_2) to help in pediatric SAHS diagnosis.

Table 5.9: Diagnostic performance of LR binary classifiers and single $ODI3$ in the children database. Se: sensitivity (%); Sp: specificity (%); Acc: accuracy; PPV: positive predictive value (%); NPV: negative predictive value (%); $LR+$: positive likelihood ratio; $LR-$: negative likelihood ratio; AROC: area under ROC curve; κ: Cohen's kappa coefficient.

Classifier	Se	Sp	Acc	PPV	NPV	$LR+$	$LR-$	AROC	κ
$ODI3$	70.9	80.3	75.3	81.9	72.8	3.60	0.36	0.676	0.482
LR_{AF}	79.2	79.4	79.1	81.2	78.8	3.84	0.26	0.875	0.599
$LR_{AF+ODI3}$	85.9	87.4	86.3	88.4	85.8	6.82	0.16	0.947	0.720

5.2. Multiclass classification

Only nasal prong AF has been evaluated in the multiclass task (no-SAHS, mild-SAHS, moderate-SAHS, severe-SAHS). This is the only database, among those used in the study, large enough (317 subjects) to ensure a minimum number of subjects in the no-SAHS class.

Feature extraction: separability of classes

The spectral band of interest for nasal prong AF used in the multiclass task was the same than in the case of binary classification task, i. e., $0.025 - 0.050$ Hz., which was derived from the apneic event typical duration. Table 5.10 displays the values of each of the features extracted for the four SAHS severity degrees (mean $\pm$ standard deviation). Four out of the 9 spectral features (MA, mA, M_{f1B}, and M_{f2B}), as well as CTM, showed statistically significant differences among classes after the Bonferroni correction (p-value < 0.01). These spectral features showed higher values as the SAHS severity increased. An opposite tendency was shown by CTM values. Hence, the variability also increased with the severity of SAHS.

Table 5.10: Features extracted from **AF** signal obtained through a **nasal prong** sensor for the four severity degrees of SAHS (mean $\pm$ standard deviation). p-values between the four groups are also shown for each feature (significance level p-value < 0.01). ns: not significant (p-value ≥ 0.01). X_B refers to features extracted from the band of interest.

Features	no-SAHS	mild	moderate	severe	p-value
$CTM(10^{-1})$	9.993±0.007	9.988±0.015	9.987±0.009	9.963±0.023	< 0.01
LZC	0.057±0.009	0.057±0.007	0.057±0.006	0.058±0.007	ns
$SampEn$	0.059±0.012	0.063±0.014	0.062±0.016	0.058±0.014	ns
$M_{f1B}(10^{-4})$	1.670±0.912	2.296±1.131	3.900±1.886	9.400±7.295	< 0.01
$M_{f2B}(10^{-5})$	2.140±1.424	3.193±2.428	7.418±8.268	24.86±27.77	< 0.01
M_{f3B}	0.190±0.540	0.259±0.512	0.149±0.619	0.429±0.689	ns
M_{f4B}	2.154±0.590	2.269±0.569	2.298±0.637	2.608±1.115	ns
$MA_B(10^{-4})$	2.012±1.091	2.854±1.460	5.148±3.134	13.74±11.36	< 0.01
$mA_B(10^{-4})$	1.359±0.729	1.849±0.930	2.903±1.294	6.225±4.498	< 0.01
WD_B	0.046±0.019	0.052±0.029	0.063±0.041	0.086±0.056	ns
MF_B	0.038±0.001	0.038±0.002	0.037±0.002	0.036±0.002	ns
$SpecEn_B$	0.996±0.003	0.996±0.005	0.992±0.017	0.988±0.013	ns

Feature selection: optimum set feature subsets

The optimum feature subsets for nasal prong AF in the multiclass task was the same than in the case of the binary classification task: M_{f1}, MA, mA, M_{f2}, $SpecEn$, MF, M_{f4}, CTM, LZC, $SampEn$. It was automatically obtained through the FCBF algorithm.

Multiclass pattern recognition: diagnostic performance when estimating SAHS severity degrees

Tables 5.11 to 5.17 show the confusion matrix and the corresponding diagnostic performance statistics for the classic event-detection algorithm, LR, single LDA, single CART, AB-LDA, and AB-CART classifiers when predicting the four SAHS severity degrees. The diagnostic performance statistics were computed for the three AHI cutoffs corresponding to the thresholds of the severity degrees, i. e., 5 e/h, 15 e/h, and 30 e/h. Since LR is essentially a binary classifier, its performance was evaluated following the well-known *one vs. all* strategy.

As displayed in the confusion matrices, the overall accuracy of the models, derived from the corresponding main diagonals, was low: Event-detection 39.7%, LR 57.4%, LDA 47.6 %, CART 54.8 %, AB-LDA 60.3%, and AB-CART 57.4%. Classification of mild and moderate subjects were particularly poor for all the models. Consistent with these overall accuracies, κ values were also low. By contrast, the diagnostic performance increased when assessing the predictions of the models in each of the AHI severity cutoffs (5 e/h, 15 e/h, and 30 e/h). Thus, high diagnostic accuracies were reached by AB-LDA and AB-CART. They outperformed LR and the Event-detection algorithm in terms of Acc and κ when assessing the three AHI cutoffs. Moreover, AB-LDA widely improved the overall performance of single LDA and the Acc for each AHI cutoff. AB-CART also improved the overall performance of CART, as well as the Acc for 5 e/h and 30 e/h. However, single CART outperformed the Acc of AB-CART when considering 15 e/h as the AHI cutoff.

Table 5.11: Four-class confusion matrix for the classic **Event-detection** algorithm.

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**2**	4	3	1
	mild	12	**16**	5	5
	moderate	1	5	**5**	5
	severe	3	17	15	**27**

Table 5.12: Four-class confusion matrix for the **LR** classifier (one vs. all strategy).

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**8**	0	2	0
	mild	14	**8**	10	6
	moderate	3	3	**10**	6
	severe	2	1	7	**52**

Table 5.13: Four-class confusion matrix for the single **LDA** classifier.

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**8**	0	2	0
	mild	13	**7**	13	5
	moderate	5	2	**6**	3
	severe	4	5	14	**39**

Table 5.14: Four-class confusion matrix for the single **CART** classifier.

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**7**	2	1	0
	mild	16	**11**	9	2
	moderate	4	3	**6**	3
	severe	3	0	14	**45**

Table 5.15: Four-class confusion matrix for the **AB-LDA** classifier.

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**8**	0	2	0
	mild	11	**16**	8	3
	moderate	3	4	**6**	3
	severe	1	3	12	**46**

Table 5.16: Four-class confusion matrix for the **AB-CART** classifier.

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**8**	1	1	0
	mild	14	**8**	12	4
	moderate	3	2	**6**	5
	severe	0	3	9	**50**

5.3. Regression

AF and RRV features from thermistor were also evaluated in a regression task, i. e., when estimating AHI. As previously explained, MLR, RBF, and MLP pattern recognition techniques were assessed for this purpose. The event-detection algorithm was also assessed in this database.

Feature extraction and spectral bands of interest

The features extracted were the same than in the binary classification task both for thermistor AF and RRV (Tables 5.1 and 5.2). Accordingly, the bands of interests were also the same (0.022-0.059 Hz. for AF and 0.09-0.13 Hz. for RRV).

Table 5.17: Diagnostic performance of the multiclass methods for the AHI cutoffs = 5, 15, and 30 e/h. Se: sensitivity (%); Sp: specificity (%); Acc: accuracy; κ: Cohen's kappa coefficient.

Classifier	**AHI cutoff**	**Se**	**Sp**	**Acc**	κ
$Event-det.$	5	86.2	20.0	81.0	0.152
	15	66.7	70.8	68.3	
	30	43.5	82.8	63.5	
LR (one *vs.* all)	5	83.6	80.0	83.3	0.370
	15	88.5	62.5	78.6	
	30	83.9	81.3	82.5	
LDA	5	81.0	80.0	81.0	0.281
	15	79.5	58.3	71.4	
	30	62.9	87.5	75.4	
$CART$	5	82.8	70.0	81.7	0.369
	15	87.2	75.0	**82.5**	
	30	72.6	92.2	82.5	
$AB-LDA$	5	87.1	80.0	**86.5**	**0.432**
	15	85.9	72.9	81.0	
	30	74.2	90.6	82.5	
$AB-CART$	5	85.3	80.0	84.9	0.381
	15	89.7	64.6	80.2	
	30	80.6	85.9	**83.3**	

Feature selection: optimum feature subsets

Since several pattern recognition techniques were evaluated, the FCBF algorithm was chosen to automatically select the 3 optimum sets from AF, RRV, and AF+RRV features. As previously stated, this algorithm ensures a selection process independent from subsequent analyses. Table 5.18 shows the optimum feature subsets in each case. As in the case of the SLR-FSBE selection algorithm used for binary classification, the optimum feature subsets obtained with FCBF highlighted the complementarity of thermistor AF and RRV data, as well as the complementarity of linear and non-linear analyses.

Table 5.18: Features automatically selected by the FCBF algorithm in the regression task.

Signal	**#Features**	**Features**
AF	7	WD_B, M_{f1B}, $ApEn$, CTM, M_{f3B}, WD, M_{f1}
RRV	5	CTM, M_{f1B}, M_{t3}, M_{f3}, M_{f1}
$AF+RRV$	10	CTM^{RRV}, WD_B^{AF}, M_{f1B}^{RRV}, M_{t3}^{RRV}, M_{f1B}^{AF}, M_{f3}^{RRV}, M_{f1}^{RRV}, $ApEn^{AF}$, CTM^{AF}, LZC^{RRV}

Pattern recognition: AHI regression, agreement, and diagnostic performance

Each of the optimum feature subsets were used to train MLR, RBF, and MLP models in order to estimate AHI. The ICC was used to measure agreement between actual and estimated AHIs. Then, trained models showing the highest ICCs were also assessed in terms of diagnostic ability in a test set for AHI cutoffs = 5, 10, 15, and 30 e/h. Tables 5.19 to 5.22 show the confusion matrices for each AHI estimation method in this test set. Additionally, Table 5.23 displays the diagnostic statistics for the Event-detection algorithm as well as the MLR, RBF, and MLP models which showed the highest ICC in the training set.The corresponding ICC in the test set were:

- **MLR**$_{AF+RRV}$: ICC= 0.809.
- **RBF**$_{AF}$: ICC= 0.748.
- **MLP**$_{AF+RRV}$: ICC= 0.849.
- **Event-detection**: ICC= 0.840.

Table 5.19: Confusion matrix for the AHI estimation provided by the classic **Event-detection** algorithm.

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**2**	9	0	0
	mild	1	**7**	6	3
	moderate	1	8	**6**	7
	severe	1	1	3	**4**

Table 5.20: Confusion matrix for the AHI estimation provided by the **MLR** method.

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**5**	4	2	0
	mild	8	**5**	2	2
	moderate	4	7	**9**	2
	severe	1	2	1	**5**

Table 5.21: Confusion matrix for the AHI estimation provided by the **RBF** artificial neural network.

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**5**	4	1	1
	mild	1	**7**	9	0
	moderate	0	3	**15**	4
	severe	0	0	2	**7**

Table 5.22: Confusion matrix for the AHI estimation provided by the **MLP** artificial neural network.

		Predicted			
		no-SAHS	mild	moderate	severe
Gold standard	no-SAHS	**3**	7	1	0
	mild	4	**7**	4	2
	moderate	0	5	**13**	4
	severe	0	0	1	**8**

Table 5.23: Diagnostic performance of the AHI estimations for the AHI cutoffs = 5, 10, 15, and 30 e/h. Se: sensitivity (%); Sp: specificity (%); Acc: accuracy; PPV: positive predictive value (%); NPV: negative predictive value (%); $LR+$: positive likelihood ratio; $LR-$: negative likelihood ratio; AROC: area under ROC curve; κ: Cohen's kappa coefficient.

Classifier	**Opt. set**	**AHI cutoff**	**Se**	**Sp**	**Acc**	**AROC**	κ
Event − det.	AF	5	93.8	18.2	79.7	0.823	0.073
		10	87.5	57.9	78.0	0.833	
		15	64.5	67.9	66.1	0.867	
		30	44.4	80.0	74.6	**0.982**	
MLR	AF, RRV	5	72.9	45.5	67.8	0.653	0.202
		10	90.0	63.2	81.4	0.607	
		15	54.8	78.6	66.1	0.504	
		30	55.5	92.0	86.4	0.612	
RBF	AF	5	97.9	45.5	**88.1**	0.882	0.404
		10	92.5	57.9	81.4	0.885	
		15	90.3	60.7	76.3	0.900	
		30	77.7	90.0	**88.1**	0.954	
MLP	AF, RRV	5	91.7	27.3	79.7	**0.903**	0.349
		10	92.5	89.5	**91.5**	**0.956**	
		15	83.9	75.0	**79.7**	**0.904**	
		30	88.8	88.0	**88.1**	0.973	

It can be observed that the AHI estimation provided by MLP outperformed MLR and RBF models, as well as the Event-detection algorithm, in terms of ICC. As in the case of multiclass classification, overall accuracies and κ values were low: 32.0% and 0.073 for Event-detection, 40.7% and 0.202 for MLR, 57.6% and 0.404 for RBF, and 52.5% and 0.349 for MLP. Nonetheless, very high diagnostic ability was shown when evaluating the common AHI cutoffs.

Consistent with its highest overall accuracy, RBF also obtained the highest κ value. However, MLP achieved the highest AROC and Acc for 3 out of the 4 AHI cutoffs, including all AROCs > 0.900 and 91.5% Acc for AHI = 10 e/h. According to confusion matrices of RBF and MLP, this higher diagnostic ability is related to a lower number of no-SAHS and mild-SAHS with overestimated severity degree.

Both MLP and RBF outperformed MLR and the Event-detection algorithm. The low diagnostic ability of MLR may be due to the underlying linearity assumption. The classic Event-detection method showed high diagnostic ability in terms of AROC. Indeed, it slightly improved AROC of MLP for the AHI cutoff = 30 e/h. However, accuracies were low for all AHI cutoffs.

Chapter 6

Discussion

In this study, the simplification of the SAHS diagnostic test has been assessed. The automated analysis of single-channel AF signals has been carried out for this purpose. Methodologies based on feature extraction, feature selection, and pattern recognition has been applied to AF in order to properly characterize SAHS, automatically detect it, and establish its severity. In this regard, binary and multiclass classification, as well as estimation of AHI, have been approached. This methodology has been compared with the classic event-by-event approach, both in the state of the art and using an event detection algorithm in our databases. Moreover, our approach has been applied to the widely studied HRV signal for comparison purposes as well. Next, the main results obtained during the study are discussed. The three first sections are focused on discussing results obtained in the case of adults. Then, an specific section is devoted to pediatric database. Finally, the main limitations of the study are presented.

6.1. Spectral bands of interest of the signals under study

Different spectral bands of interest have been used depending on the signal under study (thermistor AF and RRV, nasal prong AF, or HRV). In the case of thermistor AF and RRV (148 subject database), they were statistically determined using the whole set (148 subjects) [65] as well as only a training set (100 subjects) [64]. In either case results were consistent for both AF ($\approx$ 0.020-0.060 Hz.) and RRV ($\approx$ 0.09-0.13 Hz.).

The AF band of interest was also consistent with the reported typical (more common) apneic event duration, 20 to 40 seconds [39]. This duration mainly affects the 0.025-0.050 Hz. band in the frequency domain. Notice that the band of interest also falls within 0 to 0.1 Hz., which is the frequency band which meets the minimum apneic event duration criterion established by the AASM (10 seconds). These results lead us to directly use 0.025-0.050 Hz as band of interest for nasal-pressure AF too. This choice was supported by the reported results [62], where changes in the values of several features extracted from this band were observed while SAHS severity changed as well.

Physiological interpretation of the RRV band of interest was difficult since it is not a commonly analyzed signal and, to the best of our knowledge, it had not been previously used in SAHS studies. Moreover, interpolation was needed before conducting the spectral analysis, which is equivalent to the addition of estimated artificial data while complicates its interpretation. A range between 0.09 and 0.13 Hz. corresponds to events being repeated every 7.5 to 11 seconds. Certainly, this frequency range mostly falls above the minimal apneic event duration. However, other common and recurrent respiratory SAHS-related events, such as loud snoring, gasping, and choking, may affect time between breaths in higher frequencies than apneas and hypopneas.

As previously stated, a different case concerns HRV (or RR time series). It is a widely studied signal investigated in a wide range of physiological conditions, including SAHS. Indeed, this is a major reason for using it to compare with AF. Hence, well-known spectral bands of interest had already been reported in the literature, mainly related to the autonomic nervous system behavior. In contrast to spectral entropy, spectral power features extracted from these classic bands (VLF, LF, and HF) did not show significant differences between SAHS-negative and SAHS-positive subjects. However, a clear increase in the PSD of SAHS-positive subjects was found in the range 0.015-0.060 Hz [63],

covering part of the VLF and LF bands. Moreover, a recent study has reported an increased cardio-respiratory coordination during the apneic events [106]. These findings suggest a HRV band of interest related to the AF's one, as well as indicates that HRV classic bands are suboptimal for SAHS discrimination. Thus, further investigation is needed to find a HRV spectral band specific for SAHS.

6.2. Usefulness and complementarity of frequency and time domain analyses (linear and non-linear approaches)

Spectral and time domain features were used to characterize SAHS in each analyzed signal. In the case of thermistor AF, significant differences were mainly found in the features from the spectral band of interest. None of the time domain features showed differences between SAHS-negative and SAHS-positive subjects. Additionally, the SLR-FSBE algorithm only selected 4 spectral features as optimum (3 from the band of interest). By contrast, the FCBF method selected 7 features: 5 spectral (3 from the band of interest) and 2 non-linear. Since the underlying mechanisms of both selection methods are different, no conclusions can be drawn from the differences in the features selected. However, it is worth noting that two features were selected by the two methods and both of them were from the spectral band of interest (WD_B, M_{f1B}). All these data suggest that, in thermistor AF, the main information about SAHS is comprised within the spectral band of interest. A different behavior was found in the case of thermistor RRV. Frequency and time domain features, (from linear and non-linear approaches), showed statistically significant differences. In the case of spectral analysis, there were features from the spectral band of interest as well as from the whole spectrum which reached p-values < 0.01. The automatic feature selection stage supported these results. The SLR-FSBE algorithm selected 2 features as optimum (M_{f3}, CTM), whereas FCBF selected 5 (CTM, M_{f1B}, M_{t3}, M_{f3}, M_{f1}), which were spectral (outside of and within the spectral band) and temporal (statistics and non-linear). This indicates that, in RRV, the information about SAHS is contained in the entire signal as well as highlights the complementarity of the frequency and time-domain analyses. However, it has to be mentioned that, in spite of a higher number of features showing statistical differences, less features were selected by the SLR-FSBE (2) and FCBF (5) algorithms comparing with thermistor AF. This suggests a higher degree of redundancy in the information extracted from RRV. Finally, features from thermistor AF and RRV were automatically selected by both SLR-FSBE (2 out of 3 from AF) and FCBF (6 out of 10 from RRV) when including all of them in the selection process. This indicates that thermistor AF and RRV are able to provide complementary information about SAHS.

Only spectral features from the band of interest were extracted in the case of nasal-pressure AF. Non-linear features in time domain were also obtained. In the case of the binary classification task, CTM along with 7 out of the 9 extracted spectral features showed statistically significant differences. However, in the multiclass problem, only CTM and 4 out of the 9 spectral features reached p-value < 0.01 after Bonferroni's correction for multiple comparisons. As expected, these results suggest higher difficulty when characterizing SAHS severity instead of only the presence or absence of SAHS. The FCBF automatically selected the five features which showed the statistical differences

(M_{f1B},MA_B,mA_B,M_{f2B}, and CTM), as well as another five which did not show these differences ($SpecEn$,MF_B,M_{f4B}, LZC, and $SampEn$), suggesting their usefulness by providing complementary information.

Although the features extracted from nasal-pressure AF were not exactly the same as the extracted from thermistor AF, it is worth mentioning that M_{f1B}, CTM, and the non-linear entropy measure ($ApEn$ for thermistor and $SampEn$ for nasal prong) were selected by the FCBF in both cases. Additionally, SLR-FSBE also selected M_{f1B} from thermistor AF. M_{f1B} represents the mean PSD value of the band of interest, which is closely related to the power (or area) of that band, differing only on a constant factor. The spectral power at each frequency, in turn, is associated with the occurrence of time events in such a band. Since our proposed AF spectral band of interest is consistent with the occurrence of apneic events, M_{f1B} is most probably associated with them too. This reasoning is supported by data in Tables 5.1, 5.3, and 5.7, where it can be observed that statistically significant higher values of M_{f1B} are found in SAHS-positive subjects (both in adults and children), as well as in Table 5.10, where it is found that M_{f1B} increases as SAHS is more severe, i. e., as more number of apneic events occur. An illustration of this behavior can be also observed in figures 5.1 to 5.3. These data suggest that M_{f1B} is one major feature to characterize SAHS and its severity.

As mentioned in the previous section, power-based spectral features from typical HRV bands of interest did not evidenced differences between SAHS-negative and SAHS-positive subjects. By contrast, $SpecEn$ from VLF and LF bands did it both in women and men. This suggests these features as transverse when characterizing SAHS, as well as supports the idea pointed in the previous section regarding the need for searching new SAHS-specific HRV bands of interest, covering part of VLF and LF. MsE analysis, by contrast, showed different behaviors in men and women, suggesting it as being able to catch gender specificities. Thus, 15 out of the 25 $SampEn$ scales showed statistically significant differences in the case of women, whereas only one did it for men. Additionally, only low scales were included in the optimum set of features obtained for women by SLR-FSBE, whereas high scales were predominant (10 out of 12) in the case of men. In both cases, spectral and non-linear features were automatically selected, suggesting again the complementarity of these approaches.

6.3. Diagnostic ability: signals performance, classic approach, and state of the art

Information from signals was used as input to different pattern recognition methodologies focused on binary classification, multiclass classification, and regression. Thus, AB-CART (nasal-pressure AF) achieved the highest diagnostic ability in the binary classification task. It reached higher Acc (86.5%) and AROC (0.950) than LR trained with HRV features (72.3% Acc and 0.885 AROC), suggesting more generalization ability when gender is not taken into account. However, LR trained with HRV features from women reached higher κ and similar AROC (0.672 vs. 0.703 and 0.951 AROC). In the multiclass task, AB-LDA (nasal prong AF) achieved the highest diagnostic performance (86.5%, 81.0%, 82.5% Acc for an AHI cutoff = 5, 15, 30 e/h, respectively, and 0.432 κ). Finally, an MLP artificial neural network, feed with thermistor AF and RRV features, obtained the most accurate estimation of AHI in terms of

agreement (ICC=0.849) and diagnostic ability (91.5% Acc, 0.956 AROC, and 0.809 κ for an AHI cutoff = 10 e/h).

Thermistor, nasal-prong pressure, and HRV

As stated above, logistic regression acts as an standard in binary classification tasks. This is the reason why it was applied to features from all the signals under study. Consequently, it can be used to compare the diagnostic potential of each of them. Thus, LR was assessed for thermistor AF and RRV, as well as nasal prong AF. LR from nasal prong AF features outperformed LR from thermistor AF features (82.5% Acc, 0.917 AROC, and 0.593 κ vs. 79.7% Acc, 0.889 AROC, and 0.543 κ). The former, however, was built from 10 features whereas the latter only used 4. Performance of LR from RRV features was clearly lower. Nonetheless, diagnostic ability was similar when comparing LR from nasal prong AF with LR from thermistor AF and RRV features joined (82.4% Acc, 0.903 AROC, and 0.595 κ, 3 features). These results suggest that similar diagnostic ability can be reached when using single-channel AF information, regardless it is obtained through a thermistor or a nasal prong pressure sensor.

All LR models trained with AF information outperformed the LR model trained with HRV features from both men and women (72.3% Acc., 0.885 AROC, and 0.397 κ, 15 features). This result indicates higher diagnostic ability of AF signal than HRV signal when using data not separated by gender. LR only trained with features from men reached moderate diagnostic ability (77.6% Acc, 0.895 AROC, and 0.450 κ, 12 features). However, LR only trained with features from women reached higher diagnostic performance (85.2% Acc, 0.951 AROC, and 0.703 κ, 5 features). This suggests HRV as a useful signal to automatically screen SAHS in female subjects. In this regard, more research is needed in order to ensure causes motivating the different behavior found in men and women.

Direct comparison with the classic approach

As previously mentioned, AB-CART (nasal-pressure AF) achieved the highest diagnostic ability in the binary classification task (86.5% Acc, 0.993 AROC, and 0.672 κ). These results widely outperformed the classic approach, i. e., the Event-detection algorithm (69.0% Acc, 0.635 AROC, and 0.286 κ), which was applied to the same nasal-pressure AF database. AB-LDA (nasal-pressure AF), which achieved the highest diagnostic performance in the multiclass task (86.5%, 81.0%, 82.5% Acc in AHI = 5, 15, 30 e/h, and 0.432 κ), also overcame the Event-detection algorithm (81.0%, 68.3%, 63.5% Acc in AHI = 5, 15, 30 e/h, and 0.152 κ). This was also applied to the thermistor AF database to estimate AHI, reaching good agreement with actual AHI (ICC = 0.840). However, it showed moderated diagnostic ability (78.0% Acc, 0.805 AROC, and 0.474 κ for an AHI cutoff = 10 e/h), and MLP widely outperformed it (91.5% Acc, 0.956 AROC, and 0.809 κ for an AHI cutoff = 10 e/h). These direct comparisons suggest that our proposal, based on a comprehensive analysis of the signals, can improve the event-by-event approach usually followed to automatically detect SAHS.

Comparison with the state of the art

Table 6.1 shows results reported in a wide range of relevant state-of-the-art studies. These works focused on simplifying SAHS diagnosis by the use of single-channel AF (thermistor or nasal-pressure), HRV, or SpO_2. Table 6.2 displays the main results achieved during this study in the adults database.

Table 6.1: Summary of the diagnostic ability reported in the state-of-the-art main studies. Se: sensitivity (%); Sp: specificity (%); Acc (%): accuracy; AROC: area under ROC curve; Th: thermistor; NP: nasal prong. PSG: polysomnography. *: computed from reported data; -: not enough data to estimate; H-O: hold-out validation (training and test); loo: leave-one-out cross-validation; k-fold: k-fold cross-validation. SVM: support vector machine; MLP: multi-layer perceptron; LDA: linear discriminant analysis; QDA: quadratic discriminant analysis; KNN: K-nearest neighbors. E-D: event-detection

Study	Meth.	Sign.	n	AHI cutoff	Valid.	Se	Sp	Acc	AROC
Shochat et al [116]	E-D	AF(Th)	288	10	PSG	86.0	57.0	-	-
Gergely et al [54]	E-D	AF(Th)	83	15	PSG	71.9	73.1	72.3*	-
Nakano et al [89]	E-D	AF(Th)	216	5	H-O	88.0	80.0	-	0.950
				10		92.0	90.0	-	0.960
				15		86.0	90.0	-	0.950
Nakano et al [89]	E-D	AF(NP)	217	5	H-O	97.0	77.0	-	0.950
				10		97.0	76.0	-	**0.970**
				15		97.0	73.0	-	**0.980**
De Almeida et al [35]	E-D	AF(NP)	30	5	PSG	86.4	75.0	83.3*	0.886
				10		85.7	87.5	86.7*	0.915
				15		83.3	83.3	83.3*	0.898
Erman et al [41]	E-D	AF(NP)	59	5	PSG	85.4	50.0	74.6*	0.863
				10		82.1	83.9	83.1*	0.862
				15		90.9	94.6	**93.2***	0.977
Chen et al [28]	E-D	AF(NP)	50	5	PSG	97.7	66.7	94.0*	**0.951**
				15		87.5	88.9	88.0*	0.944
				30		88.2	93.9	**90.0***	0.955
Rofail et al [111]	E-D	AF(NP)	200	5	PSG	94.0	62.0	87.0*	0.840
				30		90.0	89.0	89.5*	**0.960**
BaHammam et al [14]	E-D	AF(NP)	95	5	PSG	79.0	68.0	77.9*	0.854
				10		70.0	89.0	75.8*	0.856
				15		65.0	94.0	75.8*	0.805
				30		63.0	98.0	83.2*	0.878
Roche et al [107]	*Tree*	HRV	147	10	k-fold	64.2*	75.6*	69.3*	-
Al-Angari et al [5]	*SVM*	HRV	100	5	-	79.6	78.4	79.0	-
Ravelo-García et al [103]	*LR*	HRV	97	10	k-fold	88.7	82.9	86.6*	0.941
Marcos et al [85]	*MLP*	SpO_2	187	10	H-O	89.8	79.4	85.5	0.900
Marcos et al [84]	*LDA*	SpO_2	187	10	H-O	86.6	80.4	84.1	0.925
	QDA					91.1	78.3	85.8	0.913
	KNN					88.1	84.8	86.7	0.822
	LR					85.1	87.0	85.8	0.930
Álvarez et al [9]	*LR*	SpO_2	148	10	loo	92.0	85.4	**89.7**	0.967
Marcos et al [83]	*MLP*	SpO_2	240	5	H-O	91.8	58.8	84.0	-
				10		89.6	81.3	86.8	-
				15		94.9	90.9	93.1	-
Álvarez et al [10]	*SVM*	SpO_2	320	10	H-O	95.2	80.0	84.5	-
Al-Angari et al [5]	*SVM*	SpO_2	100	5	-	91.8	98.0	**95.0**	-

Table 6.2: Summary of the methods which showed the highest diagnostic ability in the adults database, for each signal and pattern recognition approach (binary classification, multiclass classification, and regression). Se: sensitivity (%); Sp: specificity (%); Acc (%): accuracy; AROC: area under ROC curve; Th: thermistor; NP: nasal prong. PSG: polysomnography. w: women.*: computed from reported data; -: not enough data to estimate; H-O: hold-out validation (training and test); loo: leave-one-out cross-validation; k-fold: k-fold cross-validation.

Method	**Signal**	$\boldsymbol{n}$	**AHI cutoff**	**Valid.**	**Se**	**Sp**	**Acc**	**AROC**
$AB-CART$ (binary)[62]	AF(NP)	317	10	H-O	89.0	80.0	86.5	0.935
LR^w [63]	HRV	54	10	loo	80.8	89.3	85.2	0.951
$AB-LDA$ (multi)[62]	AF(NP)	317	5	H-O	87.1	80.0	86.5	-
			15		85.9	72.9	81.0	-
			30		74.2	90.6	82.5	-
MLP [64]	AF, RRV(Th)	148	5	H-O	91.7	27.3	79.7	0.903
			10		92.5	89.5	**91.5**	0.956
			15		83.9	75.0	79.7	0.904
			30		88.9	88.0	88.1	**0.973**

Studies involving **AF** were focused on applying an event-detection approach. Hence, the common methodology was to estimate AHI by scoring events, in order to evaluate its diagnostic performance according to one or several AHI cutoff thresholds. Most of them were intended to evaluate a portable device as a surrogate for complete PSG. Consequently, they were designed as validation studies in front of PSG, and no further validation was required. Only Nakano et al [89] detected the apneic events with the support of automatic spectral analysis. This approach required some adjustments which justified a hold-out validation procedure.

For AHI = 5 e/h, the Acc reported in these studies ranged 74.6% - 94.0%. Our multi AB-LDA (nasal-pressure AF) and MLP (thermistor AF and RRV) proposals were within this range (86.5% and 79.7%, respectively). Similarly, MLP reached 0.903 AROC, which is also in the range 0.840 - 0.951 reported in the literature. The highest Acc and AROC was shown by Chen et al (2009) [28] (94.0% and 0.951). However, their database was small ($\boldsymbol{n}$ = 50) and an imbalanced Se/Sp pair was reported (97.7% and 66.7%, respectively). Our multi AB-LDA showed higher Acc and a more balanced Se/Sp pair than the studies of De Almeida et al (2006) [35], Erman et al (2007) [41], and BaHammam et al (2011) [14]. A slightly higher Acc (87.0%) was reported by Rofail et al (2010) [111] in a large database ($\boldsymbol{n}$ = 200). However, Se and Sp were also imbalanced (94.0% and 62.0%). Finally, Nakano et al (2007) [89] evaluated both thermistor and nasal-prong AF. The latter showed imbalanced Se and Sp (97.0% and 77.0%) as well, whereas the former showed results very similar to multi AB-LDA (88.0% Se and 80.0% Sp).

In the case of AHI = 10 e/h, the AF studies reported AROC in the range 0.856 - 0.970, with our MLP and binary AB-CART proposals reaching high values within this range (0.956 and 0.935, respectively). Both of them reached higher AROC than the studies of De Almeida et al (2006) [35], Erman et al

(2007) [41], and BaHammam et al (2011) [14]. State-of-the-art studies also showed accuracies ranging 75.8% to 86.7%. AB-CART Acc was 86.5%, which was only overcame by De Almeida et [35] in a small database ($\boldsymbol{n}$ = 30). MLP performance was even higher, reaching 91.5% and a balanced Se/Sp pair (92.5% / 89.5%). Nakano et al [89] also showed high and balanced Se and Sp values (92.0% / 90.0%), as well as 0.960 AROC. These results are very similar to those from our MLP proposal. However, authors did not report enough data to estimate accuracy.

The evaluation of our proposals in the AHI cutoff = 15 e/h showed the lowest results. However, MLP and multi AB-LDA Acc and AROC were also included within the ranges reported in the literature for AF studies: 72.3% - 93.2% Acc and 0.805 - 0.980 AROC. Additionally, 3 out of the 5 studies showing higher Acc or AROC used small databases ($\boldsymbol{n}$ = 30 [35], 59 [41], 50 [28]).

Finally, in the case of AHI = 30 e/h, studies involving AF reported Acc in the range 83.2% - 90.0% and AROC ranging 0.878 to 0.960. The highest Acc was reached by Chen et al (90.0%) [28], achieved in a small database ($\boldsymbol{n}$ = 50). Multi AB-LDA did not reach the minimum Acc of the range (82.5%). However, MLP achieved 88.1% with a balanced Se/Sp pair (88.9% / 88.0%). Additionally, it showed 0.973 AROC, which overcame the upper limit of the range.

Regarding studies focused on **HRV**, our methodology applied to AF outperformed the state-of-the-art studies for AHI = 5 e/h and 10 e/h cutoffs. Moreover, it also showed high diagnostic ability when applied to HRV recordings from women. In this case, results similar to the highest reported in the literature were reached. Thus, Ravelo-García et al [103] achieved 86.5% Acc and 0.941 AROC when applying features from symbolic dynamics analysis to a LR model. Our LR proposal for women reached slightly lower Acc (85.2%) and higher AROC (0.951) in a smaller database ($\boldsymbol{n}$ = 54). However, they included 4 clinical variables in the LR model. Finally, several works not included in Table 6.1 reported 100% Acc when classifying 30 subjects from the PhysioNet Apnea-ECG database (AHI cutoff = 10 e/h), which was used in the Computers in Cardiology Challenge 2000 [95]. However, comparison with studies using this database is difficult since borderline subjects were deliberately removed from the competition [96].

In the case of the $\boldsymbol{SpO_2}$ signal, previous studies of our own group showed that spectral and nonlinear analyses, as well as pattern recognition techniques, are able to outperform conventional oximetric indexes based on event-detection, such as 3% and 4% oxygen desaturation index (ODI) [7, 8, 9, 84, 85]. Thus, results displayed in Table 6.1 show several pattern recognition algorithms which reached outstanding diagnostic ability when applied to data from SpO_2. Two studies evaluated their proposals for AHI = 5 e/h. The highest Acc was reached by Al-Angari et al (2012) [5], 95%, by means of a support vector machine (SVM) applied to oximetric data. However, the authors did not report any model validation procedure, which in the case of pattern recognition is mandatory to not overrate performance due to overfitting [20]. Marcos et al (2012) [83] reported 84% Acc along with an imbalanced Se/Sp pair (91.8% / 58.8%) through an MLP model used to estimate AHI. Our multi AB-LDA proposal reached higher Acc (86.5%) as well as a more balanced Se/Sp pair (87.1% / 80.0%). In the case of an AHI cutoff = 10 e/h, our MLP proposal reached high AROC (0.956) as well as higher Acc (91.5%) than all the SpO_2 studies in Table 6.1, including two MLP models fed with linear and non-linear features

(one for classification and the other one for regression). However, our database was smaller. For an AHI cutoff = 15 e/h, Marcos et al [83] reported high Acc (93.1%), outperforming both our AB-LDA and MLP proposals. However, no further comparison is possible since AROC was not reported.

6.4. An at-home study: pediatric SAHS

Children are particularly uncomfortable when undergoing conventional PSG. They do not tolerate well the equipment involved, which interferes in their sleep routine [75]. Hence, pediatric patients are of special interest to develop at-home and simplified diagnostic tests. As previously explained, we recorded thermistor AF and SpO_2, at children's homes, by means of a 6-channel polygraph with ability to record thoracic movements, body sensor, snoring sounds, and heart rate as well. After analyzing spectral information from AF, we answered three questions:

- **How does SAHS modify the spectral information of airflow recordings from children?**

  We found that the spectral power of AF was significantly higher in SAHS-positive subjects at 2 novel frequency bands below (BW1) and above (BW2) the typical respiratory range in children reported in previous studies ($0.220 - 0.430$ Hz) [45, 59, 121]. As in the case of adults, the relationship of BW1 with apneas and hypopneas can be explained on the basis of the definition of these apneic events. In children, apneas and hypopneas require at least 2 missed breaths of length in order to be scored [18]. Missing 2 cycles means that the recurrence of these apneic events is every 2 normal breaths, at most. Therefore, their frequency has to be located below the half of the normal respiratory frequency range, modifying the spectrum of AF in such band. Since BW1 is located below the half of the normal respiratory band, it is consistent with the occurrence of apneas and hypopneas. On the other hand, differences in the high frequency band, BW2 ($0.784 - 0.890$ Hz.), may be explained as the typical respiratory overexertion after an apneic event, which increases the respiratory rate [11].

- **Are these changes useful to distinguish SAHS in children from at-home recordings?**

  Seven out of the 13 extracted features were significantly different in SAHS-positive than in SAHS-negative subjects, (6 out of 12 from AF as well as ODI3). Two LR models, the first only fed with AF spectral features (LR_{AF}) and the second one fed with AF spectral features and ODI3 ($LR_{AF+ODI3}$), outperformed all single extracted features. Particularly high was the diagnostic ability of $LR_{AF+ODI3}$ (85.9% Se, 87.4% Sp, 86.3% Acc, and 0.947 AROC), which widely improved the performance of an in-lab 6-channel respiratory polygraph (74.2% Se, 81.8% Sp, 77.4% Acc, and 0.852 AROC) [11]. Moreover, our approach required only 2 of the channels recorded at patients' home. Additionally, LR_{AF} also outperformed this 6-channel respiratory polygraph (79.2% Se, 79.4% Sp, 79.1% Acc, and 0.875 AROC).

- **Is the airflow spectral information complementary to the classic oxygen desaturation index in pediatric SAHS detection?**

Our results showed complementarity between features in two cases. First, complementarity was highlighted between features from the two novel AF bands, since the SLR-FSBE algorithm automatically selected features from both of them to build the LR_{AF} and the $LR_{AF+ODI3}$ models. Second, features from the two spectral bands BW1-BW2 and the ODI3 also showed complementarity, since the latter was also selected for the $LR_{AF+ODI3}$ model.

Comparison with the state of the art

Other recent studies also analyzed physiological signals to help in pediatric SAHS diagnosis. Table 6.3 summarizes their results as well as displays again the results from our LR_{AF+ODI} proposal. Shouldice et al. [117] used 50 HRV recordings, and reached 85.7% Se, 81.8% Sp, and 84.0% Acc in a test set (AHI $\geq$ 1), by applying a quadratic linear discriminant to 23 features. Gil et al. [56] investigated the diagnostic usefulness of the information contained in 21 PRV time series, reporting 75.0% Se, 85.7% Sp, and 80.0% Acc after a leave-one-out cross-validation procedure (AHI $\geq$ 5). Garde et al. [53] reported 83.6% Se, 88.4% Sp, 84.9% Acc, and 0.860 AROC in a 146 subject database by combining 8 features from SpO_2 and pulse rate variability (PRV) in a linear discriminant (AHI $\geq$ 5). The relationship of high frequency inspiratory sounds (HFIS) to OSAS in children has been evaluated as well. Rembold and Suratt [104] reported data to estimate that 10 HFIS events per hour can be useful to discriminate SAHS in children for AHI $\geq$ 3 (61.5% Se, 100.0% Sp., and 80.8% Acc). Questionnaires and common symptoms have been also involved in screening tools for SAHS and sleep-disordered breathing. Spruyt and Gozal [119] proposed a severity scale based on the answers of 1133 children from general population to 6 sleep-related questions. They used a predictive score which reached 59.0% Se, 82.9% Sp, 0.790 AROC, 35.4% PPV, and 92.7% NPV (AHI $\geq$ 3). Kadmon et al. [73] validated this 6-item questionnaire in a sample of 85 children referred to a pediatric sleep clinic, reaching 83.0% Se, 64.0% Sp, 0.650 AROC, 28.0%PPV, and 96% NPV (AHI $\geq$ 5). Finally, Chang et al. [25] combined symptoms (observable apnea, restless sleep, and mouth breathing) with ODI from 141 children to assess a new discriminative score, reaching 60.0% Se, 86.0% Sp, 71.6% Acc, 84.0% PPV, and 64.0% NPV (AHI $\geq$ 5). Our $LR_{AF+ODI3}$ outperformed the reported diagnostic ability in these studies, even though we used recordings obtained from an unsupervised environment. However, Shouldice et al. [117] used a more restrictive AHI cutoff to differentiate patients from control subjects and Gil et al. [56], as well as Rembold and Suratt [104], worked with one single channel.

6.5. Limitations of the study

We have shown the utility of our proposal. However, there exist some limitations which need to be addressed. The first one is related to the sample size. Thus, in spite of using several databases involving a great number of subjects, a larger sample would enhance the statistical power of our results. More subjects would be particularly beneficial in the case of some of the subgroups analyzed, such as women, children, and subjects showing AHI $\leq$ 5 e/h. However, since SAHS is more prevalent in adult men, and high-risk population is prioritized, these kind of subjects are much less common in sleep units.

Table 6.3: Summary of the diagnostic ability reported in the state-of-the-art main studies focused on pediatric SAHS. Se: sensitivity (%); Sp: specificity (%); Acc: accuracy; AROC: area under ROC curve; PSG: polysomnography; HRV: heart rate variability; PPG: photopletysmography; PRV: pulse rate variability. *: computed from reported data; -: not enough data to estimate; loo: leave-one-out cross-validation; k-fold: k-fold cross-validation.

Study	**Signal**	n	**AHI cutoff**	**Valid.**	**Se**	**Sp**	**Acc**	**AROC**
Shouldice et al. [117]	HRV	50	1	loo	85.7	81.8	84.0	0.830
Rembold and Suratt [104]	Sounds	26	3	-	61.5*	100*	80.8*	-
Gil et al. [56]	PRV	21	5	loo	75.0	85.7	80.0	-
Spruyt and Gozal [119]	-	1133	3	PSG	59.0	82.9	-	0.790
Kadmon et al. [73]	-	85	5	PSG	83.0	64.0	70.6*	0.650
Chang et al. [25]	SpO_2	141	5	PSG	60.0	86.0	71.6*	-
Garde et al. [53]	SpO_2+PRV	146	5	k-fold	83.6	88.4	84.9	0.860
LR_{AF+ODI}	AF+ SpO_2	50	3	loo	**85.9**	87.4	**86.3**	**0.947**

Consequently, their data is also less available. This lack of data did not let us conduct multiclass or regression studies in the cases of HRV recordings from women and AF recordings from children. Despite the sample size limitation, several actions were taken in order to minimize its effect. Thus, appropriate validation methodologies were chosen according to the sample size at each case. Additionally, for the multiclass task involving nasal-pressure AF, SMOTE was applied to decrease the imbalance in the number of subjects with AHI ≤ 5.

Other limitations relate to the comprehensive analyses of the signals required to develop our proposal. In contrast to the classic approach, we did not look for each of the apneic events present in a recording. Conversely, we characterized each recording by summarizing its SAHS-related information in one or several features. As a consequence, the direct relationship between this information and the events was lost, leading to more difficult clinical interpretation of the results. By contrast, our approach used more information to characterize SAHS than the classic approach, in which data other than the apneic events is not exploited. Additionally, our approach does not depend on apnea and hypopnea definitions. Finally, we also showed that several of the extracted features were consistent with SAHS pathophysiology, which highlighted their clinical meaning. On the other hand, spectral analysis was common in all the studies conducted. It requires stationarity of the signal to be properly used. This is only partially fulfilled in the case of physiological recordings. However, Welch's periodogram was used to conduct all the spectral analyses in order to minimize this issue.

Our approach, based on minimizing the complexity of the diagnostic process, presents another limitation. Since we tried to use as few channels as possible, no data about thoracic movements were available. Consequently, it was not possible for us to distinguish among obstructive, central, and mixed events. However, central and mixed events are much less frequent than obstructive ones, and the most effective treatments do not take into account the specific cause of each respiratory disturbance. Additionally, it is not possible to recognize the sleep stages or to know whether patients are actually asleep, since EEG is not either used. Hence, when implementing our methods to estimate AHI, recording time was used instead of sleep time. Moreover, the single

use of thermistor or nasal-pressure AF is another limitation, since the AASM recommends using thermistor to score apneas and a nasal prong to score hypopneas. However, our research has shown that the comprehensive analysis of single-channel AF can achieve high diagnostic performance regardless the sensor used to acquire the signal.

A final limitation concerns the place where the recording of the signals of the adult databases are carried out. As in the case of our children database, at-home studies for adults would complement our findings about the diagnostic ability of the signals under study.

Chapter 7

Conclusions

In this Doctoral Thesis, the automated and comprehensive analysis of single-channel AF has been proposed as a simplified alternative to PSG in the SAHS-diagnosis process. Feature extraction and selection stages, as well as pattern recognition, formed the methodological core developed during the study. Feature extraction was used to exhaustively analyze AF in order to obtain as much complementary information as possible to characterize SAHS. Linear and non-linear approaches, implemented as spectral and time-domain analyses, were used for this purpose. Then, an automated feature selection step was developed in order to discard non-relevant and redundant information among the previously extracted features. Two approaches were also used for feature selection implementation: SLR-FSBE, which is dependent on logistic regression; and FCBF, which is independent of subsequent analyses. Finally, pattern recognition was used to transform the information obtained after feature extraction and selection into a diagnosis for each subject under study. Three independent strategies were followed: binary classification (presence or absence of SAHS), multiclass classification (determination of one out of the four SAHS severity degrees), and regression (estimation of the AHI).

Our pattern recognition approaches derived into high diagnostic ability models. Thus, for binary classification, an AB-CART model obtained 89.0% Se, 80.0% Sp, 86.6%, Acc, 0.935 AROC, and 0.672 κ. Similarly, in a children database, a LR model obtained 85.9% Se, 87.4% Sp, 86.3% Acc, 0.947 AROC, and 0.720 κ. In the multiclass classification task, an AB-LDA model reached 86.5%, 81.0%, and 82.5% accuracies for the AHI cutoffs = 5 e/h, 15 e/h, and 30 e/h, respectively, as well as 0.432 κ. Moreover, when estimating AHI, a MLP model achieved 79.7%, 91.5%, 79.7%, 88.1%, accuracies as well as 0.903, 0.956, 0.904, and 0.973 AROC for the AHI cutoffs = 5 e/h, 10 e/h, 15 e/h, and 30 e/h, respectively. These results highlighted the performance of our proposal comparing with the state-of-the-art studies, mainly focused on an event-by-event scoring approach. Additionally, our proposal widely outperformed a classic event-detection algorithm applied to our databases.

7.1. Contributions

Next, the main original contributions provided by the compendium of publications of this Doctoral Thesis are listed:

- To the best of our knowledge, this is the first time that single-channel AF is studied to help in SAHS diagnosis by means of the comprehensive analytical approach proposed in this study [61, 62, 64, 65]. In contrast to the classic event-by-event approach, our proposal takes into account the whole information from each recording. Additionally, it does not depend on apnea and hypopnea definitions. This general contribution can be divided into the next more specific ones:

    - AF analyses by means of features from different contexts. As mentioned in Chapter 1, physiological signals often present both stationary and chaotic behaviors. Thus, we extracted statistical features in time and frequency domain, as well as spectral and non-linear features. The main purpose of such analyses was to achieve a proper characterization of SAHS in AF by obtaining as useful and complementary information as possible. No similar methodology was found in the literature to characterize SAHS in AF recordings.

  - Automated selection of optimum sets of AF features. We followed two different approaches to carry out the selection process. We implemented a wrapper method (forward-selection backward-elimination, SLR-FSBE) and a filter method (fast correlation-based filter, FCBF). Both of them were useful to discard non-relevant and redundant information previously extracted, as well as to highlight the complementarity of the different approaches followed in the feature extraction stage. To the best of our knowledge, this is the first time that the FCBF algorithm is used in SAHS context. The SLR-FSBE method is a widely used algorithm which has been already applied to help in SAHS diagnosis. However, no other studies were found applying this method to features extracted from AF.
  - Automated detection of SAHS by means of pattern recognition techniques focused on binary classification, multiclass classification, and regression. A wide range of pattern recognition techniques were assessed to help in the automated SAHS diagnosis. Artificial neural networks and ensemble learning approaches showed the highest diagnostic ability. Thus, AdaBoost and Multi-layer perceptron (MLP) novel models outperformed the more common methods, such as linear discriminant, logistic regression, or multiple linear regression. In addition, they reached high diagnostic ability comparing with the state of the art. To the best of our knowledge, AdaBoost and MLP had not been applied to AF information in order to detect SAHS and its severity.

- In contrast to suggestions from the AASM, and by the use of pattern recognition techniques, we showed that it is possible to reach high diagnostic ability from single-channel AF, regardless it was acquired with a thermal sensor or a nasal pressure sensor [61, 62, 64, 65].

- Respiratory rate variability (RRV), derived from AF, was analyzed in relation to SAHS for the first time during this study [64, 65]. We showed that RRV contains relevant information about SAHS which complements the information from AF to improve its diagnostic ability.

- We defined novel spectral bands of interest in AF recordings from adults and children, following an automatic statistical approach, whose limits were consistent with the pathophysiology of SAHS [61, 64, 65]. Additionally, features extracted from these bands showed ability to discriminate SAHS, as well as usefulness in its automatic detection.

- It was shown that, when using information from HRV, SAHS could be more easily modeled when data is separated by gender, especially in the case of women [63]. No other studies involving HRV and SAHS were found which considered gender differences.

- An automated analysis of at-home AF and SpO_2 recordings from children was conducted. Spectral information from AF showed higher pediatric SAHS diagnosis ability than classic ODI. Moreover, the combination of ODI and this spectral information showed higher diagnostic ability than results from the state of the art [61].

7.2. Main conclusions of the study

The analysis of the obtained results lead to the next main conclusions of this Doctoral Thesis:

1. Pattern recognition applied to single-channel AF is useful to improve the automated SAHS diagnostic process.

2. High diagnostic ability can be reached from the automated analysis of single-channel AF, regardless it is obtained from a thermistor or a nasal prong. In this study, higher diagnostic performance was reached by nasal-pressure AF in the classification approaches, whereas thermistor AF showed very high performance when estimating AHI. Logistic regression was applied to both of them, reaching similar diagnostic ability and indicating that the diagnostic potential is similar for the two signals.

3. Our proposal, based on comprehensive and automatic analyses of single-channel AF, outperforms the classic event-by-event approach when both of them are applied to our database. Additionally, our proposal showed high diagnostic ability comparing with state-of-the-art studies which also adopt this event-by-event methodology.

4. Ensemble learning-based *AdaBoost* outperforms single LDA, CART, and LR classifiers, both in the binary classification task (AB-CART) and the multiclass classification (AB-LDA). The neural network-based MLP reaches the highest diagnostic performance when estimating AHI, outperforming MLR and another neural network, RBF.

5. The FCBF method may be more helpful to obtain optimum subsets of relevant and non-redundant AF features than the SLR-FSBE algorithm. The optimum subsets of features from thermistor AF and RRV obtained by FCBF reached higher diagnostic ability after the pattern recognition stage than those obtained from SLR-FSBE.

6. Thermistor AF and RRV provide complementary information about SAHS. Both SLR-FSBE and FCBF automatically selected features from the two signals when these were simultaneously evaluated, highlighting the singularity of the information provided by them.

7. The limits of the AF spectral band of interest in adults (0.020-0.060 Hz.), which were statistically obtained, are consistent with the pathophysiology of SAHS. Information extracted from the band of interest is related to both the presence of SAHS and its severity. The RRV spectral band of interest (0.090-0.130 Hz.) also provides useful information about SAHS. However, clinical interpretation is more difficult due to the novelty of the signal as well as to the interpolation required to a proper spectral analysis, which adds artificial information.

8. In AF signal, the main source of information about SAHS is comprised within the spectral band of interest. By contrast, RRV present useful information about SAHS contained in the entire spectrum and time series of the signal. However, features extracted from RRV signal are more redundant.

9. M_{f1B} from AF, closely related to the spectral power of the band of interest, is one major feature to characterize SAHS and its severity due to its

relationship to the occurrence of apneic events. It shows higher values as the SAHS severity increases. Additionally, it was automatically selected by the SLR-FSBE and FCBF algorithms, in the case of both thermistor and nasal-pressure AF.

10. Linear and non-linear approaches, implemented as frequency and time domain analyses, provide complementary information about SAHS. Both SLR-FSBE and FCBF algorithms automatically selected linear and non-linear features from AF, RRV, and HRV, which shows that both analyses provide relevant and non-redundant information.

11. Single-channel AF showed higher general diagnostic ability than HRV when applying the same analytical approach, as well as comparing with the state-of-the-art studies. However, when the gender was taken into account, a logistic regression model built with HRV features from women showed high diagnostic ability, outperforming logistic regression models built with AF features.

12. SAHS may be more easily modeled from HRV features in the case of women than in the men. A logistic regression model built with HRV spectral and non-linear features from women widely outperformed a similar model formed with features from men as well as another one built with features from both genders.

13. HRV typical spectral bands of interest are suboptimal for SAHS discrimination. Results reported in this study suggest that a different band of interest (0.015-0.060 Hz), covering part of the classic VLF and LF bands, may be more helpful in SAHS detection.

14. The spectral information contained within at-home single-channel AF recordings is useful to detect pediatric SAHS. It outperforms the conventional ODI from SpO_2 and results reported in an at-home 6-channel polygraphy.

15. The combination of the AF spectral information and ODI is helpful to accurately diagnose pediatric SAHS at home. Our 2-channel approach, based on logistic regression, outperformed the results reported in the main state-of-the-art studies.

In summary, SAHS-related information was obtained from single-channel AF. It was useful to characterize SAHS as well as to built pattern recognition models with ability to reach high diagnostic performance. Our proposal overcame the classic event-by-event approach and showed high diagnostic ability comparing with the state of the art. These results suggest that the SAHS diagnostic test can be reliably simplified by the use of automated analysis of single-channel AF.

7.3. Future research lines

There exist several questions derived from this investigation which can be the object of more research in the future since they can complement our results as well as take care of interesting topics out of the scope of this Doctoral Thesis. Below, we list those which we consider the most interesting future research lines:

- Difficulty in clinical interpretation of some of our extracted features is one of the limitations of our study. Hence, it would be helpful to conduct a more basic research regarding relationship between these features, signs, and symptoms of SAHS.
- Another limitation of our study is the impossibility of recognizing sleep stages in order to know when a patient is actually asleep. In this regard, several studies point out slightly different respiratory patterns depending on the sleep stage. Consequently, detection of these stages by the only use of single-channel AF would be helpful to complement our findings.
- The assessment of the diagnostic ability of features extracted from thermistor and nasal-pressure AF simultaneously is another interesting future line of investigation which is under development.
- One natural way to continue our investigation is the assessment of our methodologies in a large AF recording database obtained at patient's home. This, indeed, has been already planned to begin in the next months.
- Regarding pediatric SAHS, the acquisition of more recordings would be helpful to being able to conduct multiclass and AHI estimation studies. This data acquisition is currently taking place.
- In the same way, the acquisition of more HRV recordings from women would be interesting to conduct multiclass and AHI estimation studies, as well as compare results with those obtained from men. Finding SAHS-related gender specificities will be also the object of future research lines in our group.
- Another interesting future research is the automatic estimation of the quality of the signal, previous to its processing. Since the final goal is to develop simple and not supervised at-home diagnostic tests, it would be helpful to ensure the quality of the signal without the need for a of visual inspection.
- Finally, although we have implemented a wide range of methodologies during the study, the use of different feature extraction and selection techniques, as well as pattern recognition algorithms, is also a natural way to continue this research.

Appendices

Appendix A: compendium of publications

Next, they are included the full texts of the 4 published papers and the accepted paper which form the compendium of publications. They can be found in the corresponding publisher websites:

1. **Gutiérrez-Tobal, G. C.**, Hornero, R., Álvarez, D., Marcos, J. V., & del Campo, F. (2012). Linear and nonlinear analysis of airflow recordings to help in sleep apnoea-hypopnoea syndrome diagnosis. *Physiological measurement*, 33(7), 1261. Impact Factor: 1.496.

   `http://iopscience.iop.org/0967-3334/33/7/1261`

2. **Gutiérrez-Tobal, G. C.**, Álvarez, D., Marcos, J. V., Del Campo, F., & Hornero, R. (2013). Pattern recognition in airflow recordings to assist in the sleep apnoea-hypopnoea syndrome diagnosis. *Medical & biological engineering & computing*, 51(12), 1367-1380. Impact Factor: 1.500.

   `http://link.springer.com/article/10.1007/s11517-013-1109-7`

3. **Gutiérrez-Tobal, G. C.**, Álvarez, D., Gomez-Pilar, J., del Campo, F., & Hornero, R. (2015). Assessment of Time and Frequency Domain Entropies to Detect Sleep Apnoea in Heart Rate Variability Recordings from Men and Women. *Entropy*, 17(1), 123-141. Impact Factor: 1.502.

   `http://www.mdpi.com/1099-4300/17/1/123/htm`

4. **Gutiérrez-Tobal, G. C.**, Alonso-Álvarez, M. L., Álvarez, D., del Campo, F., Terán-Santos, J., & Hornero, R. (2015). Diagnosis of pediatric obstructive sleep apnea: Preliminary findings using automatic analysis of airflow and oximetry recordings obtained at patients' home. *Biomedical Signal Processing and Control*, 18, 401-407. Impact Factor: 1.419.

   `http://www.sciencedirect.com/science/article/pii/S1746809415000348`

5. **Gutiérrez-Tobal, G. C.**, Álvarez, D., del Campo, F., & Hornero, R. (2015). Utility of AdaBoost to Detect Sleep Apnea-Hypopnea Syndrome from Single-Channel Airflow. *IEEE Transactions on Biomedical Engineering*, *In Press*. Accepted August 2015. Impact Factor: 2.347.

   `http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7185342&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D7185342`

IOP PUBLISHING PHYSIOLOGICAL MEASUREMENT
Physiol. Meas. **33** (2012) 1261–1275 doi:10.1088/0967-3334/33/7/1261

Linear and nonlinear analysis of airflow recordings to help in sleep apnoea–hypopnoea syndrome diagnosis

G C Gutiérrez-Tobal[1], R Hornero[1], D Álvarez[1], J V Marcos[1] and F del Campo[2]

[1] Biomedical Engineering Group, ETSI de Telecomunicación, University of Valladolid, Paseo Belén 15, 47011, Valladolid, Spain
[2] Hospital Universitario Río Hortega, Servicio de Neumología, c/Dulzaina 2, 47012, Valladolid, Spain

E-mail: gguttob@ribera.tel.uva.es, robhor@tel.uva.es, dalvgon@ribera.tel.uva.es, jvmarcos@gmail.com and fsas@telefonica.net

Received 16 February 2012, accepted for publication 6 June 2012
Published 27 June 2012
Online at stacks.iop.org/PM/33/1261

Abstract
This paper focuses on the analysis of single-channel airflow (AF) signal to help in sleep apnoea–hypopnoea syndrome (SAHS) diagnosis. The respiratory rate variability (RRV) series is derived from AF by measuring time between consecutive breathings. A set of statistical, spectral and nonlinear features are extracted from both signals. Then, the forward stepwise logistic regression (FSLR) procedure is used in order to perform feature selection and classification. Three logistic regression (LR) models are obtained by applying FSLR to features from AF, RRV and both signals simultaneously. The diagnostic performance of single features and LR models is assessed and compared in terms of sensitivity, specificity, accuracy and area under the receiver-operating characteristics curve (AROC). The highest accuracy (82.43%) and AROC (0.903) are reached by the LR model derived from the combination of AF and RRV features. This result suggests that AF and RRV provide useful information to detect SAHS.

Keywords: sleep apnoea–hypopnoea syndrome, airflow, respiratory rate variability, feature extraction, feature selection

1. Introduction

The sleep apnoea–hypopnoea syndrome (SAHS) is characterized by repetitive events of apnoea (complete cessation of breathing) and hypopnoea (significant breathing reduction) during sleep (Flemons *et al* 2003). SAHS has been associated with other diseases such as hypertension, atrial fibrillation, stroke, cardiac failure, aortic dissection and sudden cardiac death (López-Jiménez *et al* 2008). Furthermore, daytime sleepiness caused by SAHS is a risk factor for occupational accidents and motor-vehicle collisions (Lindberg *et al* 2001, Sassani *et al* 2004).

The prevalence of SAHS has been estimated at 1%–5% of adult men and 2% women in western countries. However, studies reported up to 5% of adult population remaining undiagnosed (Young *et al* 2002).

The gold standard for SAHS diagnosis is polysomnography (PSG) (Flemons *et al* 2003). PSG is an overnight test in which many physiological signals are monitored. The apnoea–hypopnoea index (AHI) from PSG is used to characterize its severity (Patil *et al* 2007). Despite its effectiveness, PSG is an expensive and complex test, since it needs the supervision of specialists and a visual inspection of signals to compute AHI. This results in longer waiting lists and increased delay time for a final diagnosis (Flemons *et al* 2004). Therefore, there is a demand of new helping methods of diagnosis capable of overcoming PSG drawbacks (Penzel *et al* 2002). Many studies have focused on analysing a reduced set of signals from overnight PSG. Typically, the diagnostic ability of electrocardiogram (Penzel *et al* 2002), electroencephalogram (Poyares *et al* 2002), airflow (AF) (Nakano *et al* 2007, Han *et al* 2008), and blood oxygen saturation (SpO_2) (Álvarez *et al* 2010) has been evaluated.

The AF waveform is directly affected by the occurrence of respiratory events (Flemons *et al* 2003). Apnoeas are reflected by near-zero values, whereas hypopnoeas cause amplitude reduction. In contrast, clear oscillations are observed for normal breathing periods. Therefore, an intensive analysis of the information from the single-channel AF signal is proposed to help in SAHS detection. In addition to the AF signal, the respiratory rate variability (RRV) series is also analysed. RRV is computed by measuring the time between consecutive breathings in AF, similar to the well-known heart rate variability series (Cysarz *et al* 2008). The normal pattern for RRV also reflects alterations in the presence of SAHS, since sleep apnoea modifies the respiratory oscillation (Cysarz *et al* 2008).

The main purpose of the current study is to evaluate the diagnostic usefulness of AF and RRV series in SAHS detection. In order to characterize SAHS, the extraction of statistical, spectral and nonlinear features from AF and RRV is proposed. Common parameters such as statistical moments have shown to be useful in SAHS detection (Roche *et al* 1999, de Chazal *et al* 2003). Furthermore, frequency analysis has been successfully applied to study different diseases (Casolo *et al* 1991, Penzel *et al* 2002, Poza *et al* 2007). Moreover, nonlinear methods have recently proved high capability to help in SAHS diagnosis (Álvarez *et al* 2006, Hornero *et al* 2007, Morillo *et al* 2009). After feature extraction, a feature selection stage is implemeted. It is carry out by means of the forward stepwise logistic regression (FSLR) methodology (Hosmer and Lemeshow 1999), which has been successfully used in prior studies of SAHS (Álvarez *et al* 2010). The logistic regression (LR) models obtained through the FSLR procedure combine the non-redundant information from the features extracted (Hosmer and Lemeshow 1999). Finally, the diagnostic performance of the single features and the LR models are assessed and compared in terms of sensitivity, specificity, accuracy and area under the receiver-operating characteristic curve (AROC).

2. Subjects and signals

2.1. Subjects under study

In this study, 148 subjects suspected of suffering from SAHS were involved (79% males and 21% females). The recordings were obtained in the sleep unit of Hospital Universitario Río Hortega in Valladolid, Spain. All subjects presented common symptoms such as daytime hypersomnolence, loud snoring, nocturnal choking and awakenings or referred apnoeic events. The subjects were free from any medication which could influence the respiratory centre. Neither patients suffering from hypothyroidism (two out of the total subjects) nor those

Table 1. Demographic and clinical data of the population under study. Data are presented as mean ± SD or *n* (%). SAHS-positive: subjects with sleep apnoea–hypopnoea syndrome; SAHS-negative: subjects without sleep apnoea–hypopnoea syndrome; BMI: body mass index; time: recording time; AHI: apnoea–hypopnoea index.

	All subjects	SAHS-positive	SAHS-negative
Subjects (n)	148	100 (67.6%)	48 (32.4%)
Age (years)	50.87 ± 11.68	51.89 ± 11.41	48.75 ± 12.07
Males (n)	117 (79.0%)	85 (85.0%)	32 (66.7%)
BMI (kg m^{-2})	29.1 ± 4.6	29.9 ± 4.7	27.6 ± 4.9
Time (h)	7.24 ± 0.38	7.23 ± 0.36	7.27 ± 0.43
AHI (events/h)	–	32.9 ± 24.3	4.0 ± 2.4

suffering from chronic obstructive pulmonary disease (COPD) (six out of the total subjects) were excluded. Physicians considered 100 subjects affected (positive) and 48 not affected (negative) by SAHS. The AHI threshold for a positive diagnosis was 10 events/h at least. Apnoea was defined as the cessation of AF for 10 s or more. Hypopnoea was defined as a minimum of 30% of amplitude reduction for at least 10 s accompanied by a 4% or more decrease in the saturation of haemoglobin. The Review Board on Human Studies accepted the protocol, and all subjects gave their informed consent to participate in the study. Demographic and clinical data of the participants are summarized in table 1.

2.2. AF and RRV signals

The AF recordings were obtained from overnight PSG (Alice 5, Respironics, Philips Healthcare, the Netherlands). The sensor used to register AF was a thermistor (Pro-Tech, Respironics, Philips Healthcare, the Netherlands) and the sampled rate was 10 Hz. Previous to the automatic analysis, a visual inspection of the signals was carried out to assess their quality. Four recordings were excluded due to prolonged malfunction of the thermistor. Thus, the remaining 148 AF recordings were entirely analysed.

A peak detection algorithm was implemented to locate inspiratory onsets in AF signal (Korten and Haddad 1989). Then, RRV was computed by measuring the time between consecutive locations (Cysarz *et al* 2008).

Figure 1(a) shows an example of the AF signal and figure 1(b) shows the corresponding RRV signal. The first 34 s of the AF signal corresponds to a normal breathing pattern. Consequently, the time between breathings remains around 4.2 s in the RRV signal. Then, a hypopnoea is shown in the AF signal which is reflected by a decrease in the RRV signal amplitude. Finally, since the AF normal breathing pattern is recovered, the time between breathings begins to increase.

3. Methods

The proposed methodology started with a spectral analysis of AF and RRV recordings to determine those frequency bands associated with SAHS. Afterwards, spectral, nonlinear and statistical features were extracted from AF and RRV. Then several LR models were obtained by means of the FSLR method. Finally, diagnostic performance of single features and LR models was assessed.

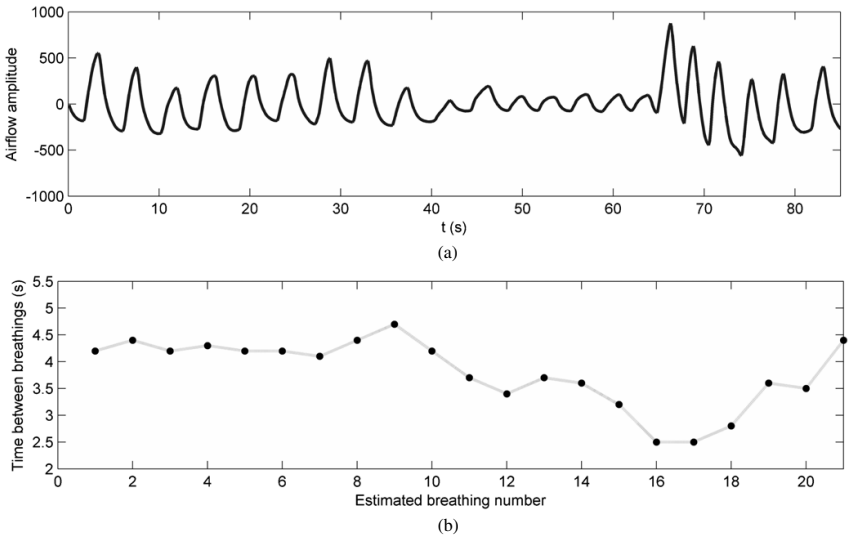

Figure 1. Normal breathing pattern followed by hypopnoea event in (a) AF signal and (b) corresponding RRV signal.

3.1. Definition of spectral bands of interest

The bands of interest were defined as the frequency regions of power spectral density (PSD) in which the highest statistical differences between SAHS-positive and SAHS-negative populations were found. PSD of recordings was estimated by means of a non-parametric Welch method (Welch 1967). This method divides the signals into M overlapping segments of length L. Then, a smooth time window $w[n]$ is applied, and the modified periodogram of each windowed segment $v_L[n]$ is computed by means of the discrete Fourier transform (DFT) $V[f]$ (Welch 1967):

$$\hat{P}[f] = \frac{|V[f]|^2}{f_s \mathrm{LU}}, \tag{1}$$

where f_s is the sample rate:

$$V[f] = \sum_{n=0}^{N-1} v_L[n]\, \mathrm{e}^{-j(2\pi k/N)n}, \tag{2}$$

and

$$U = \frac{1}{M} \sum_{n=0}^{M-1} |w(n)|^2. \tag{3}$$

Finally, the average of all DFTs is calculated to obtain the PSD function. A 2048-sample Hamming window, with 50% overlap and 4096-point DFTs, was used to compute the PSD of AF and RRV recordings. A cubic spline interpolation (resampling at 10 Hz) was applied to RRV series before the computation of the PSDs.

The representations of the joint PSDs of each SAHS-positive and SAHS-negative populations were obtained. The median of the PSD values at each frequency component

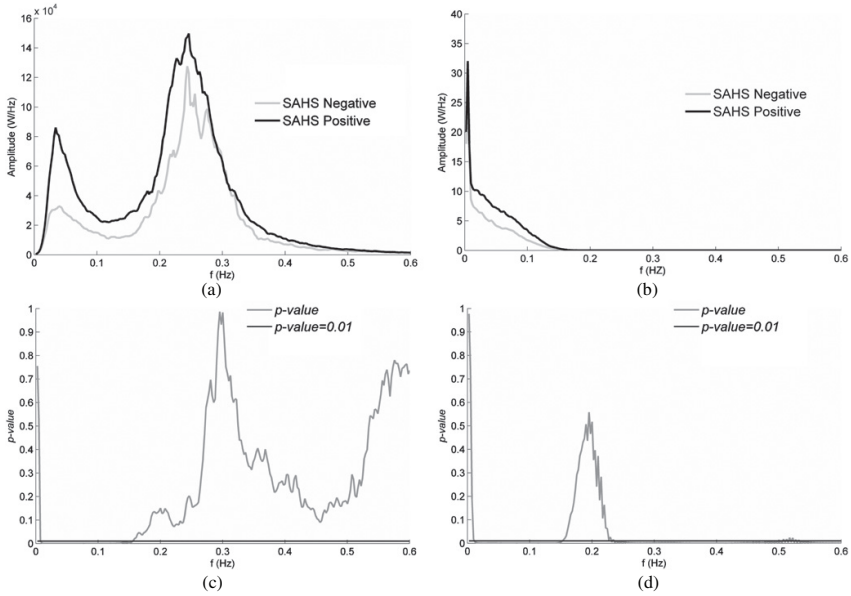

Figure 2. Spectral bands of interest of AF and RRV. AF (a) and RRV (b) median-based representations of PSDs from SAHS-positive (black line) and SAHS-negative (grey line) populations. AF (c) and RRV (d) whole spectrum *p*-value versus frequency representations (solid line).

was applied due to its robustness to outliers. Figures 2(a) and (b) show this median-based representation for AF and RRV, respectively. The *p*-value from the Kruskal–Wallis test was used to find statistically significant differences along frequencies (p-value <0.01). Figures 2(c) and (d) display the *p*-value versus frequency representations for AF and RRV, respectively. Figure 2(c) only shows significant differences in the very low frequency band (0.002–0.151 Hz). This agrees with figure 2(a), which displays the greatest qualitative differences in the same band of the spectrums of AF signals. The spectral band of interest was that with the highest statistically significant differences, i.e. 0.022–0.059 Hz. Moreover, the plot in figure 2(d) indicates significant differences in most of the frequency components of the RRV spectrum. However, the highest differences were also found at the very low band. This is also consistent with the corresponding median-based representation of the PSD (figure 2(b)). The spectral band of interest was 0.095–0.132 Hz.

3.2. Feature extraction

3.2.1. Statistical moments. The distributions of AF and RRV values are expected to differ from SAHS-positive and SAHS-negative populations. In order to typify the statistical behaviour of these distributions, the first-to-fourth statistical moments (Mt_1–Mt_4) were obtained from time series. The arithmetic mean (Mt_1), standard deviation (SD) (Mt_2), skewness (Mt_3) and kurtosis (Mt_4) quantify the central tendency, dispersion, asymmetry and peakedness of data, respectively.

3.2.2. Spectral features. The reccurrent nature of apnoeic events can be characterized by means of spectral analysis. Seven spectral features were extracted from the frequency bands of interest and the same features were obtained from the full PSDs.

Peak amplitude (*PA*) is the maximum of PSD in a given frequency interval. Band power (*BP*) represents the spectral power of a region. Both of them are conventional parameters and can be computed as follows:

$$PA = \max_{\mathrm{PSD}}\{\mathrm{PSD}(f)\}, \quad f(Hz) \in [f_i, f_N], \quad i = 1, 2, \ldots, N, \tag{4}$$

$$BP = \sum_{f_i=f_1}^{f_N} \mathrm{PSD}(f_i), \quad i = 1, 2, \ldots, N, \tag{5}$$

where N is the number of points in the band and f_i are the frequency components of the spectrum. Figures 2(a) and (b) indicate that higher *PA* and *BP* values are expected for the SAHS-positive population.

The Wootters distance (*WD*) is a disequilibrium measurement (Wootters 1981). This parameter requires the PSD to be normalized (PSD_n) in order to consider it as a probability density function (pdf). It is possible to measure the distance between the pdf and the uniform distribution (Wootters 1981):

$$WD = \arccos\left\{\sum_{f_i=f_1}^{f_2} \sqrt{\mathrm{PSD}_n(f)} \cdot \sqrt{1/N}\right\}, \tag{6}$$

with f_1 and f_2 being the limits of the frequency range where *WD* is applied and N is the number of the corresponding PSD_n points. If PSD_n is equal to a uniform distribution along frequencies (as in white noise), then *WD* will be equal to zero. Moreover, if the normalized spectrum is condensed into a narrow frequency band (as in a sum of sinusoids), *WD* reaches the highest values. According to figures 2(a) and (b), differences between *WD* values from both populations are expected.

Finally, first-to-fourth statistical moments (Mf_1–Mf_4) of the amplitude values of PSDs were also obtained. Differences between the distributions of PSD values are reflected by these parameters.

3.2.3. Nonlinear features. The high recurrence of apnoeas and hypopnoeas in SAHS-positive subjects modifies the corresponding AF and RRV waveforms. Since central tendency measure (*CTM*), Lempel–Ziv complexity (*LZC*) and approximate entropy (*ApEn*) are applied in time domain, it is expected that these parameters can reflect the differences in the variability, complexity and irregularity of time series from populations.

The *CTM* quantifies the degree of variability or chaos in a time series (Cohen *et al* 1996). *CTM* is based on the plots of the first-order differences representing $x[n+2]-x[n+1]$ versus $x[n+1]-x[n]$, where $x[n]$ are the time serie values (Abásolo *et al* 2006). It is calculated by counting the points that fall within a circle of radius ρ around the origin and dividing it by the total number of points (Cohen *et al* 1996):

$$CTM = \frac{1}{N-2}\sum_{n=1}^{n-2} \delta(n), \tag{7}$$

where

$$\delta(n) = \begin{cases} 1 & \text{if } \{(x[n+2]-x[n+1])^2 + (x[n+1]-x[n])^2\}^{1/2} < \rho \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

with N being the points of the time series. *CTM* achieves values between 0 and 1, reaching values closer to 1 when a given series is less variable (values more concentrated around centre) and closer to 0 when it has more variability. Radius ρ has to be selected experimentally, depending on the character of signals (Cohen *et al* 1996). A method based on p-value was used to select ρ (Hornero *et al* 1999). First, *CTM* of time series was computed by fixing several radii. Then a statistical significance test was applied to select the ρ which ensured the most significant differences between populations, i.e. the lowest p-value. In this study, two radii were used: $\rho_1 = 31$ for the AF signal and $\rho_2 = 6.61$ for the RRV signal.

The complexity of finite sequences can be estimated by means of *LZC* (Lempel and Ziv 1976). Larger values of the parameter correspond to higher complexity in these sequences (Zhang *et al* 2001). The first step in *LZC* estimation is to convert the time series into finite sequences of simbols, $s(i)$ (Zhang *et al* 2001, Abásolo *et al* 2006). Binary sequences have been commonly proposed. Due to its robustness to outliers, we assumed the median value as the threshold to assign a simbol to each value of time series. Once the sequence is obtained, it is scanned from left to right, and a complexity counter $c(n)$ is increased every time a new subsequence of consecutive characters is encountered (Zhang *et al* 2001). Finally, $c(n)$ is normalized to make the method independent of the length of sequences:

$$LZC = \frac{c(n)}{b(n)}, \tag{9}$$

where

$$b(n) = \frac{n}{\log_\alpha(n)}, \tag{10}$$

and $\alpha = 2$ since the sequence is binary.

ApEn is an irregularity measure in time series which was originally developed to be applied over short and noisy data sets (Pincus 1991). *ApEn* can assess both dominant and subordinates patterns in data for which other methods cannot make the feature recognition easy (Pincus 2001). *ApEn* has two user-specified parameters: a length m and a tolerance window r. Theoretically, the *ApEn* is defined as

$$ApEn(m, r) = \lim_{N\to\infty} [\phi^m(r) - \phi^{m+1}(r)], \tag{11}$$

where N is the total number of points of the original time series and $\phi^m(r)$ is the average of the logarithmic likelihood patterns of length m that are repeated along the original sequence. The tolerance parameter r is used to determine the similarity between patterns. Since N is finite, the *ApEn* is commonly applied as the statistic (Pincus 1991):

$$ApEn(m, r, N) = \phi^m(r) - \phi^{m+1}(r). \tag{12}$$

Larger values of *ApEn* correspond to more irregularity in the data (Pincus 2001). Despite their influence in the *ApEn* outcome, there are no guidelines to optimize the m and r values (Hornero *et al* 2005). Thus, $m = 1$, $m = 2$ and $r = 0.1$, 0.15, 0.2, 0.25 times the SD of the original data sequence have been proposed as input parameters. These values produce good statistical reproducibility of *ApEn* for time series of length $N \geqslant 60$ (Pincus 2001). In order to choose between these values, the p-value-based methodology previously described was used, and $m = 1$ and $r = 0.25$ SD were selected for both AF and RRV signals.

3.3. Feature selection

The features described in the previous subsections measure different properties of AF and RRV. The information contained in these parameters may be complementary. For simultaneous analysis of these features, several LR models were obtained. The method used to automatically

select the features was FSLR which was proposed by Hosmer and Lemeshow (1999). This procedure was applied to the features from AF, RRV and both signals (AF-RRV).

3.3.1. Forward stepwise logistic regression (FSLR). A regression-based method was used to describe the relationship between a response variable (outcome) and the explanatory variables. In this study, the response is a dichotomous variable codifying the diagnosis of a subject ('0' non-affected, and '1' affected by SAHS), and the explanatory variables are the features explained previously. For this outcome variable, the LR model has become the standard method of analysis:

$$\pi(x) = \frac{e^{\beta_0 + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}}}{1 + e^{\beta_0 + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}}}, \tag{13}$$

where $\pi(x)$ values range between 0 and 1, and can be interpreted as the probability of membership to the SAHS-positive population. β_0 is a constant for each model and $\boldsymbol{\beta}$ is a vector with coefficients for each component of $\mathbf{x}$. Both β_0 and $\boldsymbol{\beta}$ are estimated according to the maximum-likelihood criterion (Hosmer and Lemeshow 1999).

The more variables included in a LR model (higher dimensionality), the more dependent the model becomes on the observed data due to overfitting. Thus, feature selection was used in order to obtain models with higher capability of generalization. The FSLR procedure has been proposed for this purpose. It checks the relevance of features, including or excluding them according to a fixed decision rule. In this work, the decision rule chosen has been the p-value of the likelihood ratio (Hosmer and Lemeshow 1999). FSLR is characterized by a forward selection followed by the backward elimination of variables at each step.

3.4. Statistical analysis

The non-parametric Kruskal–Wallis test was used to assess the differences between the SAHS-positive and the SAHS-negative populations, with a p-value < 0.01 considered as significant. To ensure statistical validity of results, a leave-one-out cross-validation approach was applied. Sensitivity (percentage of SAHS-positive subjects correctly diagnosed), specificity (percentage of SAHS-negative subjects correctly diagnosed) and accuracy (proportion of total subjects under study correctly classified) were computed by averaging all results from the cross-validation process. Additionally, the AROC was computed to quantify the diagnostic performance of a given method (Zweig and Campbell 1993).

4. Results

4.1. Diagnostic performance of single features

A total of 21 features were extracted from each of the two series. Table 2 summarizes the measurements (mean $\pm$ SD) obtained for each feature in SAHS-positive and SAHS-negative populations. The p-value from the non-parametric Kruskal–Wallis significance test is also shown.

In the case of the AF signal, six out of seven spectral features obtained from the band of interest showed statistically significant differences between populations (p-value < 0.01). Only one out of seven (Mf_3) spectral features computed from the full PSD also showed p-value < 0.01. Neither the statistical moments in time domain nor the nonlinear features achieved statistically significant differences. In contrast, two statistical moments (Mt_2 and Mt_3) and two nonlinear features (*CTM* and *LZC*) obtained from RRV showed p-value < 0.01.

Table 2. Average values (mean ± SD) for the features extracted from AF and RRV signals in the SAHS-positive and the SAHS-negative populations. M_1–M_4: statistical moments of recordings in time domain; *CTM*: central tendency measure; *LZC*: Lempel—Ziv complexity; *ApEn*: approximate entropy; Mf_1–Mf_4: statistical moments obtained from full PSDs; *PA*: maximum of the full PSDs; *BP*: total spectral power; *WD*: Wootters distance; Mf_{1b}–Mf_{4b}: statistical moments obtained from the spectral band of interest (AF: 0.022–0.059 Hz, RRV: 0.095–0.132 Hz.); PA_b: maximum of *PSD* at the spectral band of interest; BP_b: spectral power at the band of interest; WD_b: Wootters distance at the spectral band of interest.

	AF signal			RRV signal		
Feature	SAHS-positive	SAHS-negative	p-value	SAHS-positive	SAHS-negative	p-value
Mt_1	0.06 ± 0.21	0.04 ± 0.11	$p > 0.01$	3.66 ± 0.51	3.64 ± 0.50	$p > 0.01$
Mt_2	190.38 ± 80.30	179.87 ± 91.42	$p > 0.01$	1.04 ± 0.33	0.85 ± 0.28	$\mathbf{p < 0.01}$
Mt_3	0.28 ± 0.24	0.28 ± 0.31	$p > 0.01$	0.81 ± 1.39	0.03 ± 1.11	$\mathbf{p < 0.01}$
Mt_4	8.25 ± 15.83	11.74 ± 21.40	$p > 0.01$	12.9 ± 12.7	9.9 ± 8.0	$p > 0.01$
CTM	0.635 ± 0.184	0.628 ± 0.185	$p > 0.01$	0.989 ± 0.017	0.998 ± 0.002	$\mathbf{p < 0.01}$
LZC	0.279 ± 0.029	0.283 ± 0.027	$p > 0.01$	0.992 ± 0.035	0.975 ± 0.037	$\mathbf{p < 0.01}$
ApEn	0.412 ± 0.073	0.435 ± 0.074	$p > 0.01$	1.46 ± 0.072	1.44 ± 0.075	$p > 0.01$
Mf_1	$8.4 \times 10^3 \pm 8.0 \times 10^3$	$7.9 \times 10^3 \pm 9.4 \times 10^3$	$p > 0.01$	0.22 ± 0.14	0.15 ± 0.10	$\mathbf{p < 0.01}$
Mf_2	$4.7 \times 10^4 \pm 5.1 \times 10^4$	$4.9 \times 10^4 \pm 6.410^4$	$p > 0.01$	1.69 ± 1.03	1.40 ± 1.04	$p > 0.01$
Mf_3	8.02 ± 1.77	8.94 ± 2.19	$\mathbf{p < 0.01}$	12.58 ± 4.28	17.51 ± 4.51	$\mathbf{p < 0.01}$
Mf_4	80.09 ± 35.18	96.98 ± 46.25	$p > 0.01$	232.60 ± 158.77	418.02 ± 184.65	$\mathbf{p < 0.01}$
PA	$59.9 \times 10^4 \pm 68.2 \times 10^4$	$68.5 \times 10^4 \pm 96.7 \times 10^4$	$p > 0.01$	37.55 ± 23.52	39.60 ± 32.36	$p > 0.01$
BP	$1.71 \times 10^7 \pm 1.64 \times 10^7$	$1.63 \times 10^7 \pm 1.93 \times 10^7$	$p > 0.01$	451.15 ± 285.91	307.2 ± 205.49	$\mathbf{p < 0.01}$
WD	0.798 ± 0.022	0.808 ± 0.018	$p > 0.01$	0.904 ± 0.009	0.908 ± 0.008	$\mathbf{p < 0.01}$
Mf_{1b}	$9.9 \times 10^4 \pm 12.9 \times 10^4$	$3.4 \times 10^4 \pm 2.3 \times 10^4$	$\mathbf{p < 0.01}$	2.96 ± 2.53	1.39 ± 0.91	$\mathbf{p < 0.01}$
Mf_{2b}	$39.6 \times 10^3 \pm 10.5 \times 10^3$	$7.1 \times 10^3 \pm 0.8 \times 10^3$	$\mathbf{p < 0.01}$	0.98 ± 0.86	0.56 ± 0.42	$\mathbf{p < 0.01}$
Mf_{3b}	0.042 ± 0.689	-0.451 ± 0.634	$\mathbf{p < 0.01}$	0.16 ± 0.39	0.23 ± 0.38	$p > 0.01$
Mf_{4b}	2.402 ± 0.905	2.675 ± 1.026	$p > 0.01$	2.0 ± 0.43	1.99 ± 0.40	$p > 0.01$
PA_b	$16.8 \times 10^4 \pm 30.7 \times 10^4$	$4.4 \times 10^4 \pm 3.710^4$	$\mathbf{p < 0.01}$	4.61 ± 3.72	2.34 ± 1.48	$\mathbf{p < 0.01}$
BP_b	$1.1 \times 10^5 \pm 20.4 \times 10^5$	$5.3 \times 10^5 \pm 3.5 \times 10^5$	$p < 0.01$	49.1 ± 41.6	23.3 ± 15.1	$\mathbf{p < 0.01}$
WD_b	0.109 ± 0.0608	0.063 ± 0.0372	$p < 0.01$	0.14 ± 0.07	0.16 ± 0.08	$p > 0.01$

Table 3. Results from the diagnostic assessment of single features extracted from AF and RRV recordings, derived from the leave-one-out cross-validation procedure. Sen.: sensitivity; Spe.: specificity; Acc.: accuracy; AROC: area under receiver-operating characteristics curve. AROCs > 0.800 are in bold.

	AF signal				RRV signal			
Feature	Sen.(%)	Spe.(%)	Acc.(%)	AROC	Sen.(%)	Spe.(%)	Acc.(%)	AROC
Mt_1	49.00	58.33	52.03	0.520	48.00	52.08	49.39	0.538
Mt_2	53.00	50.00	52.03	0.555	65.00	62.5	64.19	0.669
Mt_3	59.00	37.50	52.03	0.543	60.00	77.08	65.54	0.704
Mt_4	60.00	50.00	56.76	0.537	73.00	47.92	64.86	0.595
CTM	54.00	47.92	52.03	0.510	68.00	75.00	70.27	**0.800**
LZC	57.00	58.33	57.43	0.553	62.00	56.25	60.14	0.654
ApEn	61.00	56.25	59.49	0.585	52.00	50.00	51.35	0.577
Mf_1	58.00	50.00	55.41	0.554	65.00	62.50	64.19	0.676
Mf_2	50.00	47.92	49.32	0.502	61.00	52.08	58.11	0.618
Mf_3	68.00	56.25	64.19	0.634	77.00	68.75	74.32	**0.809**
Mf_4	63.00	56.25	60.81	0.612	79.00	70.83	76.35	**0.807**
PA	48.00	56.25	50.68	0.513	56.00	41.67	51.35	0.528
BP	58.00	50.00	55.41	0.561	65.00	62.50	64.19	0.676
WD	69.00	54.47	64.19	0.631	64.00	56.25	61.49	0.633
Mf_{1b}	71.00	83.33	75.00	**0.826**	61.00	79.17	66.89	0.745
Mf_{2b}	74.00	87.50	78.38	**0.851**	61.00	62.50	61.49	0.702
Mf_{3b}	58.00	68.75	61.49	0.676	50.00	52.08	50.68	0.559
Mf_{4b}	63.00	56.25	60.81	0.581	47.00	52.08	48.65	0.508
PA_b	74.00	83.33	77.03	**0.840**	59.00	66.67	61.49	0.745
BP_b	71.00	83.33	75.00	**0.838**	62.00	79.17	67.57	0.756
WD_b	65.00	70.83	66.89	0.797	61.00	47.92	56.76	0.567

Table 4. Results from the diagnostic assessment of the LR models, derived from the leave-one-out cross-validation procedure. Sen.: sensitivity; Spe.: specificity; Acc.: accuracy; AROC: area under receiver-operating characteristics curve. The number of features introduced as independent variables at each model are in parentheses. AROCs>0.800 are in bold.

Model	Selected features	Sen.(%)	Spe.(%)	Acc.(%)	AROC
AF (21)	WD_b, BP_b, *PA*, Mf_{2b}	84.00	70.83	79.73	**0.889**
RRV (21)	Mf_3, *CTM*	84.00	58.33	75.68	**0.850**
AF-RRV (42)	Mf_3^{RRV}, PA^{AF}, BP_b^{AF}	88.00	70.83	82.42	**0.903**

Furthermore, nine out of 14 RRV spectral features obtained from the full PSD and the band of interest presented statistically significant differences.

The results of the individual diagnostic assessment of the features are shown in table 3. Consistent with the statistical significance test analysis, those features with p-value < 0.01 improved the diagnostic performance of the others. For the AF signal, the highest accuracy (78.38%) and AROC (0.851) were reached by Mf_{2b}. In the case of RRV, the highest accuracy (76.35%) was obtained by Mf_4, whereas Mf_3 achieved the highest AROC (0.809).

4.2. Performance of the FSLR procedure

Table 4 shows the diagnostic results provided by the LR models. The features automatically selected are also specified. The order of appearance of the selected features in the table is the same as the order obtained from the FSLR method.

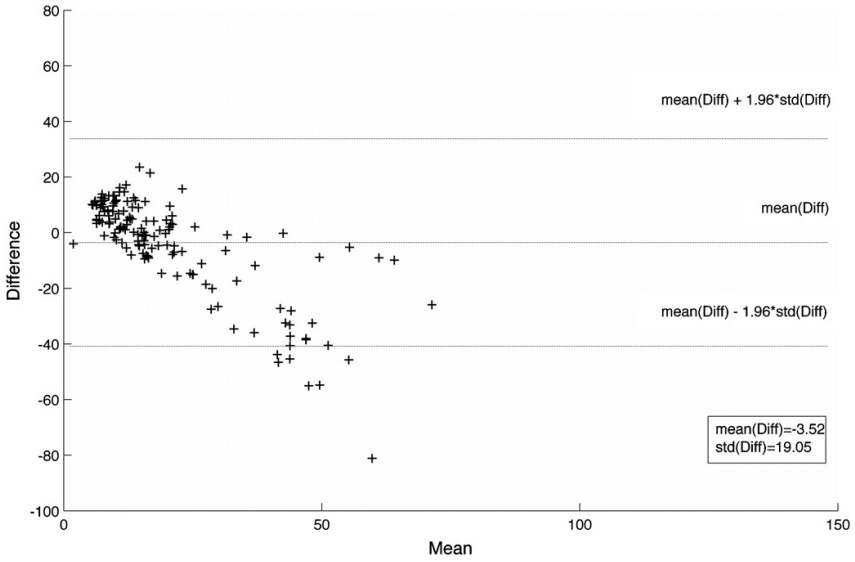

Figure 3. Bland–Altman plot comparing the AHI from $\text{MLR}_{\text{AF\_RRV}}$ with the AHI from PSG.

Table 5. Linear correlation analysis between features selected to the models and BMI/Age. ρ: Pearson's correlation coefficient.

Feature	ρ (BMI)	ρ (Age)
WD_b^{AF}	0.222	0.115
BP_b^{AF}	0.314	0.021
PA^{AF}	0.156	−0.071
Mf_{2b}^{AF}	0.269	−0.003
Mf_3^{RRV}	−0.054	−0.051
CTM^{RRV}	−0.225	−0.262

Four parameters (WD_b, BP_b, Mf_{2b} and PA) were automatically selected by the FSLR procedure when all 21 AF features were introduced as independent variables. The AF model reached 79.73% accuracy and 0.889 AROC. Mf_3 and CTM were automatically selected in the case of the RRV model, obtaining 75.68% accuracy and 0.850 AROC. Finally, the highest sensibility (88.00%), specificity (70.83%), accuracy (82.43%) and AROC (0.903) were achieved by the *AF-RRV* model. Three parameters were automatically selected: Mf_3 from RRV and PA and BP_b from AF. These features were used to obtain a multivariate linear regression model ($\text{MLR}_{\text{AF\_RRV}}$). The output of $\text{MLR}_{\text{AF\_RRV}}$ (estimated AHI) was compared to the AHI from PSG. The Bland–Altman plot shown in figure 3 was used for this purpose. It displays an overestimation tendency for lower AHI values whereas an underestimation tendency is shown when the AHI becomes higher. Finally, table 5 shows the assessment of linear correlation between all the features automatically selected and the BMI and age. None of the features obtained high values of Pearson's correlation coefficient with BMI or age.

5. Discussion and conclusions

The utility of AF signals in SAHS detection was assessed. The information from RRV series, which was derived from AF, was also evaluated. Statistical, spectral and nonlinear features were used to characterize the behaviour of AF and RRV recordings in SAHS-positive and SAHS-negative populations. AF and RRV frequency bands of interest within the very low frequency region were proposed for SAHS detection. Regarding the AF signal, since the normal breathing rate at rest is set close to 15 breaths per minute, i.e. 0.25 Hz. (Farré *et al* 1998), the spectral components at very low frequencies of AF correspond to an abnormal respiratory behaviour. Moreover, the selected spectral band of interest was located below 0.1 Hz. (0.022–0.059 Hz.). Since apnoea and hypopnoea events last 10 s or more (Flemons *et al* 2003), the band is consistent with pathophysiology. However, further analysis is required in order to assess the cause-motivating differences in frequencies higher than 0.1 Hz. The occurrence of larger number of short-time respiratory events in SAHS-positive subjects is proposed as a cause for this behaviour. In the case of RRV spectrum, significant differences were found in most of the spectral components, indicating major changes in time between breathings caused by SAHS. Most of the AF features that achieved significant differences between populations were extracted from the spectral band of interest. In contrast, p-value <0.01 was achieved by statistical, spectral and nonlinear features from RRV. This suggested that RRV signal contains useful information about SAHS in time and frequency domains. The diagnostic performance of the features reinforced the ideas exposed above: several spectral features from the AF band of interest reached higher values of AROC (0.825–0.851) than any other single feature. This confirmed the usefulness of the spectral information contained in AF signals to help in SAHS diagnosis.

None of the features selected for the *AF*, *RRV* and *AF-RRV* models presented high linear correlation with BMI or age of subjects under study. The *AF-RRV* model achieved the highest diagnostic performance (82.42% accuracy and 0.903 AROC). Mf_3^{RRV}, PA^{AF} and BP_b^{AF} were automatically selected, containing information from both *AF* and *RRV* signals. One out of the two hypothyroidism patients (false positive) and none of the COPD patients were misclassified. These results showed that the FSLR procedure improved the diagnostic performance of single features and suggested that information contained in AF and RRV signals could be complementary.

The AF signals obtained from thermistor have been recently analysed to help in SAHS diagnosis. Two hundred and eighty eight subjects participated in a multi-centre study to evaluate a screening device based on the detection of respiratory events (Shochat *et al* 2002). Thus, 86% sensitivity, 57% specificity and 0.81 AROC were achieved in the classification of subjects. The same methodology was applied to a different population, comparing the screening performance of the device to the performance of nocturnal pulsioximetry (Gergely *et al* 2009). The results of the study achieved 71.9% sensitivity and 73.1% specificity using AHI $= 15$ as a cut-off threshold. The *AF-RRV* model obtained in this work improved the diagnostic performance of both studies.

There exists an extensive literature focused on the analysis of AF from nasal pressure (NP) sensor. Most of them aimed to locate respiratory events in AF. Subsequently, a respiratory disturbance index (RDI) is computed in order to assess its diagnostic performance (De Almeida *et al* 2006, Erman *et al* 2007, Nakano *et al* 2007, Grover and Pittman 2008, Wong *et al* 2008, Tonelli *et al* 2009, Chen *et al* 2009, Rofail *et al* 2010). Populations involved in these studies ranged from 25 to 200 subjects (83.5 ± 64.6, mean $\pm$ SD). Sensitivity, specificity and AROC reached ranged 82%–97%, 62%–90% and 0.84–0.98, respectively. The best results achieved in this study are included in these intervals.

Some limitations have to be taken into account. The population under study could be larger, with a more balanced proportion of SAHS-positive and SAHS-negative subjects. Furthermore, all subjects were suspected of having SAHS before PSG test. A control group (subjects without any symptoms) should be analysed in order to assess the universal application of the methodology. It would provide additional information to complete this study. The use of a thermistor to acquire the AF signal, instead of a NP sensor, is al.so a limitation. Measurements from thermistor are only indirectly related to the AF, resulting in the underdetection of hypopnoeas (Farré *et al* 1998). The NP sensor has shown a better performance for detecting obstructive respiratory events (Bahammam 2004). However, it has a roughly quadratic relationship with the flow, causing AF changes to be exaggerated and, consequently, resulting in an overestimation of apnoea events (Bahammam 2004). The American Academy of Sleep Medicine recommends the use of both types of sensors due to these disadvantages (Iber *et al* 2007). The comparison of features extracted from the signals acquired with the two sensors and the joint analysis of the information extracted from them are future goals. Another limitation has to be considered. The variability of AHI from PSG in successive nights is well known (Carlile and Carlile 2008, Levendowski *et al* 2009). However, night-to-night variability is not often assessed due to economics and time limitations (Levendowski *et al* 2009). Repeated sleep studies in successive nights would be necessary to complete the assessment of this methodology. Finally, the use of an AHI threshold $= 10$ events/h to discriminate SAHS is also a limitation since subjects in the range 5–10 events/h could benefit from the continuous positive airway pressure (CPAP) treatment. CPAP is the most widely used treatment for severe SAHS (Lindberg *et al* 2006, Marshall *et al* 2006). Further work is needed to assess the accuracy of the methodology for screening those patients who would benefit from CPAP.

In summary, AF and RRV signals were analysed. A spectral band of interest was located in a region of the AF spectrum corresponding to anomalous respiration. The statistical significance test and the diagnostic performance assessment of the features confirmed the usefulness of the information contained in AF and RRV. Results from the FSLR procedure suggested that data extracted from them can complement each other. Moreover, the *AF-RRV* model improved the diagnostic performance of all the single features. The best results obtained from this study improved the results from those studies which involved thermistor and are comparable to those involving NP. Therefore, the proposed methodology could be useful to help in SAHS diagnosis.

Acknowledgments

This work has been partially supported by the project VA111A11-2 from Consejería de Educación de la Junta de Castilla y León, by the Proyectos Cero on Ageing from Fundación General CSIC, by Consejería de Educación de la Junta de Castilla y León (Orden EDU/1204/2010) and by the European Social Found.

References

Abásolo D, Hornero R, Gómez C, García M and López M 2006 Analysis of background activity in Alzehimer's disease patients with Lempel–Ziv complexity and central tendency measure *Med. Eng. Phys.* **28** 315–22

Álvarez D, Hornero R, Abásolo D, del Campo F and Zamarrón C 2006 Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection *Physiol. Meas.* **27** 399–412

Álvarez D, Hornero R, Marcos J V and del Campo F 2010 Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis *IEEE Trans. Biomed. Eng.* **57** 2816–24

BaHammam A 2004 Comparison of nasal prong pressure and thermistor measurements for detecting respiratory events during sleep *Respiration* **71** 385–90

Carlile J and Carlile N 2008 Repeat study of 149 patients suspected of having sleep apnea but with an AHI<5 *Sleep* **31** A153

Casolo G, Balli E, Fazi A, Gori C, Freni A and Gensini G 1991 Twenty-four-hour spectral analysis of heart rate variability in congestive heart failure secondary to coronary artery disease *Am. J. Cardiol.* **67** 1154–8

de Chazal P, Heneghan H, Sheridan E, Reilly R, Nolan P and O'Malley 2003 Automated processing of single-lead electrocardiogram for the detection of obstructive sleep apnoea *IEEE Trans. Biomed. Eng.* **50** 686–96

Chen H, Lowe A A, Bai Y, Hamilton P, Fleetham J A and Almeida F R 2009 Evaluation of a portable recording device (ApneaLink^{TM}) for case selection of obstructive sleep apnea *Sleep Breath* **13** 213–9

Cohen M E, Hudson D L and Deedwania P C 1996 Applying continuous chaotic modelling to cardiac signal analysis *IEEE Eng. Med. Biol. Mag.* **15** 97–102

Cysarz D, Zerm R, Bettermann H, Frühwirth M, Moser M and Kröz M 2008 Comparison of respiratory rates derived from heart rate variability, ECG amplitude, and nasal/oral airflow *Ann. Biomed. Eng.* **36** 2085–94

De Almeida F R, Ayas N T, Ueda H, Hamilton P, Ryan F C and Lowe A A 2006 Nasal pressure recordings to detect obstructive sleep apnea *Sleep Breath* **10** 62–69

Erman M K, Stewart D, Einhorn D, Gordon N and Casal E 2007 Validation of the ApneaLink^{TM} for the screening of sleep apnea: a novel and simple single-channel recording device *J. Clin. Sleep Med.* **3** 387–92

Farré R, Montserrat J M, Rotger M, Ballester E and Navajas D 1998 Accuracy of thermistors and thermocouples as flow-measuring devices for detecting hypopnoeas *Eur. Respir. J.* **11** 179–82

Flemons W W, Douglas N J, Kuna S T, Rodenstein D O and Wheatley J 2004 Access to diagnosis and treatment of patients with suspected sleep apnea *Am. J. Respir. Crit. Care Med.* **169** 668–72

Flemons W W, Littner M R, Rowley J A, Gay P, Anderson W M, Hudgel D W, McEvoy D and Loube D I 2003 Home diagnosis of sleep apnea: a systematic review of the literature *Chest* **124** 1543–79

Gergely V, Pallos H, Mashima K, Miyazaki S, Tanaka T, Okawa M and Yamada N 2009 Evaluation of the usefulness of the SleepStrip for screening obstructive sleep apnea–hypopnea syndrome in Japan *Sleep Biol. Rhythms* **7** 43–51

Grover S S and Pittman S D 2008 Automated detection of sleep disordered breathing using a nasal pressure monitoring device *Sleep Breath* **12** 339–45

Han J, Shin H B, Jeong D U and Park K S 2008 Detection of apnoeic events from single channel nasal airflow using 2nd derivative method *Comput. Methods Programs Biomed.* **98** 199–207

Hornero R, Aboy M, McNames J and Goldstein B 2005 Interpretation of approximate entropy. Case studies in the analysis of intracranial pressure during elevations in traumatic brain injury *IEEE Trans. Biomed. Eng.* **53** 1671–80

Hornero R, Alonso A, Jimeno N, Jimeno A and López M 1999 Nonlinear analysis of time series generated by schizophrenic patients *IEEE Eng. Med. Biol. Mag.* **3** 84–90

Hornero R, Álvarez D, Abásolo D, del Campo F and Zamarrón C 2007 Utility of approximate entropy from overnight pulse oximetry data in the diagnosis of the obstructive sleep apnea syndrome *IEEE Trans. Biomed. Eng.* **54** 107–13

Hosmer D W and Lemeshow S 1999 *Applied Logistic Regression* (New York: Wiley)

Iber C, Ancoli-Israel S, Chesson A and Quan S F 2007 *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications* (Westchester, IL: American Academy of Sleep Medicine)

Korten J B and Haddad G G 1989 Respiratory waveform pattern recognition using digital techniques *Comput. Biol. Med.* **19** 207–17

Lempel A and Ziv J 1976 On the complexity of finite sequences *IEEE Trans. Inf. Theory* **24** 530–6

Levendowski D, Steward D, Woodson B T, Olmstead R, Popovic D and Westbrook P 2009 The impact of obstructive sleep apnea variability measured in-lab versus in-home on sample size calculations *Int. Arch. Med.* **2** 1–8

Lindberg E, Berne C, Elmasry A, Hedner J and Janson C 2006 CPAP treatment of a population-based sample—What are the benefits and the treatment compliance? *Sleep Med.* **7** 553–60

Lindberg E, Carter N, Gislason T and Janson C 2001 Role of snoring and daytime sleepiness in occupational accidents *Am. J. Respir. Crit. Care Med.* **164** 2031–5

López-Jiménez F, Kuniyoshi F H S, Gami A and Somers V K 2008 Obstructive sleep apnea: implications for cardiac and vascular disease *Chest* **133** 793–804

Marshall N S, Barnes M, Travier N, Campbell A J, Pierce R J, McEvoy R D, Neill A M and Gander P H 2006 Continuous positive airway pressure reduces daytime sleepiness in mild to moderate sleep apnoea: a meta-analysis *Thorax* **61** 430–4

Morillo D S, Rojas J L, Crespo L F, León A and Gross N 2009 Poincaré analysis of an overnight arterial oxygen saturation signal applied to the diagnosis of sleep apnea hipopnea syndrome *Physiol. Meas.* **30** 405–20

Nakano H, Tanigawa T, Furukawa T and Nishina S 2007 Automatic detection of sleep-disordered breathing from single-channel airflow record *Eur. Respir. J.* **29** 728–36

Patil S P, Schneider H, Schwartz A R and Smith P L 2007 Adult obstructive sleep apnea: pathophysiology and diagnosis *Chest* **132** 325–37

Penzel T, McNames J, de Chazal P, Raymond B, Murray A and Moody G 2002 Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings *Med. Biol. Eng. Comput.* **40** 402–7

Pincus S M 1991 Approximate entropy as a measure of system complexity *Proc. Natl Acad. Sci.* **88** 2297–301

Pincus S M 2001 Assessing serial irregularity and its implications for health *Ann. New York Acad. Sci.* **954** 245–67

Poyares D, Guilleminault C, Rosa A, Ohayon M and Koester U 2002 Arousal, EEG spectral power and pulse transit time in UARS and mild OSAS subjects *Clin. Neurophysiol.* **113** 1598–606

Poza J, Hornero R, Abásolo D, Fernández A and García M 2007 Extraction of spectral based measures from MEG background oscillations in Alzheimer's disease *Med. Eng. Phys.* **29** 1073–83

Roche F, Gaspoz J M, Court-Fortune I, Minini P, Pichot V, Duverney D, Costes F, Lacour J R and Barthélémy J C 1999 Screening of obstructive sleep apnea syndrome by heart rate variability analysis *Circulation* **100** 1411–5

Rofail L M, Wong K K H, Unger G, Marks G B and Grunstein R R 2010 The role of single-channel nasal airflow pressure transducer in the diagnosis of OSA in the sleep laboratory *J. Clin. Sleep Med.* **6** 349–56

Sassani A, Findley L J, Kryger M, Goldlust E, George C and Davidson T M 2004 Reducing motor-vehicle collisions, cost, and fatalities by treating obstructive sleep apnea syndrome *Sleep* **27** 453–8

Shochat T, Hadas N, Kerkhofs M, Herchuelz A, Penzel T, Peter J H and Lavie P 2002 The SleepStrip^{TM}: an apnoea screener for the early detection of sleep apnoea syndrome *Eur. Respir. J.* **19** 121–6

Tonelli A C, Martinez D, Vasconcelos L F T, Cadaval S, Lenz M C, Costa S, Gus M, Abreu-Silva O, Beltrami L and Fuchs F D 2009 Diagnostic of obstructive sleep apnea syndrome and its outcomes with home portable monitoring *Chest* **135** 330–6

Welch P D 1967 The use fast Fourier transform of the estimation of power spectra: a method based on time averaging over short, modified periodograms 1967 *IEEE Trans. Audio Electroacoust.* **15** 70–3

Wong K K H, Jankelson D, Reid A, Unger G, Dungan G, Hedner J A and Grunstein R R 2008 Diagnostic test evaluation of a nasal flow monitor for obstructive sleep apnea detection in sleep apnea research *Behav. Res. Methods* **40** 360–6

Wootters W K 1981 Statistical distance and Hilbert space *Phys. Rev.* D **23** 357–62

Young T, Peppard P E and Gottlieb D J 2002 Epidemiology of obstructive sleep apnea *Am. J. Respir. Crit. Care* **165** 1217–39

Zhang X S, Roy R J and Jensen E W 2001 EEG complexity as a measure of depth anesthesia for patients *IEEE Trans. Biomed. Eng.* **48** 1424–33

Zweig M H and Campbell G 1993 Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine *Clin. Chem.* **39** 561–77

Med Biol Eng Comput
DOI 10.1007/s11517-013-1109-7

ORIGINAL ARTICLE

Pattern recognition in airflow recordings to assist in the sleep apnoea–hypopnoea syndrome diagnosis

Gonzalo C. Gutiérrez-Tobal · Daniel Álvarez · J. Víctor Marcos · Félix del Campo · Roberto Hornero

Received: 11 March 2013 / Accepted: 1 September 2013

Abstract This paper aims at detecting sleep apnoea–hypopnoea syndrome (SAHS) from single-channel airflow (AF) recordings. The study involves 148 subjects. Our proposal is based on estimating the apnoea–hypopnoea index (AHI) after global analysis of AF, including the investigation of respiratory rate variability (RRV). We exhaustively characterize both AF and RRV by extracting spectral, nonlinear, and statistical features. Then, the fast correlation-based filter is used to select those relevant and non-redundant. Multiple linear regression, multi-layer perceptron (MLP), and radial basis functions are fed with the features to estimate AHI. A conventional approach, based on scoring apnoeas and hypopnoeas, is also assessed for comparison purposes. An MLP model trained with AF and RRV selected features achieved the highest agreement with the true AHI (intra-class correlation coefficient = 0.849). It also showed the highest diagnostic ability, reaching 92.5 % sensitivity, 89.5 % specificity and 91.5 % accuracy. This suggests that AF and RRV can complement each other to estimate AHI and help in SAHS diagnosis.

G. C. Gutiérrez-Tobal (✉) · D. Álvarez · J. V. Marcos · R. Hornero
Biomedical Engineering Group, E.T.S.I. de Telecomunicación, University of Valladolid, Paseo Belén 15, 47011 Valladolid, Spain
e-mail: gonzalo.gutierrez@gib.tel.uva.es

F. del Campo
Servicio de Neumología, Hospital Universitario Río Hortega, c/Dulzaina 2, 47012 Valladolid, Spain

F. del Campo
Facultad de Medicina, University of Valladolid, Avenida Ramón y Cajal 7, 47005 Valladolid, Spain

Published online: 22 September 2013

Keywords Sleep apnoea–hypopnoea syndrome · Airflow · Respiratory rate variability · AHI estimation · Pattern recognition

1 Introduction

The sleep apnoea–hypopnoea syndrome (SAHS) is a disease characterized by recurrent episodes of total absence (apnoeas) or significant reduction (hypopnoeas) in airflow (AF) during sleep. SAHS is highly prevalent since up to 5 % of adults are affected [41]. It has been usually related to cardiovascular illnesses [25], motor vehicle collisions [35], and occupational accidents [24]. Recently, it has been also associated with cancer incidence [8].

The current diagnostic standard test is nocturnal polysomnography (PSG). It requires monitoring and recording multiple physiological signals from patients [32]. The origin of the signals can be electrical or mechanical, and each of them can involve one or several channels. The apnoea–hypopnoea index (AHI), which is derived from PSG, is used to establish SAHS. Physicians have to perform an offline inspection of signals such as electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG), oxygen saturation (SpO2), or AF to obtain AHI. Thus, PSG is technically complex and time-consuming [6, 14]. Moreover, it is also costly since requires expensive equipment as well as expert workforce overnight [14]. These restrictions limit the availability of specialized sleep units, leading to long waiting lists and increasing the time until diagnosis and treatment [11]. Thereby, simplifying SAHS diagnosis has become a major concern.

New alternative methods have been proposed to overcome the PSG drawbacks. A common approach is to

analyse reduced sets of signals from PSG in order to decrease complexity, cost, and diagnostic time [14]. We propose evaluating the utility of single-channel AF data to assist in SAHS diagnosis. The respiratory rate variability (RRV), derived from AF [10], is also investigated. The waveform of both signals is directly modified by the occurrence of apnoea and hypopnoea events [10, 19]. Hence, their study is a natural way of dealing with the problem. There exist many recent works focused on processing AF to determine SAHS. Most of them are aimed at scoring apnoeic events to estimate AHI [5, 11, 30, 36, 39]. By contrast, our proposal performs a direct estimation of AHI after a comprehensive analysis of AF and RRV. Thus, the first step is the extraction of statistical moments, nonlinear measures and spectral parameters from the recordings in order to characterize them [2, 15, 26]. This exhaustive characterization of AF and RRV may lead to obtain redundant or non-relevant features. Hence, we include a second step consisting of a feature selection procedure using the fast correlation-based filter (FCBF) [42]. FCBF relies on symmetrical uncertainty (SU) and has been already involved in biomedical applications for cancer recognition [18], neonatal seizure detection [1], or gene classification [13]. Its purpose is to filter data according to their relevancy and redundancy. A final step is included to estimate AHI. Thus, we feed three pattern recognition techniques with the extracted features: multiple linear regression (MLR), multi-layer perceptron neural network (MLP), and radial basis function neural network (RBF). They represent common linear (MLR) and nonlinear (MLP, RBF) methodologies to perform regression tasks [7]. We evaluate the agreement between these estimations and the true AHI of subjects as well as their diagnostic ability. Additionally, we also conduct a conventional approach (scoring apnoeas and hypopnoeas) for comparison purposes. Our hypothesis is that relevant and non-redundant features from single-channel AF could help in SAHS diagnosis by estimating AHI.

2 Materials and methods

Figure 1 presents a scheme of the general methodology carried out in this study. It includes the feature extraction, the feature selection, and the AHI estimation steps, as well as the two kinds of evaluations applied to the estimations from each pattern recognition method and the conventional approach.

2.1 Subjects and signals

This study involved recordings from 148 subjects (100 SAHS-positive and 48 SAHS-negative). The AF data were

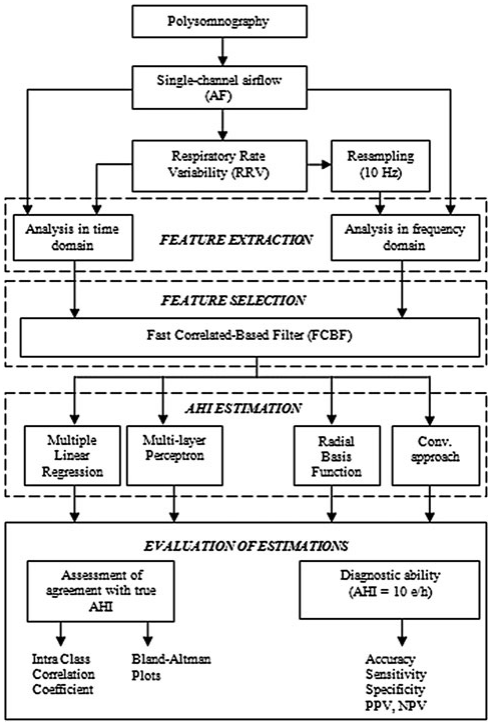

Fig. 1 General scheme of the methodology carried out in the study. *AHI* apnoea-–hypopnoea index, *PPV* positive predictive value, *NPV* negative predictive value

obtained from nocturnal PSG, which was conducted in the sleep unit of the Hospital Universitario Río Hortega (Valladolid, Spain). All subjects were suspected of suffering from SAHS before undergoing PSG due to common symptoms such as daytime sleepiness, loud snoring, nocturnal choking, awakenings, and referring apnoeic events. The physicians established the AHI threshold for a positive diagnosis in 10 events per hour (e/h). The score of apnoeic events was done following the rules of the American Academy of Sleep Medicine (AASM) [19]. Thus, apnoeas were defined as 10-s-or-more episodes of complete cessation of AF. Accordingly, hypopnoeas were defined as 10-s-or-more episodes of 30 % of AF reduction accompanied by a 4 % or more decrease in the saturation of haemoglobin. The Review Board on Human Studies accepted the protocol, and all the subjects gave their informed consent to participate in the study.

The proportion of male subjects was 79 %. No statistically significant differences between SAHS-positive and SAHS-negative samples were encountered in the body

mass index (BMI) or age. The entire group was randomly divided into a training group (60 %) and a test group (40 %). Table 1 summarizes demographic and clinical data from the entire sample, the training group and the test group.

The acquisition of signals during PSG was done by means of a polygraph (Alice 5, Respironics, Philips Healthcare, The Netherlands). AF was obtained through a thermistor (Pro-Tech, Respironics, Philips Healthcare, The Netherlands) at the sample rate of 10 Hz. The length of the AF recordings was 7.24 ± 0.38 h (mean $\pm$ standard deviation). An anti-aliasing filter was applied to satisfy the Nyquist–Shannon theorem. The RRV signal was obtained from AF by measuring the time between consecutive breaths [10]. Thereby, we examined the first derivative of AF to find time intervals in which the original signal grew. We located the AF maximums at each interval. To derive RRV, consecutive locations were used as references to measure the time from one breath to the next [21].

Table 1 Demographic and clinical data for all subjects under study (mean $\pm$ standard deviation)

	All	SAHS positive	SAHS negative
Subjects	148	100	48
All subjects			
Age (years)	50.9 ± 11.7	51.9 ± 11.4	48.7 ± 12.1
Males (%)	79.0	85.0	66.7
BMI (kg/m^2)	29.2 ± 4.7	29.7 ± 4.5	28.1 ± 5.0
Recording time (h)	7.24 ± 0.38	7.23 ± 0.36	7.27 ± 0.43
AHI (h^{-1})		37.15 ± 25.81	4.13 ± 2.39
	All	SAHS positive	SAHS negative
Subjects	89	60	29
Training set			
Age (years)	51.9 ± 11.8	52.8 ± 11.9	50.2 ± 11.7
Males (%)	80.9	88.3	65.5
BMI (kg/m^2)	29.8 ± 5.0	30.5 ± 5.2	28.4 ± 5.7
Recording time (h)	7.22 ± 0.43	7.21 ± 0.38	7.24 ± 0.52
AHI (h^{-1})		37.4 ± 27.2	3.8 ± 2.4
	All	SAHS positive	SAHS negative
Subjects	59	40	19
Test set			
Age (years)	49.2 ± 11.3	50.5 ± 10.7	46.5 ± 12.5
Males (%)	76.3	80.0	68.4
BMI (kg/m^2)	28.3 ± 4.1	28.6 ± 3.5	27.7 ± 5.2
Recording time (h)	7.27 ± 0.29	7.26 ± 0.32	7.30 ± 0.23
AHI (events/h)		26.2 ± 17.2	4.3 ± 2.3

BMI body mass index, *AHI* apnoea–hypopnoea index

2.2 Definition of spectral bands of interest

The recurrent behaviour of apnoeas and hypopnoeas can be characterized by analysing AF and RRV in the frequency domain. Moreover, according to previous studies [15], differences in the spectrum of SAHS-positive and SAHS-negative samples are expected. Thus, the power spectral density (PSD) of the recordings was computed in order to establish these differences. PSD was estimated using the nonparametric Welch method, which is suitable for non-stationary signal analysis [38]. A Hamming window of 2048 (204.8 s) samples (50 % overlap and 4,096-point DFTs) was used. Cubic spline interpolation was previously applied to RRV series in order to resample the recordings to a constant sample rate (10 Hz). The interpolation is not needed to perform the analysis in time domain, and therefore, the resampled version of the RRV recordings was not used in that case.

Spectral bands of interest were defined for AF and RRV. The Mann–Whitney test was applied to each SAHS-positive and SAHS-negative full PSD from the training group. Thus, a p value was computed for each frequency. We located those frequencies at which the lowest p value for AF and RRV was reached (p value $\ll 0.01$). We set the corresponding band limits around these frequencies. In order to minimize type I errors, we chose those frequencies with a corresponding p value smaller than one order of magnitude. Thereby, we maximized the likelihood of defining bands in which truly exist significant differences. According to this procedure, the following spectral bands of interest were determined: [0.022–0.058] Hz for AF and [0.085–0.134] Hz for RRV. Figure 2a, b shows the averaged PSD of SAHS-positive and SAHS-negative samples for AF and RRV, respectively, in the training set.

2.3 Feature extraction

Up to 19 features were used to exhaustively characterize AF and RRV. Statistical moments, nonlinear measures, and spectral parameters were extracted from each full AF and RRV recordings. Thus, subjects were described by patterns composed of the corresponding values for each feature.

2.3.1 Statistical moments

We expected differences between the distribution of the time series amplitude values from SAHS-positive and SAHS-negative samples [15]. Hence, four statistical moments were extracted from AF and RRV. Mean (M_{t1}), standard deviation (M_{t2}), skewness (M_{t3}), and kurtosis (M_{t4}) were computed to quantify central tendency, dispersion, asymmetry, and peakedness of data, respectively.

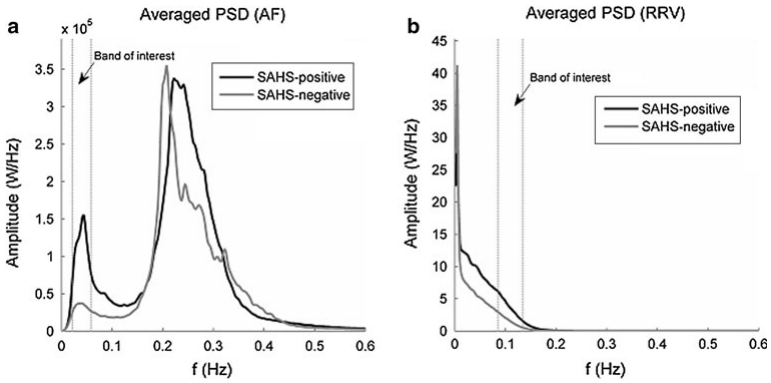

Fig. 2 Low-frequency representation of the averaged PSD for **a** AF and **b** RRV. SAHS-positive group in *solid black line*. SAHS-negative group in *solid grey line*. Corresponding bands of interest into *dashed lines*

2.3.2 *Nonlinear features*

Nonlinear features were used to measure the variability, complexity, and irregularity of the time series. We used central tendency measure (CTM), Lempel–Ziv complexity (LZC), and approximate entropy (ApEn) for this purpose. These methods have been already used to characterize SAHS in previous studies [2, 15, 26].

- Central tendency measure quantifies the degree of variability in time series [8]. It is based on first-order difference plots that can be generated representing $x[n+2]-x[n+1]$ versus $x[n+1]-x[n]$, where $x[n]$ are the time series values. CTM is computed by counting the points falling within a preselected radius ρ and dividing that count by the total number of points [9]. Values closer to 1 indicate lower variability, whereas values closer to 0 indicate higher variability.
- Lempel–Ziv complexity is a measurement of complexity in finite sequences [23]. Thus, the conversion of time series into a finite sequence of symbols is needed. Binary conversion has been commonly applied by using the median as a threshold [29]. Once the sequence is obtained, it is scanned from left to right in order to find new subsequences of consecutive characters [43]. The final number of these subsequences is normalized to make the method independent of the length of sequences. Larger values of LZC correspond to higher complexity [43].
- ApEn measures the irregularity of time series. It assigns higher values to higher irregularity [34]. ApEn was originally developed to be applied over short and noisy data sets and requires the specification of two design parameters: a length m and a tolerance window r [33]. These are used to establish the logarithmic likelihood resulting from those close patterns (within r) for m contiguous observations, which remain close (within the same r) for $m+1$ contiguous observations.

Optimum radius ρ (CTM), length m, and tolerance r (ApEn) were determined by a p value-based methodology [17]. In the case of ApEn, we evaluated $m = 1, 2$ and r ranging 0.10–0.25 times the standard deviation of the times series (with a 0.05 step). These values produce good statistical reproductibility for ApEn [34]. A wide range of values for ρ were also assessed (0.1–30, with a 0.1 step). We selected those configurations, which showed the lowest p value between SAHS-positive and SAHS-negative samples in the training group:

- AF: $\rho = 0.8$ (CTM), $m = 2$, $r = 0.2$ times standard deviation (ApEn).
- RRV: $\rho = 4.8$ (CTM), $m = 2$, $r = 0.2$ times standard deviation (ApEn).

2.3.3 *Spectral features*

A total of 12 parameters were extracted from the full PSD (6) and the band of interest (6) for every AF and RRV recording.

- First-to-fourth statistical moments, which were also extracted in the frequency domain ($M_{f1} - M_{f4}$).
- Peak amplitude (PA), taken as the maximum value of PSDs in a given frequency interval.
- The Wootters distance (WD) [40], which is a disequilibrium measure. WD assigns higher values when the PSD is concentrated into a narrow frequency band (as in sum of sinusoids). If it is uniformly distributed along frequencies (white noise), WD equals zero [27].

2.4 Automatic feature selection: FCBF

After the feature extraction stage, the FCBF algorithm automatically selected relevant and non-redundant features [42]. FCBF is a filter method, which is not dependent on posterior analysis. It relies on symmetrical uncertainty

(SU), which is a normalized measure of information gain (IG) between two variables [42]. The method is divided into two steps. First, a relevance analysis of features was done. SU between the features (X_i) and AHI (Y) was computed as follows:

$$\mathrm{SU}_i(X_i, Y) = 2\frac{\mathrm{IG}_i(X_i, Y)}{H_i(X_i) + H(Y)} \quad i = 1, 2, \ldots, N, \tag{1}$$

where H refers to Shannon's entropy [42], and N is the number of features extracted. SU is restricted to the range [0, 1]: 1 indicates that knowing one feature it is possible to completely predict the other, whereas 0 indicates that the two features are independent [42]. Once SU_i were computed, the features were ranked from more relevant (higher SU_i) to less relevant (lower SU_i). The mean of all SU_i values was used as a cut-off to perform a preselection. The second step was a redundancy analysis. SU between each pair of preselected features ($\mathrm{SU}_{i,j}$) was sequentially computed beginning from the most relevant ones. When $\mathrm{SU}_{i,j} \geq \mathrm{SU}_i$, the feature j was discarded due to redundancy and was not taken into account in successive comparisons. The final selected features were those not discarded after ending the procedure.

2.5 Pattern recognition methods

As described above, the extracted features were used to form patterns (vectors). Thus, a subject n was characterized by a pattern x_n. Each subject and its corresponding x_n are associated with an AHI value (t_n). We modelled the statistical relationship between patterns and AHI by means of pattern recognition techniques. The utility of three methods to provide a reliable estimation (y) of the AHI was evaluated.

2.5.1 Multiple linear regression (MLR)

Multiple linear regression is a traditional method to predict an output variable, y, through data from a multivariate pattern, x_1, x_2,…, x_N. It assumes a linear relationship between the former and the latter [20]:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_N x_N = \mathbf{w}^T \mathbf{x}, \tag{2}$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_N)^T$ are the regression coefficients for each input variable and the intercept (w_0). The computation of $\mathbf{w}$ is done by means of the sum of squares error (E_D) minimization [7]:

$$E_\mathrm{D} = \frac{1}{2}\sum_{n=1}^{N} [y(\mathbf{x}_n, \mathbf{w}) - t_n]^2. \tag{3}$$

2.5.2 Multi-layer perceptron (MLP) network

The MLP network is a model inspired by the human brain. The architecture of MLP is arranged in several interconnected layers (input, hidden layers, and output), which are composed of simple units known as perceptrons [7]. Each perceptron is characterized by an activation function $g(\bullet)$, and their connections to perceptrons from other layers are associated with adaptive weights (w_{ij}).

The output layer provides the response, y. Since our purpose is to estimate a continuous variable, a single output unit with a linear activation function was used [28]. Additionally, we implemented a single hidden layer composed of perceptrons with nonlinear activation functions. This configuration is known to be able of providing universal approximation [7]. Thus, y can be expressed as follows:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{N_H} \left[w_j g\left(\sum_{i=1}^{d} w_{ij} x_i + b_j \right) + b_0 \right], \tag{4}$$

where $\mathbf{w}$ is a vector with all the adaptive parameters (weights and bias), w_j is the weight connecting hidden units h_j with the output unit, b_0 is the bias associated with the output unit, w_{ij} is the weight connecting the input unit i with hidden unit h_j, and b_j is its associated bias. N_H, the number of perceptrons in the hidden layer, is a design parameter. Weights were optimized with patterns from the training group, by sum of squares error function minimization. Scaled conjugate gradient was used for this purpose [7].

Weight decay regularization was used to achieve good generalization. Thus, a penalty term (Ω) was added to the error function E_D, to favour small weights [7]:

$$\begin{aligned} E_T &= E_\mathrm{D} + \Omega \\ &= E_\mathrm{D} + \upsilon \sum_i w_i^2 = \frac{1}{2}\sum_{n=1}^{N} [y(\mathbf{x_n}, \mathbf{w}) - t_n]^2 + \upsilon \sum_i w_i^2, \end{aligned} \tag{5}$$

where Ω is the sum of squares of the network weights, and υ is known as the regularization parameter, which has to be configured.

2.5.3 Radial basis function (RBF) network

Radial basis function is a different neural network approach. This network is composed of a hidden and an output layer. The output y is computed from the responses provided by the basis functions $\psi(\bullet)$ in the hidden layer nodes. These functions only depend on the radial distance (typically the Euclidian distance) between the input vector $\mathbf{x}$ and a set of suitable centres $\mathbf{c}_j$ [7]. A single output neuron with a linear activation function was used to implement the output layer, since the problem was a single variable regression task. Thus, y is given by the following expression [7]:

$$y(\mathbf{x},\mathbf{w}) = \sum_{j=1}^{N_B} w_j \psi_j\left(||\mathbf{x}-\mathbf{c}_j||\right) + b, \quad (6)$$

where N_B is the number of basis functions (or centres), $\mathbf{c}_j$ is the centre of function ψ_j, w_j is the weight connecting ψ_j and the output neuron, and b is the bias parameter for this neuron. A Gaussian function is commonly used for $\psi(\bullet)$ [7]:

$$\psi_j(\mathbf{x}) = \exp\left(-\frac{||\mathbf{x}-\mathbf{c}_j||^2}{2\sigma_j^2}\right). \quad (7)$$

where σ_j is the standard deviation (width) of each function. Thus, the numbers of centres (N_B) and their locations $\mathbf{c}_j$ as well as the widths of radial basis functions σ_j and the weights w_j are parameters to be optimized. N_B and σ_j were experimentally determined during a design stage. *K*-means algorithm was used to optimize the location of centres [7], and w_j was computed from the solution of linear equations following the sum of squares error minimization [7]. All of them were configured by patterns from the training group.

2.6 Conventional approach

A conventional way of dealing with the problem of automatic SAHS diagnosis is to detect and score respiratory events in AF signal. Then, an estimation of AHI (AHI_c) can be derived by dividing the number of these events by the sleep time. We implemented a scoring algorithm to compare it with the proposed pattern recognition techniques. A peak detection algorithm was used to locate inspiratory onsets and endings in AF [21]. These values determined the amplitude of every inspiration. Following the rules of the AASM, we scored those respiratory events that matched 30 % or more drop from the baseline and lasted a minimum of 10 s [19]. The baseline was determined by the mean amplitude of the s previous inspirations [16]. Hence, s was a design parameter. The same methodology than in the case of the parameters of nonlinear features was used to optimize s. We computed AHI_c in the training group by varying s from 1 to 10 (with a 1 step). For each s, the Mann–Whitney test was used to obtain the p value between the AHI_c from the SAHS-positive and the SAHS-negative samples. The greatest statistical difference, i.e. the lowest p value, was obtained for $s = 3$, which was established as the optimum value.

2.7 Statistical analysis

Data did not pass the Lilliefors normality test. Hence, the nonparametric Mann–Whitney significance test was used to assess the differences in SAHS-positive and SAHS-negative samples. We used the intra-class correlation coefficient (*ICC*) and Bland–Altman plots as assessment of agreement between estimated and true AHI. The diagnostic ability of the estimations was assessed by means of sensitivity (proportion of SAHS-positive patients correctly classified), specificity (proportion of SAHS-negative subjects correctly classified), accuracy (percentage of subjects correctly classified over the entire sample), positive predictive value (proportion of positive test result which are true positives), and negative predictive value (proportion of negative test result which are true negatives).

3 Results

Three sets of complete patterns were defined: patterns composed of the 19 AF features (P^c_{AF}); patterns composed of the 19 RRV features (P^c_{RRV}); and patterns composed of the 38 AF and RRV features (P^c_{AF-RRV}). Then, we used the training group to select relevant and non-redundant features through FCBF algorithm. Thus, three new sets of reduced patterns, formed with filtered features, were obtained (P^r_{AF}, P^r_{RRV}, and P^r_{AF-RRV}). The training group was also used in the process of obtaining specific pattern recognition models. This process was divided into two stages: design and training. In the first one, the *ICC* was computed using a leave-one-out cross-validation (loo-cv) procedure to find optimum design parameters for MLP and RBF. In the second one, MLR, MLP, and RBF models were trained by the use of the entire training group.

The test group was used to evaluate our methodology. *ICC* and Bland–Altman plots were used to assess the agreement between the AHI estimations (MLR, MLP, RBF, and the conventional approach) and the actual values of AHI. Furthermore, the diagnostic ability of these estimations was also evaluated. Thus, we used the AHI threshold established by the physicians (AHI = 10 e/h) to derive Se, Sp, Acc, PPV, and NPV in each case.

3.1 Feature selection stage

The FCBF algorithm was applied to P^c_{AF}, P^c_{RRV}, and P^c_{AF-RRV}. The complete patterns were significantly filtered. Thus, the reduced patterns P^r_{AF}, P^r_{RRV}, and P^r_{AF-RRV} were, respectively, composed of: 7 out of 19 AF features (from higher to lower SU: WD_b, M_{f1b}, ApEn, CTM, M_{f3b}, WD, M_{f1}), 5 out of 19 RRV features (from higher to lower SU: CTM, M_{f1b}, M_{t3}, M_{f3}, M_{f1}), and 10 out of 38 AF and RRV features (from higher to lower SU: CTM^{RRV}, WD_b^{AF}, M_{f1b}^{RRV}, M_{t3}^{RRV}, M_{f1b}^{AF}, M_{f3}^{RRV}, M_{f1}^{RRV}, $ApEn^{AF}$, CTM^{AF}, LZC^{RRV}). All the selected features from the spectral bands of interest were more relevant than the features from the full PSDs. Linear and nonlinear features, as well as

frequency and time domain features, were selected in all cases. The presence of AF and RRV features was balanced in P^{r}_{AF-RRV}. Nonetheless, the features from RRV tended to be more relevant than those from AF. CTM from RRV was the most relevant feature in terms of SU.

3.2 Design and training stages

3.2.1 Design of MLP and RBF

A proper design of MLP and RBF networks is required to achieve high generalization ability. It refers to selecting the appropriate model complexity in order to prevent over-fitting and under-fitting effects [7]. The effective complexity of the MLP and RBF models is governed by the design parameters [7]. Thus, we experimentally determined the number of hidden nodes (N_H and N_B), the regularization parameter (υ), and a smoothing parameter (τ), which governs the widths of kernel functions (σ_j) in RBF. Only the training group was used for this purpose.

Figures 3 and 4 show the results of the experiments conducted to determine these parameters. The MLP and RBF were fed with complete (P^{c}_{AF}, P^{c}_{RRV}, P^{c}_{AF-RRV}) and reduced (P^{r}_{AF}, P^{r}_{RRV}, P^{r}_{AF-RRV}) patterns. In each case, the *ICC* was computed for N_H/υ (MLP) or N_B/τ (RBF) pairs, and it was used as selection criterion. *ICC* was estimated through loo-cv, which was repeated ten times due to random initialization of weights and centres of MLP and RBF networks. Then, we averaged the ten *ICC*s to obtain the final value.

Figure 3a–f displays the performance of the MLP networks following this procedure. Figures in the same column correspond to complete (left) or reduced (right) input patterns, respectively. Figures in the same row indicate the origin of the features included in the patterns: AF, RRV, or both signals. υ was assessed according to each set. We chose those υ for which their *ICC* was higher throughout the number of nodes. N_H was varied from 1 to 50, and the optimum value was selected for the sake of the network complexity, i.e. we chose those values from which no substantial *ICC* improvement was observed. Thus, the optimum values were $N_H/\upsilon = 18/6$ (P^{c}_{AF}), 20/11 (P^{c}_{RRV}), 22/8 (P^{c}_{AF-RRV}), 17/3 (P^{r}_{AF}), 13/7 (P^{r}_{RRV}), and 18/2 (P^{r}_{AF-RRV}). Since N_H/υ govern the effective complexity of the networks [7], less complex models were selected as optimum when using reduced patterns.

Figure 4 follows the same scheme for the RBF networks. We varied N_B from 1 to 50 and evaluated τ in 1, 2, 3, 4 and 5. Since the evolution of the *ICC* presented clear absolute maximums, we selected those pairs N_B/τ corresponding with these points. Hence, N_B/τ were the following: 21/2 (P^{c}_{AF}), 7/4 (P^{c}_{RRV}), 7/4 (P^{c}_{AF-RRV}), 18/3 (P^{r}_{AF}), 4/1 (P^{r}_{RRV}), and 5/4 (P^{r}_{AF-RRV}). The optimum models were also less complex in the case of reduced patterns, i.e. fewer nodes N_B were used.

3.2.2 Training of MLR, MLP and RBF models

Specific MLR, MLP and RBF models were obtained from the entire training group. A single MLR model was computed for each set of complete (P^{c}_{AF}, P^{c}_{RRV}, and P^{c}_{AF-RRV}) and reduced (P^{r}_{AF}, P^{r}_{RRV}, and P^{r}_{AF-RRV}) patterns. In the case of MLP and RBF, we computed 100 models for each set, due to random initializations in these networks. The optimum design parameters values, which were obtained in the previous stage, were used in the process.

3.3 Test stage

3.3.1 Intra-class correlation coefficient and Bland–Altman plots

Table 2 shows the *ICC* values reached by the MLR, MLP and RBF models for each set of patterns in the test group. The values for MLP and RBF are presented as mean ± standard deviation of the 100 models previously obtained. One model for each method was selected according to their *ICC*: MLR^{c}_{AF-RRV} (P^{c}_{AF-RRV} from MLR), MLP^{r}_{AF-RRV} (P^{r}_{AF-RRV} from MLP), and RBF^{r}_{AF} (P^{r}_{AF} for RBF). Thus, MLP^{r}_{AF-RRV} outperformed AHI_c in terms of agreement and both of them outperformed MLR^{c}_{AF-RRV} and RBF^{r}_{AF}. This tendency was also observed when applying graphical analysis. Figure 5 displays the "Bland–Altman"—(a, c, e, g)—and "estimated versus true AHI" plots—(b, d, f, h). Both graphs show smaller deviation from the target AHI in the case of MLP^{r}_{AF-RRV} and AHI_c. These models also reached less dispersion in the scatter of the points, which is reflected in the corresponding 95 % confidence interval: [−15.6, 19.9] e/h in the case of MLP^{r}_{AF-RRV} and [−16.6, 19.3] e/h for AHI_c.

3.3.2 Diagnostic performance of the models

To complete the analysis, we evaluated the diagnostic ability of the four AHI estimations obtained from the test group. Table 3 shows sensitivity (Se), specificity (Sp), accuracy (Acc), positive predictive value (PPV), and negative predictive value (NPV) for each method. The highest performance was achieved by MLP^{r}_{AF-RRV}, which reached 92.5 % Se, 89.5 % Sp, 91.5 % Acc, 94.9 % PPV, and 85.0 % NPP. MLR^{c}_{AF-RRV} and RBF^{r}_{AF} also outperformed AHI_c at each statistic.

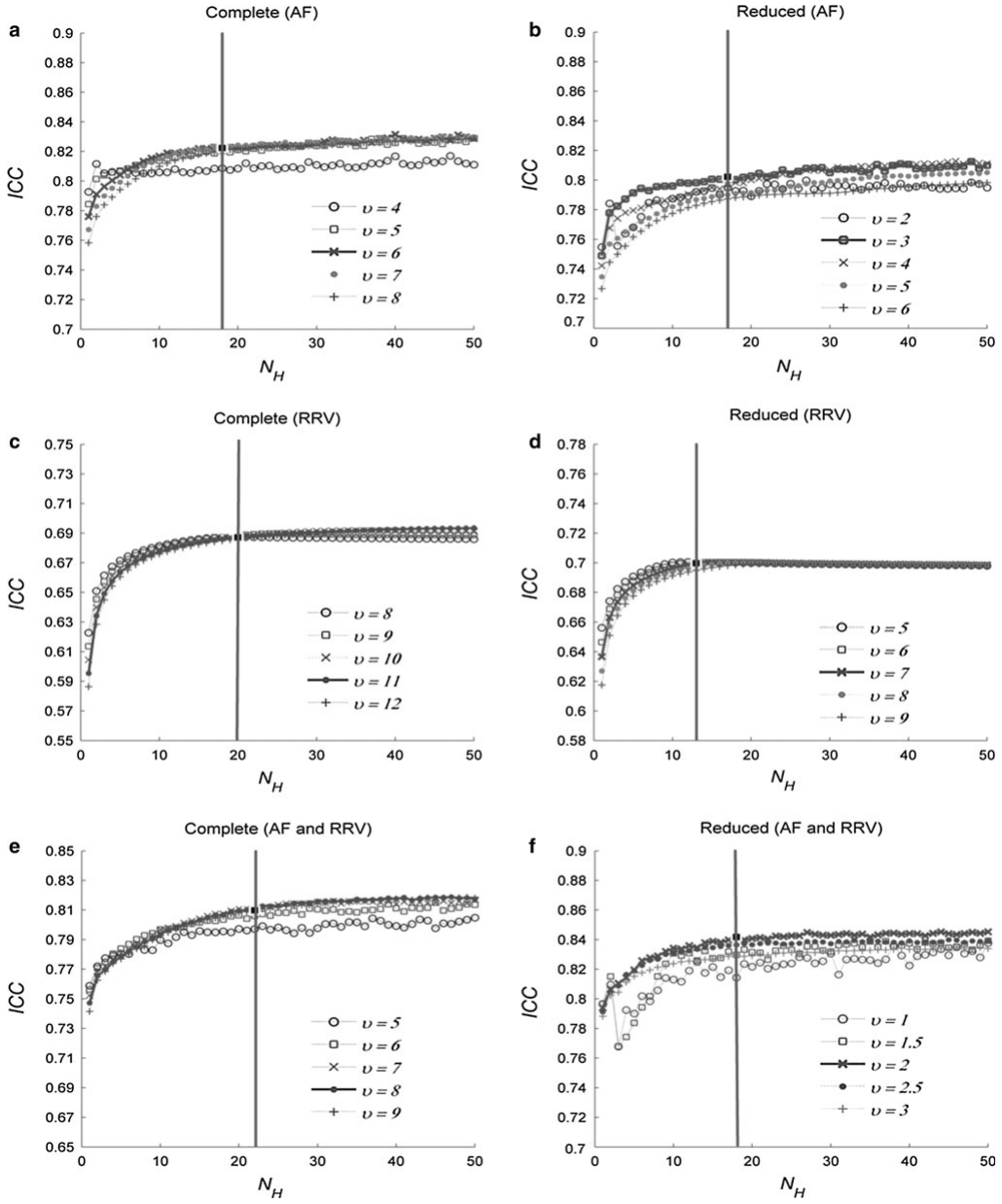

Fig. 3 MLP design stage: *ICC* for different N_H and υ values. Optimum values of: υ marked in *solid line*; N_H marked in *vertical line*

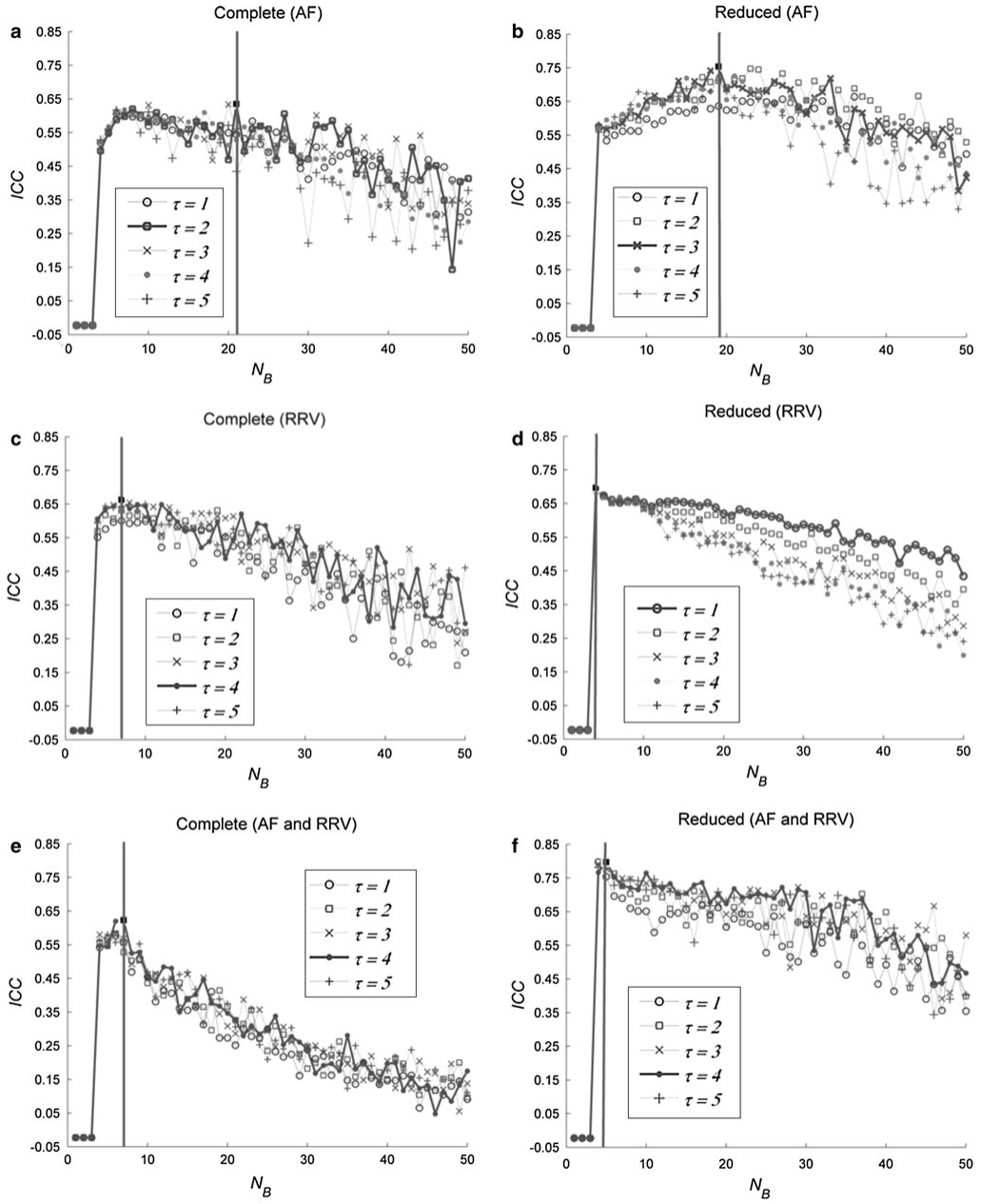

Fig. 4 RBF design stage: *ICC* for different N_B and τ values. Optimum values of: τ marked in *solid line*; N_B marked in *vertical line*

Table 2 *ICC* obtained from MLR, MLP, RBF, and the conventional approach (AHI_c)

	ICC test		
	AF	RRV	AF-RRV
AHI_c	**0.840**	–	–
MLR			
P^c	0.796	0.710	**0.809**
P^r	0.650	0.689	0.777
MLP			
P^c	0.782 ± 0.002	0.644 ± 4.3 e-4	0.808 ± 1.7-5
P^r	0.743 ± 0.002	0.685 ± 1.1 e-4	**0.849 ± 0.002**
RBF			
P^c	0.594 ± 0.094	0.617 ± 0.022	0.632 ± 0.170
P^r	**0.748 ± 0.037**	0.703 ± 0.006	0.732 ± 0.016

Best performance for each method in bold

P^c complete patterns, P^r reduced patterns

4 Discussion and conclusions

In this study, we addressed the estimation of AHI by pattern recognition in single-channel AF. Our approach focused on the exhaustive analysis of AF and RRV signals. Thus, spectral, nonlinear, and statistical features were obtained from all recordings. FCBF algorithm filtered these features, discarding those non-relevant or redundant. After filtering, both linear and nonlinear features from AF and RRV were selected. Moreover, all the features selected from the spectral bands of interest were more relevant in terms of SU than those selected from the full PSDs. The FCBF method was also useful in the design of MLP and RBF. Thereby, optimum less complex networks were selected in both cases when using reduced patterns (P^r_{AF}, P^r_{RRV}, and P^r_{AF-RRV}) instead of complete patterns (P^c_{AF}, P^c_{RRV}, and P^c_{AF-RRV}). These results support the use of the AF and RRV signals, as well as the methodology conducted to characterize them.

During the test stage, the agreement between the AHI estimations and the true AHI was evaluated. We selected specific models according to their *ICC*. Both *ICC* and graphical analysis supported MLP^r_{AF-RRV} and AHI_c as the best in terms of agreement. The conventional approach, however, systematically overestimated AHI in the SAHS-negative sample (15 out of 19 subjects) and underestimated AHI in the SAHS-positive sample (27 out of 40 subjects) (Fig. 5 b). These two effects may have caused that, despite having lower *ICC* values, MLR^c_{AF-RRV} and RBF^r_{AF} reached higher global diagnostic ability than AHI_c.

The diagnostic ability of the methods was also assessed. The highest performance was achieved by the AHI estimation derived from the MLP^r_{AF-RRV} model. This model reached high sensitivity (92.5 %), specificity (89.5 %), and accuracy (91.5 %). Only 2 out of 19 SAHS-negative subjects (false positives) and 3 out of 40 SAHS-positive subjects (false negatives) were misclassified. Additionally, three out of them have borderline true AHI values (5.7, 10, and 15.8 e/h). Thus, 94.9 % of subjects that our model estimated SAHS-positive were actually suffering from SAHS. Moreover, 85.0 % of subjects that our model predicted SAHS-negative were not SAHS patients. These findings confirmed the usefulness of combining relevant and non-redundant features from AF and RRV.

Fig. 5 Bland–Altman plots (**a**, **c**, **e**, **g**) and "estimated versus true AHI" (**b**, **d**, **f**, **h**), for the specific models and the conventional approach (AHI_c). Results derived from the test group. *TP* true positives, *FP* false positives, *TN* true negatives, *FN* false negatives ▶

Recent studies aimed at identifying SAHS (AHI threshold = 10 e/h) from single-channel AF. Most of them detected and scored respiratory events to estimate AHI. Shochat et al. [36] investigated the usefulness of SleepStrip™ for this purpose. They acquired AF through a thermistor and involved 288 subjects. Sensitivity was 86.0 %, but specificity reached low values (57.0 %). Nakano et al. [30] scored events supported by a spectral analysis of AF. The best performance was achieved using 116 AF recordings acquired with a thermocouple: 92 % Se and 90 % Sp. Their results are similar to ours from MLP^r_{AF-RRV}. Nonetheless, no further comparison was possible since no data were reported to obtain Acc, PPV or NPV. Nasal prong pressure sensor (NPP) has been widely used to acquire AF in portable diagnostic devices. Thus, De Almeida et al. assessed SleepCheck™ [11]. The authors reported 85.7 % Se and 87.5 % Sp by using a small sample size (30 subjects). Additionally, Wong et al. [39] evaluated FlowWizard™. They achieved high diagnostic performance: 92 % Se, 86 % Sp, 96 % PPV, and 75 % NPV. However, only 27 SAHS-positive subjects and 7 SAHS-negative subjects were used. Finally, ApenaLink™ was recently evaluated by BaHammam et al. [5]. The study involved 95 AF recordings. Specificity and PPV reached high values (89.0 and 91.0 %, respectively), but sensitivity (70.0 %) and NPV (63.0 %) were low. In contrast to the conventional approach conducted in these studies, our methodology took into account not only the apnoeic events but also data from the whole single-channel AF. A similar approach was performed in a recent study of our research group [15]. The utility of AF and RRV signals was assessed by the use of a logistic regression model, i.e. into a binary classification task. After a loo-cv process, the diagnostic performance reached 88 % Se, 70.8 % Sp, 82.4 % Acc, 86.3 % PPV, and 73.9 % NPV.

There also exist SAHS studies not aimed at assessing the diagnostic ability of a given methodology, but focused on evaluating how well this methodology detects apnoeas

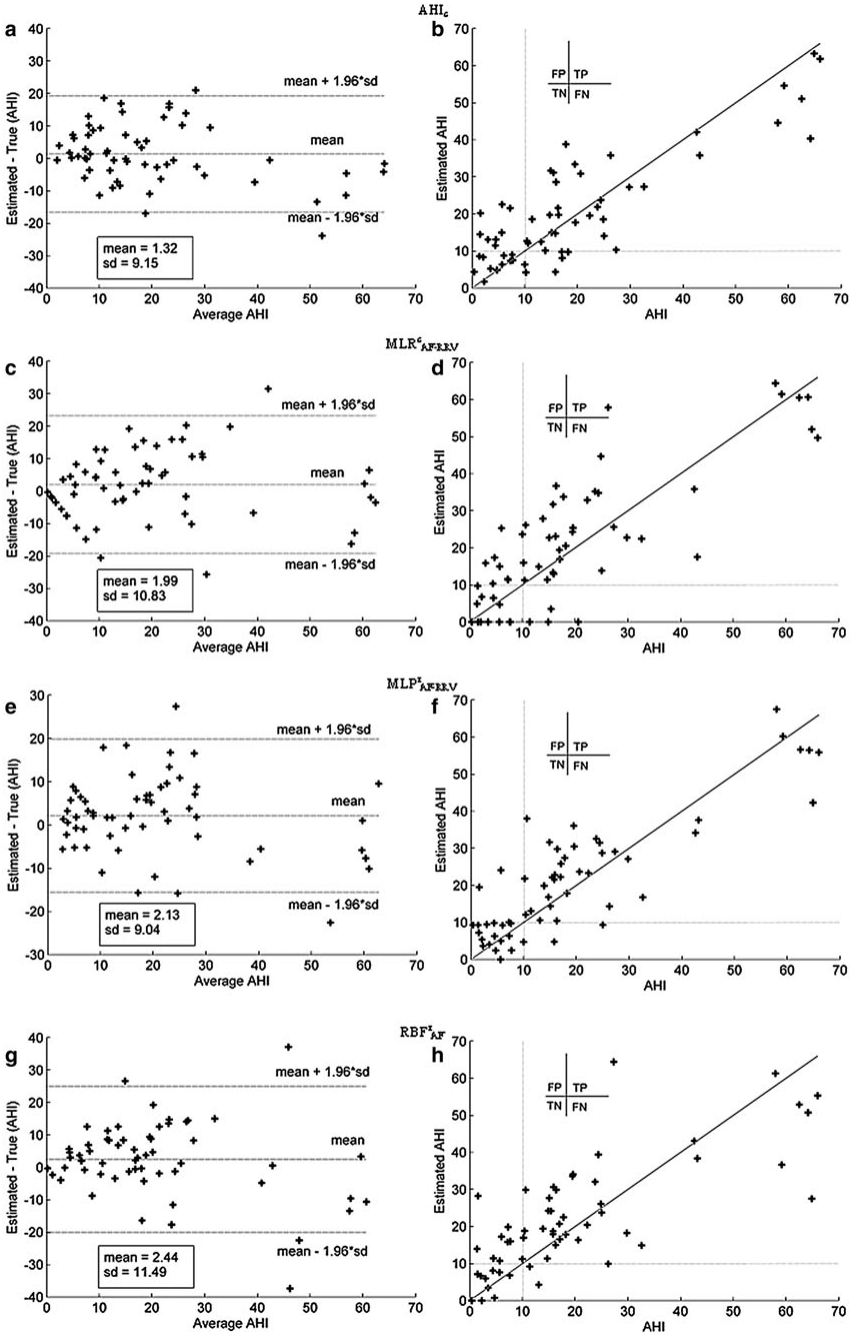

Table 3 Diagnostic performance of the specific models on the test group: comparison with the conventional approach (AHI_c)

	Se (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)
AHI_c	87.5	57.9	78.0	81.4	68.7
MLR^c_{AF-RRV}	90.0	63.2	81.4	83.7	75.0
MLP^r_{AF-RRV}	92.5	89.5	91.5	94.9	85.0
RBF^r_{AF}	92.5	57.9	81.4	82.2	78.6

Se sensitivity, *Sp* specificity, *Acc* accuracy, *PPV* positive predictive value, *NPP* negative predictive value

and/or hypopnoeas. Han et al. [16] used AF recordings from NPP, along with an automatic algorithm based on the mean magnitude of the second derivative, to detect apnoeas. They reported 92.4 % Se and 88.3 % Sp when comparing their methodology with the manual score of the events. Álvarez-Estévez and Moret-Bonillo [3] applied a fuzzy algorithm to AF, SpO_2, and respiratory movement recordings in order to detect respiratory events and classify them into apnoeas or hypopnoeas. Their results showed 87 % Se and 89 % Sp in the detection task, whereas they reported 92/85 % Se and 85/92 % Sp in the classification task (apnoeas/hypopnoeas). Otero et al. [31] propose several algorithms to detect different pathological events from polysomnographic recordings. Their results showed 97.4 and 94.0 % PPV when detecting apnoeas and hypopnoeas, respectively.

Pattern recognition techniques have been already shown to be useful in SAHS detection. Varady et al. [37] trained four feed-forward artificial neural networks to detect apnoeic segments in AF recordings. Data from AF and respiratory inductive plethysmography (RIP) were used. Up to 93 % of patterns were correctly classified into normal, apnoea, or hypopnoea categories. No assessment of diagnostic ability was performed. El-Shol et al. [12] trained a MLP network to predict AHI from demographic and clinical variables of subjects. Sensitivity and specificity reached 94.9 and 64.7 %, whereas PPV and NPV were 87.9 and 85.2 %, respectively. Additionally, in other study of our research group [26], 14 features extracted from 240 SpO_2 recordings were used along with MLR and MLP algorithms. The *ICCs* were 0.80 and 0.91, respectively. The MLP model showed the highest diagnostic performance: 89.6 % Se, 81.2 % Sp, 86.8 % Acc, 90.5 % PPV, and 79.6 % NPV.

Although our methods have revealed the usefulness of AF and RRV in SAHS detection, some limitations have to be addressed. A larger sample size would improve the generalization of our results. Accordingly, the validation of the proposed algorithms using different databases would be of great interest to enhance their statistical power [22]. Moreover, the use of subjects without previous suspects of suffering from SAHS would complement our findings. Nonetheless, this issue has no easy solution since subjects usually undergo overnight PSG after referring some symptoms. The cut-off AHI = 10 e/h is widely used to determine SAHS [5, 30, 36, 39]. Hence, our methodology was optimized according to this threshold. Future works, however, could assess our methodology for other common cut-offs such as 5 or 15 e/h. Another limitation is the use of a thermistor, instead of a thermistor and a NPP simultaneously. The AASM recommends using both sensors to acquire AF [19], due to weaknesses in the two of them [4]. Additionally, it is well known that NPP outperforms thermistor when recording respiratory events [4]. However, this work has shown that a global analysis of single-channel AF from thermistor can achieve high diagnostic performance and improve the results reported in recent studies only involving NPP [5, 11, 39]. The application of our methodology to AF recordings from NPP is a future goal. Another future goal is to assess relationships between the proposed features and the apnoeic events in order to clarify their physiological meaning. Additionally, our methodology does not offer flexibility to the physicians in order to change the AHI based on their expertise. However, the results reported in this study measure to what extent physicians can trust our AHI estimations. Finally, the main benefit of our approach would be obtained by applying our algorithms to single-channel AF recordings acquired at patient's domicile. Although there exist several portable devices to obtain AF [5, 11, 36, 39], these have limitations and need further investigation to ensure their reliability in unattended studies at home.

In summary, single-channel AF from thermistor can be used to assist in SAHS detection and simplify diagnosis. The methodology conducted over AF and RRV signals has shown its usefulness to estimate AHI. Particularly, the FCBF algorithm was successfully used to discard redundant and non-relevant information from recordings, which in turn decreased the complexity of the models obtained through neural networks. An MLP model, trained with relevant and non-redundant features from AF and RRV, achieved high results in terms of agreement with true AHI and diagnostic ability. It outperformed a conventional approach, based on scoring apnoeas and hypopnoeas, conducted over the same database. Additionally, the MLP approach also improved the diagnostic ability of the conventional one conducted in other studies. Our results suggest that AF and RRV complement each other in the AHI estimation and can help in SAHS diagnosis.

Acknowledgments This research was supported in part by the "Consejería de Educación (Junta de Castilla y León)" under project VA111A11-2, the Project Cero 2011 on Ageing from Fundación General CSIC, and project TEC2011-22987 from Ministerio de Economía y Competitividad and FEDER. G. C. Gutiérrez-Tobal was

in receipt of a PIRTU grant from the Consejería de Educación de la Junta de Castilla y León and the European Social Fund (ESF).

References

1. Aarabi A, Wallois F, Grebe R (2006) Automated neonatal seizure detection: a multistage classification system through feature selection based on relevancy and redundancy analysis. Clin Neurophysiol 117:328–340
2. Álvarez D, Hornero R, Marcos JV, del Campo F (2010) Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. IEEE Trans Biomed Eng 57:2816–2824
3. Álvarez-Estévez D, Moret-Bonillo V (2009) Fuzzy reasoning used to detect apneic events in the sleep apnea-hypopnea syndrome. Expert Syst Appl 36:7778–7785
4. BaHammam A (2004) Comparison of nasal prong pressure and thermistors measurements for detecting respiratory events during sleep. Respiration 71:385–390
5. BaHammam A, Sharif M, Gacuan DE, George S (2011) Evaluation of the accuracy of manual and automatic scoring of a single airflow channel in patients with a high probability of obstructive sleep apnea. Med Sci Monit 17:MT13–MT19
6. Bennet JA, Kinnear WJM (1999) Sleep on the cheap: the role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome. Thorax 54:958–959
7. Bishop CM (1996) Neural networks for pattern recognition. Oxford University Press, Oxford, UK
8. Campos-Rodríguez F, Martínez-García MA, Martínez M, Duran-Cantolla J, de la Peña M, Masdeu MJ, González M, del Campo F, Gallego I, Martín JM, Barbe F, Monstserrat JM, Farre R (2013) Association between obstructive sleep apnea and cancer incidence in a large multicenter Spanish cohort. Am J Respir Crit Care Med 187:99–105
9. Cohen ME, Hudson DL, Deedwania PC (1996) Applying continuous chaotic modelling to cardiac signal analysis. IEEE Eng Med Biol Mag 15:97–102
10. Cysarz D, Zerm R, Bettermann H, Frühwirth M, Moser M, Kröz M (2008) Comparison of respiratory rates derived from heart rate variability, ECG amplitude, and nasal/oral airflow. Ann Biomed Eng 36:2085–2094
11. De Almeida FR, Ayas NT, Otsuka R, Ueda H, Hamilton P, Ryan FC, Lowe AA (2006) Nasal pressure recordings to detect obstructive sleep apnea. Sleep Breath 10:62–69
12. El-Solh AA, Mador MJ, Ten-Brock E, Shucard DW, Abul-Khoudoud M, Grant BJB (1999) Validity of neural network in sleep apnea. Sleep 22:105–111
13. Fernández-Navarro F, Hervás-Martínez C, Ruiz R, Riquelme JC (2012) Evolutionary generalized radial basis function neural networks for improving prediction accuracy in gene classification using feature selection. Appl Soft Comput 12:1787–1800
14. Flemons WW, Littner MR, Rowley JA, Gay P, Anderson WM, Hudgel DW, McEvoy RD, Loube DI (2003) Home diagnosis of sleep apnea: a systematic review of the literature. Chest 124:1543–1579
15. Gutiérrez-Tobal GC, Hornero R, Álvarez D, Marcos JV, del Campo F (2012) Linear and nonlinear analysis of airflow recordings to help in sleep apnoea-hypopnoea syndrome diagnosis. Physiol Meas 33:1261–1275
16. Han J, Shin HB, Jeong DU, Park KS (2008) Detection of apnoeic events from single channel nasal airflow using 2nd derivative method. Comput Methods Progr Biomed 98:199–207
17. Hornero R, Alonso A, Jimeno N, Jimeno A, López M (1999) Nonlinear analysis of time series generated by schizophrenic patients. IEEE Eng Med Biol Mag 3:84–90
18. Hu Q, Pan W, An S, Ma P, Wei J (2010) An efficient gene selection technique for cancer recognition based on neighborhood mutual information. Int J Mach Learn Cybern 2:63–74
19. Iber C, Ancoli-Israel S, Chesson A, Quan SF (2007) The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. American Academy of Sleep Medicine, Westchester, IL
20. Jobson JD (1991) Applied multivariate data analysis. Regression and experimental design, vol I. Springer, New York
21. Korten JB, Haddad GG (1989) Respiratory waveform pattern recognition using digital techniques. Comput Biol Med 19:207–217
22. Lado MJ, Vila XA, Rodríguez-Liñares L, Méndez AJ, Oliveri DN, Félix P (2011) Detecting sleep apnea by heart rate variability analysis: assessing the validity of databases and algorithms. J Med Syst 35:473–481
23. Lempel A, Ziv J (1976) On the complexity of finite sequences. IEEE Trans Inform Theory 24:530–536
24. Lindberg E, Carter N, Gislason T, Janson C (2001) Role of snoring and daytime sleepiness in occupational accidents. Am J Respir Crit Care Med 164:2031–2035
25. López-Jiménez F, Kuniyoshi FHS, Gami A, Somers VK (2008) Obstructive sleep apnea: implications for cardiac and vascular disease. Chest 133:793–804
26. Marcos JV, Hornero R, Álvarez D, Aboy M, del Campo F (2012) Automated prediction of the apnea-hypopnea index from nocturnal oximetry recordings. IEEE Trans Biomed Eng 59:141–149
27. Martín MT, Plastino A, Rosso OA (2003) Statistical complexity and disequilibrium. Phys Lett A 311:126–132
28. Nabney IT (2002) NETLAB: algorithms for pattern recognition. Springer, Berlin
29. Nagarajan R (2002) Quantifying physiological data with Lempel-Ziv complexity—certain issues. IEEE Trans Biomed Eng 49:1371–1373
30. Nakano H, Tanigawa T, Furukawa T, Nishina S (2007) Automatic detection of sleep-disordered breathing from single-channel airflow record. Eur Respir J 29:728–736
31. Otero A, Félix P, Álvarez MR (2011) Algorithms for the analysis of polysomnographic recordings with customizable criteria. Expert Syst Appl 38:10133–10146
32. Patil SP, Schneider H, Schwartz AR, Smith PL (2007) Adult obstructive sleep apnea: pathophysiology and diagnosis. Chest 132:325–337
33. Pincus SM (1991) Approximate entropy as a measure of system complexity. Proc Natl Acad Sci 88:2297–2301
34. Pincus SM (2001) Assessing serial irregularity and its implications for health. Ann N Y Acad Sci 954:245–267
35. Sassani A, Findley LJ, Kryger M, Goldlust E, George C, Davidson TM (2004) Reducing motor-vehicle collisions, cost, and fatalities by treating obstructive sleep apnea syndrome. Sleep 27:453–458
36. Shochat T, Hadas N, Kerkhofs M, Herchuelz A, Penzel T, Peter JH, Lavie P (2002) The SleepStripTM: an apnoea screener for the early detection of sleep apnoea syndrome. Eur Respir J 19:121–126
37. Várady P, Micsik T, Benedek S, Benyó Z (2002) A novel method for the detection of apnea and hypopnea events in respiration signals. IEEE Trans Biomed Eng 49:936–942
38. Welch PD (1967) The use of fast Fourier transform of the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust AU-15:70–73
39. Wong KKH, Jankelson D, Reid A, Unger G, Dungan G, Hedner JA, Grunstein RR (2008) Diagnostic test evaluation of a nasal flow monitor for obstructive sleep apnea detection in sleep apnea research. Behav Res Methods 40:360–366

40. Wootters WK (1981) Statistical distance and Hilbert space. Phys Rev D 23:357–362
41. Young T, Peppard PE, Gottlieb DJ (2002) Epidemiology of obstructive sleep apnea. Am J Respir Crit Care 165:1217–1239
42. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224
43. Zhang XS, Roy RJ, Jensen EW (2001) EEG complexity as a measure of depth anesthesia for patients. IEEE Trans Biomed Eng 48:1424–1433

Entropy **2015**, *17*, 123-141; doi:10.3390/e17010123

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Assessment of Time and Frequency Domain Entropies to Detect Sleep Apnoea in Heart Rate Variability Recordings from Men and Women

Gonzalo C. Gutiérrez-Tobal [1,*], Daniel Álvarez [1], Javier Gomez-Pilar [1], Félix del Campo [2,3] and Roberto Hornero [1]

[1] Biomedical Engineering Group, Universidad de Valladolid, Paseo Belén 15, 47011 Valladolid, Spain; E-Mails: dalvgon@gmail.com (D.A.); javier.gomez@gib.tel.uva.es (J.G.-P.); robhor@tel.uva.es (R.H.)

[2] Facultad de Medicina, Universidad de Valladolid, Avenida Ramón y Cajal 7, 47007 Valladolid, Spain; E-Mail: campo@med.uva.es

[3] Hospital Universitario Río Hortega, Calle Dulzaina 2, 47012 Valladolid, Spain

* Author to whom correspondence should be addressed; E-Mail: gonzalo.gutierrez@gib.tel.uva.es; Tel.: +34-983-423-000 (ext. 4716).

Academic Editor: Niels Wessel

Received: 30 October 2014 / Accepted: 31 December 2014 / Published: 6 January 2015

Abstract: Heart rate variability (HRV) provides useful information about heart dynamics both under healthy and pathological conditions. Entropy measures have shown their utility to characterize these dynamics. In this paper, we assess the ability of spectral entropy (SE) and multiscale entropy (MsE) to characterize the sleep apnoea-hypopnea syndrome (SAHS) in HRV recordings from 188 subjects. Additionally, we evaluate eventual differences in these analyses depending on the gender. We found that the SE computed from the very low frequency band and the low frequency band showed ability to characterize SAHS regardless the gender; and that MsE features may be able to distinguish gender specificities. SE and MsE showed complementarity to detect SAHS, since several features from both analyses were automatically selected by the forward-selection backward-elimination algorithm. Finally, SAHS was modelled through logistic regression (LR) by using optimum sets of selected features. Modelling SAHS by genders reached significant higher performance than doing it in a jointly way. The highest diagnostic ability was reached by modelling SAHS in women. The LR classifier achieved 85.2% accuracy (Acc) and 0.951 area under the ROC curve (AROC). LR for men reached 77.6% Acc and 0.895 AROC, whereas LR for the whole

set reached 72.3% Acc and 0.885 AROC. Our results show the usefulness of the SE and MsE analyses of HRV to detect SAHS, as well as suggest that, when using HRV, SAHS may be more accurately modelled if data are separated by gender.

Keywords: sleep apnoea; spectral entropy; multiscale entropy; heart rate variability

PACS Codes: 87.85.Ng; 87.19.Hh; 87.19.lo

1. Introduction

The sleep apnoea-hypopnoea syndrome (SAHS) is a highly prevalent disease which negatively impacts both the health and quality of life of affected people [1]. SAHS is mainly characterized by the recurrence of both total breathing cessation (apnoea events) and significant airflow reduction (hypopnoea events) during sleep time [2]. Apnoeas and hypopnoeas cause oxygen desaturations and sleep fragmentation [2], preventing patients from resting while sleeping, and leading to daytime symptoms such as morning headaches, excessive sleepiness, memory loss, or decreased concentration [3]. Apnoeic events are also related to challenging processes for different main body systems. In this regard, hypoxemia, hypercapnia, inspiratory overexertion, or arousals may vary the normal response of systems such as neural, cardiovascular, and metabolic [1]. Thus, SAHS has been associated with major pathological conditions such as hypertension, stroke, coronary artery disease, congestive heart failure, atrial fibrillation, or diabetes [1–3].

Simplifying SAHS diagnosis has become a major concern for experts in recent years. The standard diagnostic test is overnight polysomnography (PSG), which is technically complex and expensive [4], since it involves monitoring and recording multiple physiological signals such as electroencephalogram (EEG), electrocardiogram (ECG), electromyogram, oxygen saturation (SpO_2), and airflow [5]. PSG is also time-consuming since the physicians need an offline inspection of these recordings to diagnose SAHS. Additionally, it is well-known that SAHS is an underdiagnosed disease. As a result, there exists an increasing demand of PSG tests [6], which exceed the clinical resources in many of the Western countries [7]. The limitations of overnight PSG have led to a search for diagnostic alternatives for SAHS. In this regard, one common approach has been the analysis of reduced sets of signals chosen among those involved on PSG [6].

Heart rate variability (HRV), which is derived from ECG, has been widely investigated to assess multiple conditions related to the heart and the autonomic nervous system (ANS) [8]. This connection between the heart function and the ANS, extensively reported in the literature, provides a unique framework when studying SAHS which is not present in the case of other signals involved in PSG. In this regard, the ANS response to the apnoeic events has been associated with a recurrent progressive-bradycardia/abrupt-tachycardia pattern observed in HRV [9,10]. Therefore, HRV has been usually studied through different approaches to gain insight into SAHS and help in its diagnosis [9–14]. The recurrence of bradycardia-tachycardia patterns justifies the use of frequency analyses and the definition of spectral bands of interest. Thus, the power in the very low frequency band (VLF, 0–0.04 Hz.), in the low frequency band (LF, 0.04–0.15 Hz.), and in the high frequency band (HF, 0.15–0.4 Hz.) has

been successfully assessed in studies involving SAHS [11,12]. On the other hand, some works have reported chaotic heart beat behaviours [15,16], which suggest applying of nonlinear analyses to HRV.

Entropy measures, as a common choice to quantify nonlinear dynamics in biomedical signals, have shown to be useful in the study of different pathologies and physiological conditions like Alzheimer's disease (AD) [17], diabetes [18], atrial fibrillation [19], or SAHS [20,21]. Particularly, the multiscale entropy (MsE) analysis has been widely applied to biomedical signals in order to quantify their irregularity (or complexity) over time scales. Thus, MsE has shown its usefulness to study heart rate dynamics [22], to find differences in the HRV from healthy subjects and subjects suffering from congestive heart failure and atrial fibrillation [23], to show different behaviours in the heart rate of young and elder [23], to quantify the complexity of human gait [24], to improve the knowledge of the EEG behaviour in AD patients [25], as well as to evaluate the effects of drugs in the EEG of schizophrenia patients [26]. Moreover, spectral entropy (SE) has been helpful to quantify the depth of anaesthesia in EEG recordings from women undergoing gynaecological surgery [27], to detect endpoints in speech signals recorded in noisy environments [28], to show the changes that AD causes in the spectrum of magnetoencephalographic and EEG recordings [17,29], as well as to enhance the automatic detection of SAHS from single-channel SpO_2 recordings obtained during nocturnal oximetry [30].

We hypothesize that both SE and MsE can be useful to gain insight into the effects that SAHS causes in HRV and, consequently, to help in its detection. Thus, the main objective of this paper is the assessment of these analyses in the context of SAHS. As mentioned above, HRV has been commonly used to help in SAHS diagnosis, both in frequency and time domain. However, no studies have been found showing the behaviour of HRV in SAHS patients (SAHS-positive) and no-SAHS subjects (SAHS-negative) through SE and MsE, or combining them to automatically detect SAHS. Furthermore, HRV is well-known to be affected by gender [8]. In this regard, HRV has shown significantly lower amplitude in healthy women of all ages, along with lower standard deviation [31]. These findings justify take eventual gender specificities into account when analysing HRV. Therefore, we firstly propose to analyse SE and MsE in HRV recordings from SAHS-positive and SAHS-negative considering these potential differences, *i.e.*, considering the whole set of recordings as well as dividing it into women and men. Then, the use of the automatic forward-selection backward-elimination algorithm (FSBE) is proposed to obtain optimum sets of features from the three sets. This analysis highlights relevant features as well as allows evaluating the complementarity of SE and MsE when modelling SAHS [32]. Finally, we assess the diagnostic ability of logistic regression models built with these features and we compare the results reached for women, men, and the whole set of recordings.

2. Methodology

2.1. Subjects and Signals under Study

The study involved 188 subjects (134 men and 54 women) sent to the sleep unit of the Hospital Universitario Rio Hortega (Valladolid, Spain) due to suspicion of SAHS. All subjects underwent overnight PSG. No subjects with reported cardiac illnesses were included in the study. Apnoeas and hypopnoeas were scored by a single expert, who followed the rules of the American Academy of Sleep Medicine (AASM) [33]. An apnoea-hypopnoea index (AHI) of 10 events per hour (e/h) was established

as the threshold for a positive diagnosis. Accordingly, 93 men and 26 women were considered as SAHS-positive. The Ethics Committee of the Hospital Universitario Rio Hortega (Spain) accepted the protocol and all the subjects gave their informed consent. Table 1 shows demographical and clinical data from the subjects (mean ± standard deviation). No statistically significant differences were found in body mass index (BMI) and age between SAHS-positive and SAHS-negative groups in men and women (Mann-Whitney U test, p-value > 0.01), nor were found in AHI, BMI, and age between men and women in SAHS-positive and SAHS-negative groups (p-value > 0.01).

Table 1. Demographic and clinical data from the subjects under study.

	Women		**Men**	
	SAHS-Negative	**SAHS-Positive**	**SAHS-Negative**	**SAHS-Positive**
#Subjects	28	26	41	93
Age (years)	49.2 ± 8.6	58.3 ± 14.3	46.0 ± 13.1	51.1 ± 11.7
BMI (kg/m^2)	26.8 ± 6.9	28.8 ± 5.8	28.8 ± 5.6	29.2 ± 2.9
AHI (e/h)	3.3 ± 2.3	32.8 ± 24.7	4.1 ± 2.5	33.0 ± 22.5

PSG was carried out with a polysomnograph (Alice 5, Respironics, Philips Healthcare, The Netherlands). The HRV signals were obtained from ECG, which was recorded during overnight PSG (6 to 8 h) at a sample rate of 200 Hz. Each sample in the HRV signal is the time between two consecutive R peaks [34]. Hence, to derive HRV, we firstly applied a QRS-complex detection algorithm [35]. It was reported to reach high sensitivity (99.94%) and positive predictive value (99.93%), even in the presence of muscular noise and baseline artefacts (99.88% sensitivity and 99.73% positive predictive value, respectively) [35]. It is based on Hilbert transform and consists of two stages. Initially, the first differential of the ECG signal is computed (dECG). This is carried out to avoid baseline shifts and motion artefacts. Then the Hilbert transform is applied to dECG (h(n) = H[dECG]). Due to the properties of Hilbert transform, points around peaks in h(n) are regions of high probability of containing actual QRS peaks [35]. Since in h(n) the P and T waves are low comparing with the R waves [35], an adaptive threshold is used to establish those regions truly corresponding to R waves. In the second stage of the algorithm, these regions are used to look for the actual peaks in the original ECG. After QRS-complex detection, the difference between R-R peaks was computed. In order to deal with arrhythmia-related artefacts, we excluded those R-R intervals not fitting: (i) 0.33 s < R-R interval < 1.5 s and (ii) difference to the previous R-R interval > 0.66 s [11]. No statistically significant differences (Mann-Whitney U test) were found between women and men in the percentage of R-R intervals discarded per subject or between SAHS-positive and SAHS-negative subjects in both groups. Before performing the spectral analysis, the HRV signals were resampled at 3.41 Hz by the use of linear interpolation [11]. This sample frequency was chosen as a trade-off between not to add a large amount of estimated data and take an efficient length for the posterior fast Fourier transform computation.

2.2. Analysis in Frequency Domain: Spectral Entropy

The power spectral density (PSD) of each resampled HRV recording was computed. We used the Welch's method since it is suitable for non-stationary signals [36]. A Hamming window of 2^{10} points (50% overlap), along with a discrete Fourier transform of 2^{11} points, were used to estimate the PSDs.

Then, each PSD was normalized (PSDn) by dividing the amplitude value at each frequency by the corresponding total power. In spite of some controversy [37], it is commonly accepted that LF is associated with sympathetic activity [8,11], *i.e.*, variations in the low-frequency PSD values from HRV reflect changes in the sympathetic nervous system. On the other hand, HF has been related to the respiratory rhythms and, therefore, to the parasympathetic activity [8,11]. The physiological interpretation of the very-low-frequency PSD values remains unclear [11], and it has been simply identified with long-period rhythms [38].

SE measurements were obtained from VLF (SE_{VLF}), LF (SE_{LF}), HF (SE_{HF}), and 0–0.4 Hz (SE_{VLF-HF}) bands. SE quantifies the uniformity, or flatness, of a PSD distribution [17,39]. Thus, a uniform (flat) spectrum, whose components are equally dispersed along frequencies, gives a high SE value (SE ≈ 1) [40]. This is the case of low predictability signals like the white noise [39,40]. Conversely, a condensed spectrum gives a low SE value (SE ≈ 0), which is the case of high predictability signals like sinusoids [39,40]. Thereby, for each band, higher SE values correspond to less predictability for the associated components in time domain. SE can be computed from the following expression [17]:

$$SE = -\sum_{f=f1}^{f2} \mathrm{PSDn}(f) \cdot \log[\mathrm{PSDn}(f)] \quad (1)$$

which is the application of Shannon's entropy to the normalized values of the PSD between the *f1* and *f*2 frequency limits [17].

2.3. Nonlinear Analysis in Time Domain: Multiscale Entropy

It is accepted that biological systems tend to non-linearity. As stated above, the heart rate is supposed to behave in this way [16]. In this regard, the MsE analysis applied to HRV has showed to be helpful in determining complexity patterns of several illnesses, as well as age [22,23]. MsE was originally developed by Costa et al. [41] on the basis of approximate entropy (ApEn) or sample entropy (SampEn). ApEn was designed by Pincus as an entropy measure which quantifies irregularity in time series [42]. Richman and Moorman improved ApEn by developing SampEn to reduce the bias caused by self-matching [43].

SampEn divides time-series into consecutive vectors of length *m*. It assesses whether the maximum absolute distance between the corresponding components of each pair of vectors is less than or equal to a tolerance *r*, *i.e.*, if the vectors match each other within *r*. If so, the vectors are considered as similar. The same process is repeated for vectors of length $m + 1$. Then, it is computed the conditional probability of similar vectors of length *m* remaining similar when the length is $m + 1$. The final SampEn value is obtained as the negative logarithm of such conditional probability [29,43]. Thus, higher values of SampEn indicate less self-similarity in the times-series and, consequently, more irregularity [29].

Our MsE analysis begins by applying SampEn to the original HRV series. This is the first scale. Scale 2 is computed by applying SampEn to a time-series whose values are the original HRV values averaged every two samples, without overlapping. In the same way, scale *k* is computed applying SampEn to time-series whose values are the original HRV values averaged every *k* samples without overlapping [41].

SampEn requires fitting a vector length, *m*, and a tolerance, *r*. We used $m = 3$ and $r = 0.2$ times the standard deviation of the time-series, as common choices in the study of HRV through SampEn [44].

Our HRV recordings have an average length of 29,000 points. A proper computation of SampEn requires at least 10^m points [43]. Hence, we chose 25 as a conservative number of scales to be analysed. The SampEn values of the 25 scales were taken as features ($SampEn_1$- $SampEn_{25}$).

2.4. Logistic Regression: Automatic Feature Selection and Classification

Logistic regression (LR) is a well-known supervised learning algorithm which estimates the posterior probability that a given instance $\mathbf{x}_i$ belongs to certain class C_k. This posterior probability, $p(C_k|\ \mathbf{x}_i)$, is computed through the logistic function:

$$p(C_k \mid \mathbf{x}_i) = \frac{e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}} \quad (2)$$

where β_0 and $\boldsymbol{\beta}$ are obtained by the weighted least squares minimization procedure [32]. Then, an instance $\mathbf{x}_i$ is assigned to the class with larger posterior probability. In our case, we have two classes: SAHS-positive and SAHS-negative. Input pattern $\mathbf{x}_i$ for each subject was composed of the feature values obtained for that subject after the SE and MsE analyses.

In this study LR was used with two purposes. First, to automatically select relevant and non-redundant features among those extracted from the SE and MsE analyses. This was performed through the forward-selection backward-elimination algorithm (FSBE), proposed by Hosmer and Lemeshow [32]. Then, LR was also used to assess the joint diagnostic ability of the features selected in the previous step.

2.5. Statistical Analysis

Features did not pass the Lilliefors normality test. Hence, the non-parametric Mann-Whitney U test was used to establish eventual statistically significant differences between SAHS-positive and SAHS-negative subjects (p-value < 0.01), both in women and men. The diagnostic ability of LR was assessed in terms of sensitivity (Se, percentage of SAHS-positive subjects rightly classified), specificity (Sp, percentage of SAHS-negative subjects rightly classified), accuracy (Acc, overall percentage of subjects rightly classified), positive predictive value (PPV, proportion of positive test results which are true positives), negative predictive value (NPV, proportion of negative test results which are true negatives), positive likelihood ratio (LR+, Se/(1-Sp)), and negative likelihood ratio (LR-, (1-Se)/Sp). The area under the receiver operating-characteristic curve was also computed (AROC). All the statistics were obtained after a leave-one-out cross-validation (loo-cv) procedure.

3. Results

Our results are divided into three sections. First, we compare the PSDn of SAHS-positive and SAHS-negative subjects, along with their corresponding SE_{VLF}, SE_{LF}, SE_{HF}, and $SE_{VLF\text{-}HF}$ mean values. Then, a similar analysis is conducted by using MsE curves and the features extracted from them: $SampEn_1$-$SampEn_{25}$. Thus, 29 features are obtained from the HRV of each subject. Finally, we compare LR models obtained to detect SAHS in women, men, and the whole set of subjects, which are built with an optimum subset out of the 29 features for each case.

3.1. Spectral Entropy

Figure 1 shows the averaged PSDn for the whole sets of SAHS-positive and SAHS-negative subjects (men and women jointly). VLF (0–0.04 Hz), LF (0.04–0.15 Hz), and HF (0.15–0.4 Hz) bands are also showed. Figure 2 shows the average PSDn values in women and men for the SAHS-positive and SAHS-negative groups. PSDn in both genders follow the same pattern, with SAHS-positive curves being qualitatively higher than the SAHS-negative ones from 0.015 to 0.060 Hz, approximately, *i.e.*, covering part of VLF and LF bands. Mean values of SE for the four bands considered are displayed in Table 2, separated by genders and SAHS class.

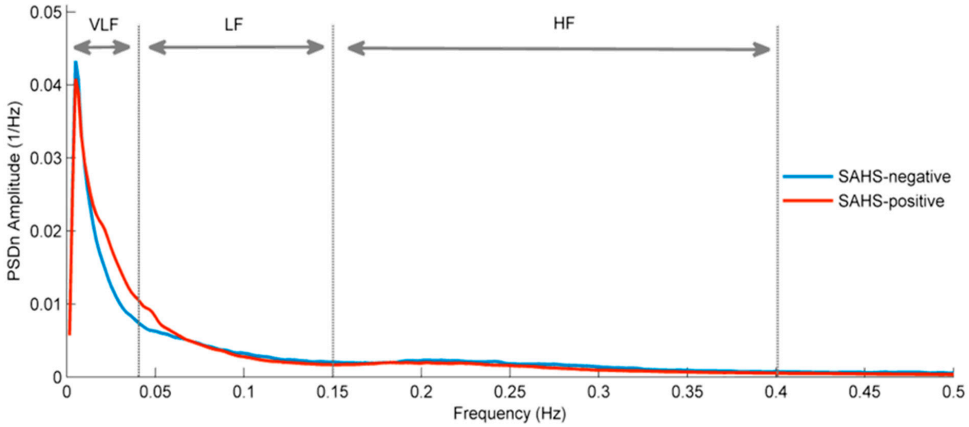

Figure 1. Normalized power spectral density of HRV in the whole set of SAHS-negative (blue) and SAHS-positive (red) subjects.

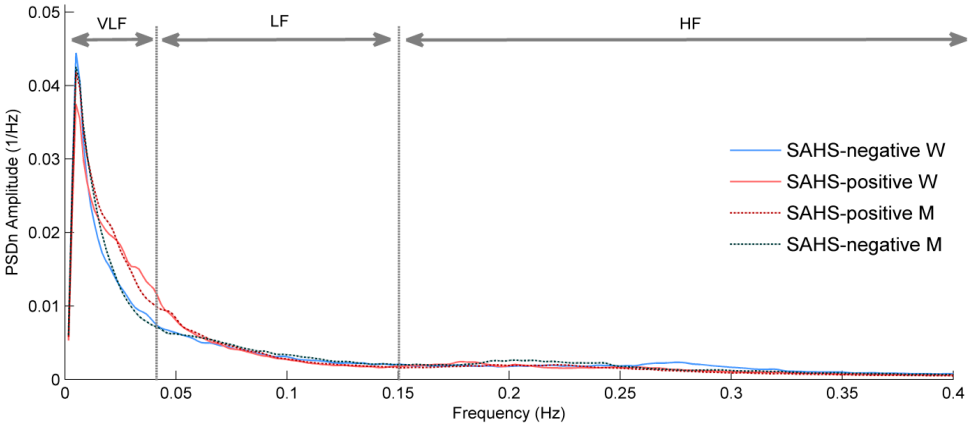

Figure 2. Normalized power spectral density of HRV in women (solid lines) and men (dashed lines) for the SAHS-positive and SAHS-negative groups.

Table 2. Spectral entropy features from VLF, LF, HF, and VLF-HF bands for women and men (mean ± standard deviation). *P*-values obtained from the Mann-Whitney U test.

	Women			Men		
	SAHS-Negative	**SAHS-Positive**	***p*-value**	**SAHS-Negative**	**SAHS-Positive**	***p*-value**
SE_{VLF}	0.959 ± 0.020	0.971 ± 0.011	**<0.01**	0.958 ± 0.020	0.966 ± 0.018	**<0.01**
SE_{LF}	0.984 ± 0.011	0.959 ± 0.028	**$<10^{-4}$**	0.983 ± 0.012	0.960 ± 0.035	**$<10^{-4}$**
SE_{HF}	0.979 ± 0.021	0.970 ± 0.022	0.158	0.983 ± 0.015	0.976 ± 0.023	0.219
SE_{VLF-HF}	0.899 ± 0.060	0.863 ± 0.051	<0.05	0.900 ± 0.053	0.873 ± 0.061	<0.05

Statistically significant differences between SAHS-positive and SAHS-negative subjects were found in the *SE* measures from VLF and LF of women and men (Mann-Whitney U test, *p*-value < 0.01). Specifically, SAHS-positive women and men present significantly higher SE_{VLF} and significantly lower SE_{LF}. On the other hand, no statistically significant differences were found in the SE of HF and the whole band. Finally, no statistically significant differences between women and men were found in any of the SE measures, either between SAHS-positive groups or between SAHS-negative ones.

3.2. Multiscale Entropy

Figure 3 displays averaged MsE curves for the 25 scales of SAHS-positive and SAHS-negative whole groups (men and women jointly). Figure 4 depicts the average MsE curves for SAHS-positive and SAHS-negative subjects, divided into women and men. It can be observed that SAHS increases SampEn values, *i.e.*, irregularity of HRV, for all the scales. A common behaviour is observed both in men and women, since the differences in low scales (1st to 7th) are lower than those in the high scales (from the 8th). This indicates higher differences in the degree of HRV irregularity of SAHS-positive subjects as more R-R intervals are averaged. However, only scale 13th reaches statistically significant differences between SAHS-positive and SAHS-negative men, whereas statistically significant differences are showed for scale 8th and from scale 12th onwards in women. No statistically significant differences were found between SAHS-negative women and men, and only scale 1 showed them in the case of the SAHS-positive groups.

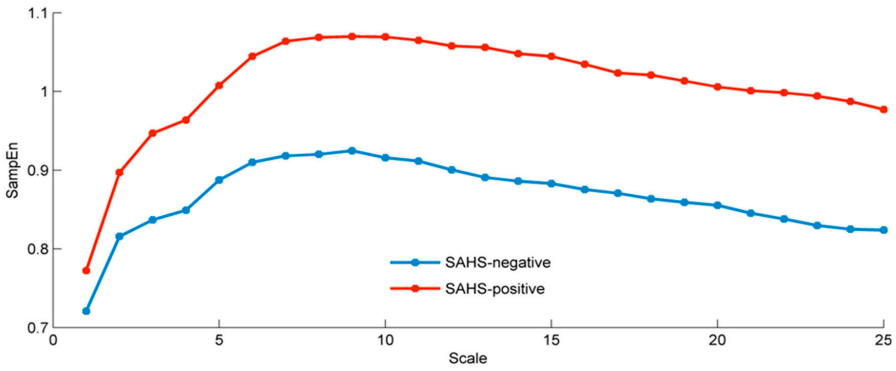

Figure 3. Averaged *MsE* curves for 25 scales of the SAHS-negative (blue) and SAHS-positive (red) whole groups.

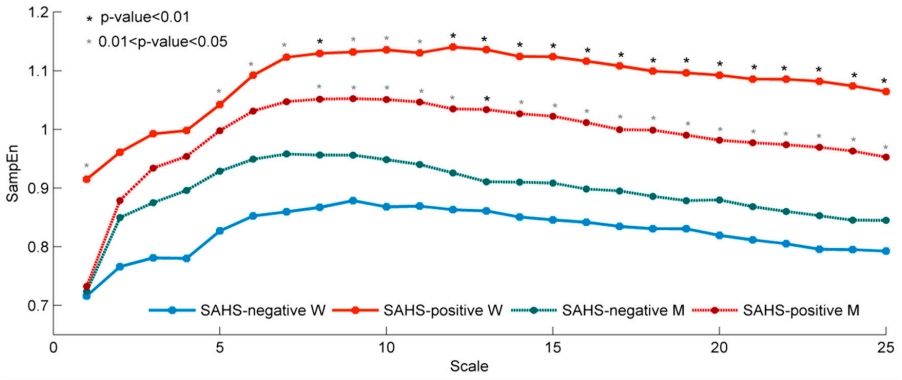

Figure 4. Average *MsE* curves of HRV in women (solid lines) and men (dashed lines) for the SAHS-positive and SAHS-negative groups. Black asterisks over SAHS-positive curves mark statistical significant differences (Mann-Whitney U test) with the corresponding SAHS-negatives (p-value < 0.01). Grey asterisks mark p-values in the range 0.01–0.05.

3.3. Feature Selection and Classification Results

The relevancy and complementarity of the features from SE and MsE analyses were assessed by the LR-based FSBE selection algorithm. In this regard, three experiments were conducted. First, only the features from women were included in the FSBE algorithm. Then, we did the same with the features from men. Finally, we applied the FSBE methodology to the 29 extracted features from the whole set of subjects. Table 3 shows the features selected in each case. SE and MsE features were selected in the three of them, showing that the information obtained from these analyses is complementary for SAHS detection. Moreover, SE_{VLF}, SE_{LF}, and $SampEn_2$ are common in the three optimum sets of features. Different tendencies can be observed for men and women in the remaining MsE features selected. For female subjects only scales below 8 were selected, whereas in the case of men the remaining features were selected from scales above 9.

Table 3. Features automatically selected by the FSBE algorithm for women, men, and the whole set of subjects. Common features for the three cases are in bold.

	Number of Features	Features Selected
Women	5	***SE*$_{VLF}$**, ***SE*$_{LF}$**, $SampEn_1$, ***SampEn*$_2$**, and $SampEn_7$
Men	12	***SE*$_{VLF}$**, ***SE*$_{LF}$**, $SE_{VLF\text{-}HF}$, ***SampEn*$_2$**, $SampEn_{10}$, $SampEn_{13}$, $SampEn_{16}$, $SampEn_{17}$, and $SampEn_{20}$- $SampEn_{23}$
All	15	***SE*$_{VLF}$**, ***SE*$_{LF}$**, $SE_{VLF\text{-}HF}$, ***SampEn*$_2$**, $SampEn_7$, $SampEn_9$, $SampEn_{11}$, $SampEn_{13}$, $SampEn_{14}$, $SampEn_{17}$,and $SampEn_{19}$- $SampEn_{23}$

Table 4. Diagnostic ability of the three LR models trained with the optimal features from women (LR_{FSBE-W}), men (LR_{FSBE-M}), and all subjects ($LR_{FSBE-All}$). Results were computed through a loo-cv procedure.

	Se(%)	Sp(%)	Acc(%)	PPV(%)	NPV(%)	LR+	LR-	AROC
LR_{FSBE-W}	80.8	89.3	85.2	87.5	83.3	7.6	0.215	0.951
LR_{FSBE-M}	87.1	56.1	77.6	81.8	65.7	1.98	0.230	0.895
$LR_{FSBE-All}$	79.8	59.4	72.3	77.2	63.1	1.97	0.340	0.885

The results of the diagnostic ability assessment of the corresponding LR models are shown in Table 4. All the statistics were computed after a loo-cv procedure. It can be observed that LR_{FSBE-W} and LR_{FSBE-M}, separately, achieve a significant higher performance than the LR model containing all the subjects. LR_{FSBE-W} achieves the highest overall performance.

4. Discussion and Conclusions

In this study, the ability of entropy measures to characterize SAHS in HRV recordings was evaluated. We also looked for eventual differences in these analyses depending on gender. Spectral entropy measurements in the VLF and LF bands characterized SAHS transversely, *i.e.*, regardless the gender. Thus, for both genders, SAHS decreased predictability of long-period rhythms in HRV (significantly higher SE_{VLF}). Also, it made more predictable the rhythms usually associated with sympathetic activity (significantly lower SE_{LF}). Spectral powers from VLF (P_{VLF}), LF (P_{LF}), and HF (P_{HF}), as well as the LF/HF spectral power ratio ($P_{LF/HF}$), have usually acted as parameters to characterize different physiological conditions in the HRV signal [8]. Table 5 shows the values of these conventional parameters reached by our SAHS-negative and SAHS-positive groups both in women and men (mean ± standard deviation). The corresponding *p*-values are also displayed (Mann-Whitney U test, *p*-value significance threshold = 0.01). No statistically significant differences were found between SAHS-negative and SAHS-positive women in any of the classical parameters and only P_{HF} showed a *p*-value < 0.01 in the case of men. Thus, SE showed higher performance when characterizing SAHS than the spectral powers obtained from the conventional VLF, LF, and HF bands. Nevertheless, a clear increase in the PSDn of SAHS-positive subjects can be observed in the range 0.015–0.060 Hz, covering part of VLF and LF bands (see Figure 3). It has been established that the typical duration of apnoeic events ranges from 20 to 40 s [45], which in the frequency domain would mainly affect the 0.025–0.050 Hz band. Latest studies confirmed this as the band with the highest statistically significant differences between SAHS-positive and SAHS-negative subjects in the airflow signal [46,47]. Additionally, a recent study has reported an increased cardio-respiratory coordination during the apnoeic events [48]. Hence, there exist strong indications that the observed changes covering part of the VLF and LF bands may be directly caused by these events, suggesting that further investigation is needed to find a specific spectral band related to SAHS in HRV. Finally, no differences were found in the SE from the HF band. This could be due to its relationship with normal breathing patterns [11], which are more predictable as well as predominant even in the presence of severe SAHS.

Table 5. Differences in the conventional spectral features of women and men (mean ± standard deviation). P_{VLF}: spectral power in the VLF band. P_{LF}: spectral power in the LF band. P_{HF}: spectral power in the HF band. $P_{LF/HF}$: P_{LF}/P_{HF} ratio.

	Women			**Men**		
	SAHS-Negative	**SAHS-Positive**	***p*-Value**	**SAHS-Negative**	**SAHS-Positive**	***p*-Value**
P_{VLF}	0.425 ± 0.153	0.489 ± 0.170	0.076	0.437 ± 0.167	0.503 ± 0.168	<0.05
P_{LF}	0.236 ± 0.052	0.241 ± 0.068	0.897	0.250 ± 0.051	0.250 ± 0.068	0.640
P_{HF}	0.234 ± 0.087	0.199 ± 0.118	0.058	0.228 ± 0.102	0.183 ± 0.108	**<0.01**
$P_{LF/HF}$	1.164 ± 0.514	1.639 ± 1.065	0.130	1.407 ± 0.874	1.898 ± 1.247	<0.05

MsE analysis also found differences in SAHS-positive and SAHS-negative subjects. Both in women and men, SAHS increased the average irregularity of HRV in the 25 time scales considered. Although it is usually accepted that the disease condition leads to a decrease of HRV irregularity [8], it has been also shown that sick sinus syndrome, characterized by bradycardia-tachycardia events, can increase the entropy measures [8]. As these bradycardia-tachycardia patterns are also recurrent after apnoeic events [9], this is consistent with the higher values of HRV irregularity showed in the MsE analysis of the SAHS-positive subjects. Moreover, in both genders, differences for the inferior scales are lower than those for the coarse-grained scales, indicating that SAHS affects more the long-term rhythms. This agrees with the differences found in the SE of VLF and LF. However, unlike the case of SE, we found different tendencies in men and women in the MsE curves. Mean values throughout the scales were higher for SAHS-positive and lower for SAHS-negative women than the corresponding for men. Thereby, 15 out of the 25 scales reached significant differences between SAHS-positive and SAHS-negative women, whereas only one did the same among men. In this regard, it has been previously reported that the R-R intervals from healthy women are significantly shorter and present less standard deviation than the corresponding from men [31]. This is reflected in HRV time series as lower mean amplitude and degree of variability. In this study the same tendency is observed since mean and standard deviation values in HRV are 0.886 ± 0.15 ms for SAHS-negative men and 0.878 ± 0.13 ms for SAHS-negative women. When comparing the two groups, in which normal heart behaviour is expected, the lower degree of variability may be one reason for the lesser mean values of entropy in women throughout the scales. In SAHS-positive subjects, the mean and standard deviation values in HRV behave in the same way (0.901 ± 0.12 ms for men and 0.888 ± 0.10 ms for women). Progressive bradycardia patterns are present in HRV from SAHS patients, *i.e.*, there are recurrent periods of increased HRV amplitude. Since the mean amplitude in women is lower, the difference to the increased HRV values may be higher than in men. This may be one reason for explaining the upper values of entropy, which tachycardia episodes would not be able to compensate because of their abrupt nature.

The FSBE algorithm showed the complementarity of SE and MsE analyses by automatically selecting features from both of them in the case of women, men, and the whole set of subjects. Additionally, since SE_{VLF} and SE_{LF} were common for the three sets of optimum features, it supported SE as a transversal characterizing of SAHS. It also supported the ability of MsE to distinguish gender specificities in HRV, since only scales below the 8th were selected in the case of women and eight out of the nine scales selected for men were above the 9th.

LR models were built with the three sets of selected features. These showed significantly higher performance when modelling SAHS by genders (85.2% Acc for $LR_{FSBE\text{-}W}$, 77.6% Acc for $LR_{FSBE\text{-}M}$) than when doing it in a jointly way (72.3% Acc for $LR_{FSBE\text{-}All}$). As a result, Acc increased 9.3% among women and 6.7% among men when comparing with the performance of the general model. Our feature selection methodology was optimized for an AHI threshold = 10 e/h. However, the outputs provided by the LR models can be also evaluated for other common thresholds. Thereby, for AHI = 5 (15) e/h, the Acc of $LR_{FSBE\text{-}W}$, $LR_{FSBE\text{-}M}$, and $LR_{FSBE\text{-}ALL}$ reaches 75.9% (79.6%), 76.1% (66.4%), and 72.3% (65.4%), respectively. Although 5 and 15 e/h are suboptimal thresholds for these models, the obtained results show the same general tendency as in the case of AHI = 10 e/h. Thus, it is suggested that SAHS may be more easily modelled from the SE and MsE analyses of HRV in the case of women.

Table 6 displays results from previous works focused on automatic SAHS classification. Data from studies involving SpO_2, airflow, snoring, respiratory effort, and HRV were included. In the case of the SpO_2 signal, Acc and AROC range from 84.1% to 95% and 0.822 to 0.967, respectively [49–52]. A database composed of 187 recordings was used to model a multi-layer perceptron (MLP) classifier, which was obtained from three non-linear features [49]. Six spectral (3) and non-linear (3) features were extracted from the same database to obtain four more classifiers by means of linear and quadratic discriminant analysis, *K*-nearest neighbours (KNN), and LR [50]. The best diagnostic ability for SpO_2 in terms of AROC (0.967) was achieved by a LR model obtained from four automatically-selected features extracted from the frequency and time domain of 147 recordings [51]. The best Acc (95.0%) was reported in the case of a support vector machine (SVM) classifier evaluated for a 5 e/h AHI threshold [52].

Up to nine features were extracted and analysed from the Hilbert transform of 41 oronasal airflow recordings [53]. The highest diagnostic ability was showed by the 25th frequency percentile of the Hilbert spectrum histogram (87.8% Acc and 0.877 AROC). However, these results were reached evaluating an AHI threshold = 5 e/h. Other recent study analysed linear and non-linear features from thermistor airflow [46]. A LR model obtained from three spectral features reached 82.4% Acc and 0.904 AROC after loo-cv (AHI threshold = 10 e/h). Moreover, five time and phase domain features from the abdominal and thoracic respiratory effort signals were used to feed a SVM classifier which reached 89.0% Acc, evaluated for AHI = 5 e/h [52]. Snoring sounds have been also assessed in the context of SAHS classification. A LR classifier modelled with nine spectral features reached 81.1% Acc and 0.850 AROC (AHI= 5 e/h threshold); and 86.5% Acc and 0.920 AROC (AHI = 5 e/h threshold) [54]. Another LR classifier obtained from 11 time and frequency domain features was evaluated for AHI = 10 e/h [55]. High diagnostic performance after loo-cv was reported (90.2% Acc and 0.967 AROC).

Table 6. Comparison with previous works focused on automatic classification of SAHS. MLP: multi-layer perceptron artificial neural network, LDA: linear discriminant analysis, QDA: quadratic discriminant analysis, KNN: *K*-nearest neighbours, LR: logistic regression, SVM: support vector machine, Loo: leave one out.

Study	Signal	#Subjects	Classifier	#Features	Validation	AHI Threshold	Se (%)	Sp (%)	Acc (%)	AROC
Roche *et al.* 2003 [56]	HRV	147	Tree	8	*k*-fold	10	64.2[+]	75.6[+]	69.3[+]	-
Marcos *et al.* 2008 [49]	SpO_2	187	MLP	3	Hold-out	10	89.8	79.4	85.5	0.900
Marcos *et al.* 2009 [50]	SpO_2	187	LDA	6	Hold-out	10	86.6	80.4	84.1	0.925
			QDA	6	Hold-out	10	91.1	78.3	85.8	0.913
			KNN	6	Hold-out	10	88.1	84.8	86.7	0.822
			LR	6	Hold-out	10	85.1	87.0	85.8	0.930
Caseiro *et al.* 2010 [53]	Airflow	41	Threshold	1	-	5	81.0	95.0	87.8[+]	0.877
Álvarez *et al.* 2010 [51]	SpO_2	148	LR	4	Loo	10	92.0	85.4	89.7	0.967
Fiz *et al.* 2010 [54]	Snoring	37	LR	9	-	5	87.0	71.4	81.1[+]	0.850
					-	15	80.0	90.9	86.5[+]	0.920
Karunajeewa *et al.* 2011 [55]	Snoring	41	LR	11	Loo	10	89.3	92.3	90.2[+]	0.967
Al-Angari *et al.* 2012 [52]	SpO_2	100	SVM	2	-	5	91.8	98.0	95.0	-
	Respiratory effort			5	-	5	85.7	92.2	89.0	-
	HRV			5	-	5	79.6	78.4	79.0	-
Gutiérrez-Tobal *et al.* 2012 [46]	Airflow	148	LR	3	Loo	10	88.0	70.8	82.4	0.903
Ravelo-Garcia *et al.* 2014 [57]	HRV	97	LR	5	*k*-fold	10	88.7	82.9	86.6[+]	0.941
This study (LR_{FSBE-W})	HRV	54	LR	5	Loo	10	80.8	89.3	85.2	0.951
This study (LR_{FSBE-M})	HRV	134	LR	13	Loo	10	87.1	56.1	77.6	0.895
This study ($LR_{FSBE-All}$)	HRV	188	LR	15	Loo	10	79.8	59.4	72.3	0.885

[+] Computed from reported data.

Decision trees, SVM, and LR classifiers have been also used to model SAHS from HRV features. Eight features from wavelet analysis were used to build a decision tree, reaching 69.3% Acc [56]. A SVM model was obtained from five time and frequency domain features [52]. Authors reported 79.0% Acc when evaluating the classifier for AHI = 5 e/h. A recent study reported 86.6% Acc and 0.941 AROC for a LR classifier modelling with four clinical variables and one symbolic dynamic feature extracted from HRV [57]. Finally, several works have reported 100% Acc when classifying 30 subjects from the PhysioNet Apnea-ECG database, which was used in the Computers in Cardiology Challenge 2000 [58]. However, comparison with studies using this database is difficult since borderline subjects were deliberately removed from the competition. Additionally, only one woman was included in the apnea group [59].

Our LR models achieved high diagnostic ability comparing with those studies involving HRV. Only the Acc reported in the study from Ravelo-García *et al.* outperformed the Acc reached by our LR_{FSBE-W} model. However, our AROC was slightly higher and we did not include clinical variables in the

modelling process. Moreover, only the results reported in the studies conducted by Álvarez *et al.* and Karunajeewa *et al.* outperformed our $LR_{FSBE\text{-}W}$ classifier both in Acc and AROC. Nonetheless, the latter does not meet the subject:feature ratio criterion which avoids bias in the logistic regression coefficients [60]. These results suggest that our proposal could be helpful to detect SAHS, especially in the case of women. Actually, none of the above mentioned studies addressed the problem of evaluating gender differences when modelling SAHS.

Some limitations need to be pointed out in this study. First, an increased number of subjects would provide our results with higher statistical power. Additionally, a larger number of subjects would let us define the optimum sets of features from an independent database. Nonetheless, we validated their performance by the use of leave-one-out cross-validation. Increasing the SAHS-negative men would be particularly helpful since there exists a clear unbalance with SAHS-positive ones. However, the prevalence of SAHS in our database reflects a realistic proportion of SAHS patients among the subjects which undergo PSG [7,61]. Another limitation is the complexity inherent to the acquisition of surface ECG as a previous step to obtain HRV. Although acquiring ECG is significantly easier than recording the whole set of signals involved during PSG, there exist studies which address the obtaining of HRV from simpler devices such as the oximeter [62,63]. Moreover, regarding the R-R time series, no specific correction for the timing of R waves associated with ectopic beats has been applied in this study. On the other hand, there are some factors with ability to change HRV dynamics. We did not address issues like tobacco and alcohol consumption, as well as differences in the fitness of the subjects. These could be the object of interesting future research. Another future goal is to carry out an exhaustive analysis of the optimal values of m and r for the computation of SampEn from HRV in the context of SAHS. Although the values used in this study have shown their usefulness in the characterization of a range of physiological conditions, we did not test other values which could enhance the characterization of SAHS in our specific database. Regarding the spectral and nonlinear analysis, other features extracted from HRV could complement our study and increase the diagnostic ability of our methodology. Finding a single feature with ability to gather the complementary information of those interdependent features present in our study would be of great interest. In addition, the use of more complex classifiers could also improve the results achieved. Finally, even though no statistically significant difference was found between the age of SAHS-positive and SAHS-negative groups (either women or men), higher average values are present in those affected by SAHS. In this regard, no statistically significant correlation was found between any of the features used in the LR models and age (all the absolute Spearman's correlation coefficients were lower than 0.18 and the corresponding p-values higher than 0.11).

In summary, we showed that SE and MsE analyses of HRV can be used to help in SAHS detection. The complementarity of the two of them was also exposed. The ability of MsE to distinguish gender specificities in HRV was suggested too. Higher diagnostic ability was reached when modelling SAHS from entropy measures of women and men separately. A LR model built with five SE and MsE features from women achieved the highest performance in SAHS detection (85.2% Acc, 0.951 AROC for an AHI threshold = 10 e/h). This suggests that SAHS may be more easily modelled from HRV in the case of women. Our results show the utility of the SE and MsE analyses to help in SAHS detection, as well as indicate that, when using HRV, SAHS may be more accurately modelled if data are separated by gender.

Acknowledgments

This research was supported by project TEC2011-22987 from Ministerio de Economía y Competitividad and FEDER, the Proyecto Cero 2011 on Ageing from Obra Social La Caixa, Fundación General CSIC and CSIC, and the project VA059U13 from the Consejería de Educación de la Junta de Castilla y León. G. C. Gutiérrez-Tobal was in receipt of a PIRTU grant from the Consejería de Educación de la Junta de Castilla y León and the European Social Fund.

Author Contributions

Gonzalo C. Gutiérrez-Tobal designed the study, analysed the data, interpreted the results, and drafted the manuscript. Daniel Álvarez and Roberto Hornero designed the study, analysed the data and interpreted the results. Javier Gomez-Pilar took part in the collection of data, analysed the data and interpreted the results. Félix del Campo took part in the diagnosis of subjects and the collection of data, and interpreted the results. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Lopez-Jiménez, F.; Kuniyoshi, F.H.S.; Gami, A.; Somers, V.K. Obstructive Sleep Apnea. *Chest* **2008**, *133*, 793–804.
2. Patil, S.P.; Schneider, H.; Schwartz, A.R.; Smith, P.L. Adult obstructive sleep apnea: Pathophysiology and diagnosis. *Chest* **2007**, *132*, 325–337.
3. Epstein, L.J.; Kristo, D.; Strollo, P.J.; Friedman, N.; Malhotra, A.; Patil, S.P.; Ramar, K.; Rogers, R.; Schwab, R.J.; Weaver, E.M.; *et al*. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. *J. Clin. Sleep Med.* **2009**, *5*, 263–276.
4. Bennett, J.A.; Kinnear, W.J. M. Sleep on the cheap: The role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome. *Thorax* **1999**, *54*, 958–959.
5. Iber, C.; Ancoli-Israel, S.; Chesson, A.L.; Quan S.F. *The AASM Manual for the Scoring of Sleep and Associated Events*; American Academy of Sleep Medicine: Westchester, IL, USA, 2007.
6. Flemons, W.W.; Littner, M.R.; Rowley, J.A.; Gay, P.; Anderson, W.M.; Hudgel, D.W.; McEvoy, R.D.; Loube, D.I. Home diagnosis of sleep apnea: A systematic review of the literature. *Chest* **2003**, *124*, 1543–1579.
7. Flemons, W.W.; Douglas, N.J.; Kuna, S.T.; Rodenstein, D.O.; Wheatley, J. Access to diagnosis and treatment of patients with suspected sleep apnea. *Am. J. Respir. Crit. Care Med.* **2004**, *169*, 668–672.
8. Acharya, U.R.; Joseph, K.P.; Kannathal, N.; Lim, C.M.; Suri, J.S. Heart rate variability: A review. *Med. Biol. Eng. Comput.* **2006**, *44*, 1031–1051.
9. Guilleminault, C.; Winkle, R.; Connolly, S.; Melvin, K.; Tilkian, A. Cyclical variation of the heart rate in sleep apnoea syndrome: Mechanisms and usefulness of 24 h electrocardiography as a screening technique. *Lancet* **1984**, *323*, 126–131.

10. Bonsignore, M.R.; Romano, S.; Marrone, O.; Chiodi, M.; Bonsignore, G. Different heart rate patterns in obstructive apneas during NREM sleep. *Sleep* **1997**, *20*, 1167–1174.
11. Penzel, T.; Kantelhardt, J.W.; Grote, L.; Peter, J.H.; Bunde, A. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *IEEE Trans. Biomed. Eng.* **2003**, *50*, 1143–1151.
12. Gula, L.J.; Krahn, A.D.; Skanes, A.; Ferguson, K.A.; George, C.; Yee, R.; Klein, G.J. Heart rate variability in obstructive sleep apnea: A prospective study and frequency domain analysis. *Ann. Noninvasive Electrocardiol.* **2003**, *8*, 144–149.
13. Penzel, T.; Wessel, N.; Riedl, M.; Kantelhardt, J.W.; Rostig, S.; Glos, M.; Suhrbier, A.; Malberg, H.; Fietze, I. Cardiovascular and respiratory dynamics during normal and pathological sleep. *Chaos* **2007**, *17*, 015116.
14. Gapelyuk, A.; Riedl, M.; Suhrbier, A.; Kraemer, J.F.; Bretthauer, G.; Malberg, H.; Kurths, J.; Penzel, T.; Wessel, N. Cardiovascular regulation in different sleep stages in the obstructive sleep apnea syndrome. *Biomed. Technik. (Biomed. Eng.)* **2011**, *56*, 207–213.
15. Goldberger, A.L. Is the normal heartbeat chaotic or homeostatic? *News Physiol. Sci.* **1991**, *6*, 87–91.
16. Wessel, N.; Riedl, M.; Kurths, J. Is the normal heart rate "chaotic" due to respiration? *Chaos* **2009**, *19*, 028508.
17. Poza, J.; Hornero, R.; Abásolo, D.; Fernández, A.; García, M. Extraction of spectral based measures from MEG background oscillations in Alzheimer's disease. *Med. Eng. Phys.* **2007**, *29*, 1073–1083.
18. Chang, Y.C.; Wu, H.T.; Chen, H.R.; Liu, A.B.; Yeh, J.J.; Lo, M.T.; Tsao, J.H.; Tang, C.-J.; Tsai, I.-T.; Sun, C.-K. Application of a Modified Entropy Computational Method in Assessing the Complexity of Pulse Wave Velocity Signals in Healthy and Diabetic Subjects. *Entropy* **2014**, *16*, 4032–4043.
19. Alcaraz, R.; Rieta, J.J. Sample entropy of the main atrial wave predicts spontaneous termination of paroxysmal atrial fibrillation. *Med. Eng. Phys.* **2009**, *31*, 917–922.
20. Hornero, R.; Álvarez, D.; Abásolo, D.; del Campo, F.; Zamarrón, C. Utility of approximate entropy from overnight pulse oximetry data in the diagnosis of the obstructive sleep apnea syndrome. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 107–113.
21. Al-Angari, H.M.; Sahakian, A.V. Use of sample entropy approach to study heart rate variability in obstructive sleep apnea syndrome. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 1900–1904.
22. Costa, M.D.; Peng, C.K.; Goldberger, A.L. Multiscale analysis of heart rate dynamics: Entropy and time irreversibility measures. *Cardiovasc. Eng.* **2008**, *8*, 88–93.
23. Costa, M.D.; Goldberger, A.L.; Peng, C.K. Multiscale entropy analysis of biological signals. *Phys. Rev. E* **2005**, *71*, 021906.
24. Costa, M.D.; Peng, C.K.; Goldberger, A.L.; Hausdorff, J.M. Multiscale entropy analysis of human gait dynamics. *Physica A* **2003**, *330*, 53–60.
25. Escudero, J.; Abásolo, D.; Hornero, R.; Espino, P.; López, M. Analysis of electroencephalograms in Alzheimer's disease patients with multiscale entropy. *Physiol. Meas.* **2006**, *27*, 1091–1106.
26. Takahashi, T.; Cho, R.Y.; Mizuno, T.; Kikuchi, M.; Murata, T.; Takahashi, K.; Wada, Y. Antipsychotics reverse abnormal EEG complexity in drug-naive schizophrenia: A multiscale entropy analysis. *Neuroimage* **2010**, *51*, 173–182.

27. Hans, P.; Dewandre, P.Y.; Brichant, J.F.; Bonhomme, V. Comparative effects of ketamine on Bispectral Index and spectral entropy of the electroencephalogram under sevoflurane anaesthesia. *Br. J. Anaesth.* **2005**, *94*, 336–340.
28. Shen, J.L.; Hung, J.W.; Lee, L.S. Robust entropy-based endpoint detection for speech recognition in noisy environments. *ICSLP* **1998**, *98*, 232–235.
29. Abásolo, D.; Hornero, R.; Espino, P.; Álvarez, D.; Poza, J. Entropy analysis of the EEG background activity in Alzheimer's disease patients. *Physiol. Meas.* **2006**, *27*, 241–253.
30. Alvarez, D.; Hornero, R.; Marcos, J.V.; Wessel, N.; Penzel, T.; Glos, M.; del Campo, F. Assessment of Feature Selection and Classification Approaches to Enhance Information from Overnight Oximetry in the Context of Apnea Diagnosis. *Int. J. Neural Syst.* **2013**, *23*, 1–18.
31. Bonnemeier, H.; Wiegand, U.K.; Brandes, A.; Kluge, N.; Katus, H.A.; Richardt, G.; Potratz, J. Circadian profile of cardiac autonomic nervous modulation in healthy subjects. *J. Cardiovasc. Electrophysiol.* **2003**, *14*, 791–799.
32. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*; John Wiley & Sons: London, UK, 1999.
33. Berry, R.B.; Budhiraja, R.; Gottlieb, D.J.; Gozal, D.; Iber, C.; Kapur, V.K.; Marcus, C.L.; Mehra, R.; Parthasarathy, S.; Quan, S.F.; *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.* **2012**, *8*, 597–619.
34. Baselli, G.; Cerutti, S.; Civardi, S.; Lombardi, F.; Malliani, A.; Merri, M; Pagani, M.; Rizzo, G. Heart rate variability signal processing: A quantitative approach as an aid to diagnosis in cardiovascular pathologies. *Int. J. Biol. Med. Comput.* **1987**, *20*, 51–70.
35. Benitez, D.; Gaydecki, P.A.; Zaidi, A.; Fitzpatrick, A.P. The use of the Hilbert transform in ECG signal analysis. *Comput. Biol. Med.* **2001**, *31*, 399–406.
36. Welch, P.D. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **1967**, *15*, 70–73.
37. Reyes del Paso, G.A.; Langewitz, W.; Mulder, L.J.; Roon, A.; Duschek, S. The utility of low frequency heart rate variability as an index of sympathetic cardiac tone: A review with emphasis on a reanalysis of previous studies. *Psychophysiology* **2013**, *50*, 477–487.
38. Sztajzel, J. Heart rate variability: A noninvasive electrocardiographic method to measure the autonomic nervous system. *Swiss Med. Wkly.* **2004**, *134*, 514–522.
39. Inouye, T.; Shinosaki, K.; Sakamoto, H.; Toi, S.; Ukai, S.; Iyama, A.; Katsuda, Y.; Hirano, M. Quantification of EEG irregularity by use of the entropy of the power spectrum. *Electroencephalogr. Clin. Neurophysiol.* **1991**, *79*, 204–210.
40. Sleigh, J.W.; Steyn-Ross, D.A.; Steyn-Ross; M.L.; Grant, C.; Ludbrook, G. Cortical entropy changes with general anaesthesia: Theory and experiment. *Physiol. Meas.* **2004**, *25*, 921–934.
41. Costa, M.; Goldberger, A.L.; Peng, C.K. Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* **2002**, *89*, 068102.
42. Pincus, S.M. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301.

43. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **1999**, *278*, H2039–H2049.
44. Alcaraz, R.; Rieta, J.J. A review on sample entropy applications for the non-invasive analysis of atrial fibrillation electrocardiograms. *Biomed. Signal. Process. Control.* **2010**, *5*, 1–14.
45. Eckert, D.J.; Malhotra, A. Pathophysiology of adult obstructive sleep apnea. *Proc. Am. Thoracic Soc.* **2008**, *5*, 144–153.
46. Gutiérrez-Tobal, G.C.; Hornero, R.; Álvarez, D.; Marcos, J.V.; del Campo, F. Linear and nonlinear analysis of airflow recordings to help in sleep apnoea-hypopnoea syndrome diagnosis. *Physiol. Meas.* **2012**, *33*, 1261–1275.
47. Gutiérrez-Tobal, G.C.; Álvarez, D.; Marcos, J.V.; del Campo, F.; Hornero, R. Pattern recognition in airflow recordings to assist in the sleep apnoea–hypopnoea syndrome diagnosis. *Med. Biol. Eng. Comput.* **2013**, *51*, 1367–1380.
48. Riedl, M.; Müller, A.; Kraemer, J.F.; Penzel, T.; Kurths, J.; Wessel, N. Cardio-Respiratory Coordination Increases during Sleep Apnea. *PLoS One* **2014**, *9*, e93866.
49. Marcos, J.V.; Hornero, R.; Álvarez, D.; del Campo, F.; Zamarrón, C.; López, M. Utility of multilayer perceptron neural network classifiers in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry. *Comput. Methods Progr. Biomed.* **2008**, *92*, 79–89.
50. Marcos, J.V.; Hornero, R.; Álvarez, D.; del Campo, F.; Zamarrón, C. Assessment of four statistical pattern recognition techniques to assist in obstructive sleep apnoea diagnosis from nocturnal oximetry. *Med. Eng. Phys.* **2009**, *31*, 971–978.
51. Alvarez, D.; Hornero, R.; Marcos, J.V.; del Campo, F. Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. *IEEE Trans. Biomed. Eng.* **2010**, *57*, 2816–2824.
52. Al-Angari, H.M.; Sahakian, A.V. Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier. *IEEE Trans. Inf. Technol. Biomed.* **2012**, *16*, 463–468.
53. Caseiro, P.; Fonseca-Pinto, R.; Andrade, A. Screening of obstructive sleep apnea using Hilbert-Huang decomposition of oronasal airway pressure recordings. *Med. Eng. Phys.* **2010**, *32*, 561–568.
54. Fiz, J.A.; Jane, R.; Solà-Soler, J.; Abad, J.; García, M.; Morera, J. Continuous analysis and monitoring of snores and their relationship to the apnea-hypopnea index. *Laryngoscope* **2010**, *120*, 854–862.
55. Karunajeewa, A.S.; Abeyratne, U.R.; Hukins, C. Multi-feature snore sound analysis in obstructive sleep apnea-hypopnea syndrome. *Physiol. Meas.* **2011**, *32*, doi:10.1088/0967-3334/32/1/006.
56. Roche, F.; Pichot, V.; Sforza, E.; Duverney, D.; Costes, F.; Garet, M.; Barthélémy, J.C. Predicting sleep apnoea syndrome from heart period: A time-frequency wavelet analysis. *Eur. Respir. J.* **2003**, *22*, 937–942.
57. Ravelo-García, A.G.; Saavedra-Santana, P.; Juliá-Serdá, G.; Navarro-Mesa, J.L.; Navarro-Esteva, J.; Álvarez-López, X.; Gapelyuk, A.; Penzel, T.; Wessel, N. Symbolic dynamics marker of heart rate variability combined with clinical variables enhance obstructive sleep apnea screening. *Chaos* **2006**, *24*, 024404.

58. Penzel, T.; McNames, J.; de Chazal, P.; Raymond, B.; Murray, A.; Moody, G. Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. *Med. Biol. Eng. Comput.* **2002**, *40*, 402–407.
59. Penzel, T.; Moody, G.B.; Mark, R.G.; Goldberger, A.L.; Peter, J.H. The apnea-ECG database. In *Computers in Cardiology 2000*, Proceedings of Conference in Computers in Cardiology, Cambridge, MA, USA, 24–27 September 2000; pp. 255–258.
60. Peduzzi, P.; Concato, J.; Kemper, E.; Holford, T.R.; Feinstein, A.R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **1996**, *49*, 1373–1379.
61. Kapsimalis, F.; Kryger, M.H. Gender and obstructive sleep apnea syndrome, part 1: Clinical features. *Sleep* **2002**, *25*, 412–419.
62. Constant, I.; Laude, D.; Murat, I.; Elghozi, J.L. Pulse rate variability is not a surrogate for heart rate variability. *Clin. Sci.* **1999**, *97*, 391–397.
63. Gil, E.; Orini, M.; Bailón, R.; Vergara, J.M.; Mainardi, L.; Laguna, P. Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions. *Physiol. Meas.* **2010**, *31*, 1271.

Biomedical Signal Processing and Control 18 (2015) 401–407

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Diagnosis of pediatric obstructive sleep apnea: Preliminary findings using automatic analysis of airflow and oximetry recordings obtained at patients' home

Gonzalo C. Gutiérrez-Tobal[a,*], M. Luz Alonso-Álvarez[b], Daniel Álvarez[a], Félix del Campo[c,d], Joaquín Terán-Santos[b], Roberto Hornero[a]

[a] Biomedical Engineering Group, E.T.S.I. de Telecomunicación, Universidad de Valladolid, Valladolid, Spain
[b] Unidad Multidisciplinar de Sueño, CIBER respiratorio, Hospital Universitario de Burgos, Burgos, Spain
[c] Facultad de Medicina, Universidad de Valladolid, Valladolid, Spain
[d] Hospital Universitario Río Hortega, Valladolid, Spain

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form
29 December 2014
Accepted 25 February 2015

Keywords:
Pediatric obstructive sleep apnea
Airflow
Oximetry
Spectral analysis
At-home assessment

ABSTRACT

The obstructive sleep apnea syndrome (OSAS) greatly affects both the health and the quality of life of children. Therefore, an early diagnosis is crucial to avoid their severe consequences. However, the standard diagnostic test (polysomnography, PSG) is time-demanding, complex, and costly. We aim at assessing a new methodology for the pediatric OSAS diagnosis to reduce these drawbacks. Airflow (AF) and oxygen saturation (SpO_2) at-home recordings from 50 children were automatically processed. Information from the spectrum of AF was evaluated, as well as combined with 3% oxygen desaturation index (ODI3) through a logistic regression model. A bootstrap methodology was conducted to validate the results. OSAS significantly increased the spectral content of AF at two abnormal frequency bands below (BW1) and above (BW2) the normal respiratory range. These novel bands are consistent with the occurrence of apneic events and the posterior respiratory overexertion, respectively. The spectral information from BW1 and BW2 showed complementarity both between them and with ODI3. A logistic regression model built with 3 AF spectral features (2 from BW1 and 1 from BW2) and ODI3 achieved (mean and 95% confidence interval): 85.9% sensitivity [64.5–98.7]; 87.4% specificity [70.2–98.6]; 86.3% accuracy [74.9–95.4]; 0.947 area under the receiver-operating characteristics curve [0.826–1]; 88.4% positive predictive value [72.3–98.5]; and 85.8% negative predictive value [65.8–98.5]. The combination of the spectral information from two novel AF bands with the ODI3 from SpO_2 is useful for the diagnosis of OSAS in children.

1. Introduction

Obstructive sleep apnea syndrome (OSAS) is a disorder characterized by recurrent episodes of apnea (complete absence of airflow) and hypopnea (significant reduction of airflow) during sleep [1]. Apneic events lead to oxygen desaturations and arousals which prevent patients from resting while sleeping, disrupting both their health and quality of life. OSAS can affect both adults and children. Common symptoms in children include overnight snoring and sleep difficulties [2], which may derive in other daytime symptoms and illnesses such as cognitive and behavioral irregularities, abnormal growth, and cardiovascular risks [3,4]. Moreover, pediatric OSAS is known to be underdiagnosed [5], and the scientific literature reports up to 6% of children affected [3]. This indicates the high prevalence of the disease which, in turn, leads to an intensive use of the healthcare services [6].

Abbreviations: Acc, accuracy; AF, airflow; AHI, apnea-hypopnea index; AROC, area under the receiver-operating characteristics curve; ECG, electrocardiogram; IQR, interquartile range; LR, logistic regression; *MA*, maximum amplitude of the power spectral density; *mA*, minimum amplitude of the power spectral density; M_{f1}-M_{f4}, first to fourth statistical moments of the power spectral density; NPV, negative predictive value; ODI, oxygen desaturation index; OSAS, obstructive sleep apnea syndrome; PPV, positive predictive value; PSD, power spectral density; PSG, polysomnography; RP, respiratory polygraphy; Se, sensitivity; SLR, stepwise logistic regression; Sp, specificity; SpO_2, oxygen saturation of the blood.

* Corresponding author at: Departamento de Teoría de la Señal y Comunicaciones e Ingeniería Telemática, Universidad de Valladolid, Paseo Belén, 15, 47011 Valladolid, Spain. Tel.: +34 983423000x4716.
E-mail address: gonzalo.gutierrez@gib.tel.uva.es (G.C. Gutiérrez-Tobal).

http://dx.doi.org/10.1016/j.bspc.2015.02.014
1746-8094/

OSAS in children is diagnosed by means of nocturnal polysomnography (PSG) test, which acts as the "gold standard" [2]. PSG requires recording a wide range of physiological signals from patients overnight, including electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), electrooculogram (EOG), thoracic and abdominal respiration movements, oxygen saturation (SpO_2), and airflow (AF) [1]. Hence, the necessary acquisition equipment is both complex and costly [6]. OSAS diagnosis is established according to the apnea-hypopnea index (AHI), which estimates the number of apneic events per hour of sleep time. To derive AHI, the physiological recordings need to be examined. Consequently, PSG is also time-consuming [7]. Furthermore, the equipment involved in PSG is often not well tolerated by children [8], interfering with their sleep routine.

To overcome these drawbacks a number of alternatives have been studied. One common approach is the use of a reduced set of signals from PSG to compute different estimations of AHI. In this regard, the respiratory disturbance index obtained from respiratory polygraphy (RP) was successfully assessed in an in-lab study with children involving 6 signals [9]: SpO_2, AF, heart rate, chest movements, body position, and snoring. The oxygen desaturation index (ODI), in combination with common symptoms, has been also recently evaluated as an alternative to PSG in pediatric patients [10]. On the other hand, the automatic analysis of physiological signals has been also proposed. In this sense, features from photoplethysmography time series have shown their usefulness in OSAS detection in children [11]. Moreover, studies conducting an automatic processing of the SpO_2 and ECG signals have been successfully performed in the context of adult and pediatric OSAS [12–16].

In this paper, a new method for OSAS diagnosis in children is assessed. Our methodology is based on the only use of spectral data from single-channel AF and the 3% ODI (ODI3), both of them obtained at patient's home. The main objective is to evaluate the diagnostic usefulness of eventual differences in the AF spectrum of OSAS patients (OSAS-positive) and no-OSAS subjects (OSAS-negative) in combination with ODI3. As stated above, ODI3 is a commonly used parameter in OSAS studies. Moreover, the study of AF is a straightforward choice since apneas and hypopneas are defined on the basis of its amplitude variations [17]. Additionally, the recurrence of apneic events naturally leads to the study of AF in the frequency domain. Recent works have shown that OSAS modifies the spectral content of AF recordings from adults at certain frequencies, and that the information contained in such frequencies is useful in OSAS detection [18,19]. However, no studies have been found applying a similar analysis to AF recordings from children. According to the above mentioned, we pose the following research questions:

i. How does OSAS modify the spectral information of airflow recordings from children?
ii. Are these changes useful to distinguish OSAS in children from at-home recordings?
iii. Is the airflow spectral information complementary to the classic oxygen desaturation index in pediatric OSAS detection?

To answer them, we conduct an exploratory analysis of the power spectral density (PSD) of the AF recordings. We look for spectral bands of interest showing differences in OSAS-positive and OSAS-negative subjects, as well as their characterization. The single diagnostic performance of both the AF spectral features and the ODI3 are assessed. We also evaluate their usefulness and complementarity through logistic regression models. Our hypothesis is that the joint use of spectral information contained in single-channel AF and ODI3 could be useful to diagnose OSAS in children.

Table 1
Demographic and clinical data.

Features	All	OSAS-positive	OSAS-negative
# Subjects	50	26	24
Age* (years)	5.3 ± 2.5	5.4 ± 2.7	5.2 ± 2.4
Male (%)	54.0	61.5	45.8
BMI+ (kg/m^2)	16.5 ± 2.5	16.9 ± 3.0	16.1 ± 1.7
Recording Time (h)	8.9 ± 0.8	8.8 ± 1.0	9.0 ± 0.5
AHI (e/h)	9.9 ± 13.8	17.9 ± 15.4	1.3 ± 0.8

BMI: body mass index; AHI: apnea hypopnea index.
* p-value = 0.76.
+ p-value = 0.94.

2. Methods and materials

2.1. Subjects and signals under study

This study involved AF and SpO_2 recordings from 50 children ranging 3–13 years old (24 OSAS-negative and 26 OSAS-positive). All of them were referred to the unit of respiratory sleep disorders of the University Hospital of Burgos (Spain), due to clinical suspicion of OSAS (snoring and/or witnessed breathing pauses). Those children suffering from serious chronic medical or psychiatric co-morbidities, those who required urgent treatment, and those with symptoms suggestive of sleep disorders other than OSAS (e.g., parasomnias, narcolepsy, or periodic leg movements), were excluded. AF and SpO_2 were acquired during a polygraphy test performed at patients' home through an eXim Apnea polygraph (Bitmed®, Sibel S.A., Barcelona, Spain). The sensor used to obtain AF was a thermistor and the sample rate was 100 Hz. SpO_2 was recorded through an oximeter at the same sample rate. The physicians used the AHI derived from PSG to establish OSAS. For the overnight PSG, the Deltamed Coherence® 3NT Polysomnograph, version 3.0 system (Diagniscan, S.A. ACH – Werfen Company; Paris, France) was used, recording EEG, right and left EOG, tibial and submental (leg and chin) EMG, ECG, AF by thermistor and nasal cannula, chest-abdomen movements with bands, body position, SpO_2 (Nellcor Puritan Bennett – NPB-290®), snoring, and a continuous transcutaneous recording of carbon dioxide ($PtcCO_2$). The American Academy of Sleep Medicine (AASM) criteria were used to evaluate sleep states and respiratory events [17]. The median time between PSG and RP was 14 days ([6,25], interquartile range, IQR). Apneas were scored after complete cessation of AF, as defined by the American Academy of Sleep Medicine [17]. Hypopneas were defined after a 50% reduction of AF accompanied by a 3% decrease in SpO_2 [17]. Amplitude cessations and reductions of AF required lasting 2 missed cycles in order to be considered as apneas or hypopneas, respectively [17]. An obstructive AHI threshold of 3 events/h was used to distinguish OSAS-positive from OSAS-negative subjects [20]. ODI3 was estimated as the number of desaturations (at least 3%) per hour of recording. The interruption of the oronasal flow secondary to movements was not accounted for either the PSG or the RP. An uninterpretable AF signal was defined as no AF during 30 s of normal respiration, while respiratory motion signals and SpO_2 remained unchanged. Data were excluded from analysis if >60% of the AF was uninterpretable. The Ethics Committee of the University Hospital of Burgos accepted the protocol (approval #CEIC 936) and an informed consent was obtained for each subject. Table 1 summarizes clinical and demographical data from the subjects under study. No statistical significant differences in body mass index or age were found between groups (p-value $\gg 0.01$).

2.2. Power spectral density of airflow

We computed the PSD of each AF recording to explore eventual differences between the spectral information of OSAS-positive and

OSAS-negative groups. The estimation of the PSDs was carried out by the non-parametric Welch method, which is suitable for non-stationary signals [21]. A Hamming window of 2^{15} samples (50% overlap) along with a discrete Fourier transform of 2^{16} samples were used. To avoid the influence of non-physiological factors, each PSD was normalized (PSDn) by dividing all their spectral components by their corresponding total power [22]. Thus, the amplitude values of the PSDns, as measured in 1/Hz, reflect the occurrence of AF events at each frequency.

In order to define the spectral bands of interest, we looked for statistical significant differences between PSDns from OSAS-positive and OSAS-negative groups. Data were not normally distributed. Hence, we used a *p*-value based methodology consisting in applying the non-parametric Mann–Whitney *U* test to the amplitude values of the PSDns from both groups, at each frequency [19]. Fig. 1 shows the median values of the PSDns from OSAS-positive (black line) and OSAS-negative (gray line) samples. It also shows the *p*-values obtained in the comparison of both groups (light gray line). We found marked drops in the *p*-values around [0.06–0.2] Hz., [0.35–0.43] Hz., and [0.7–1] Hz. However, in order to avoid type I errors, we only defined as bands of interest those spectral bands in which the *p*-value were lower than 0.01. Thus, two bands were finally defined: 0.119–0.192 Hz (BW1); 0.784–0.890 Hz (BW2). At each band, we let 10% of components have a *p*-value above 0.01 to maintain coherence with the *p*-value tendency showed in Fig. 1. This avoids the disaggregation of one single homogeneous band into several due to spurious values.

We characterized these two bands by extracting six common spectral features from each of them:

- Maximum and minimum amplitude (*MA*, *mA*), computed as the highest and the lowest PSDn values in each band. These features measure the maximum and minimum occurrence of AF events at the bands.
- First to fourth statistical moments (M_{f1}–M_{f4}). Mean (M_{f1}), standard deviation (M_{f2}), skewness (M_{f3}), and kurtosis (M_{f4}), quantify central tendency, dispersion, asymmetry, and peakedness of the spectral data, respectively.

2.3. Logistic regression: feature selection and classification

The logistic regression (LR) method is a supervised learning algorithm which estimates the posterior probability of a given instance $\mathbf{x_i}$ (in our case, a vector containing the extracted features) belonging certain class C_k (in our case, C_k = OSAS-positive or OSAS-negative). Hence, the posterior probability $p(C_k|\mathbf{x_i})$, i.e.

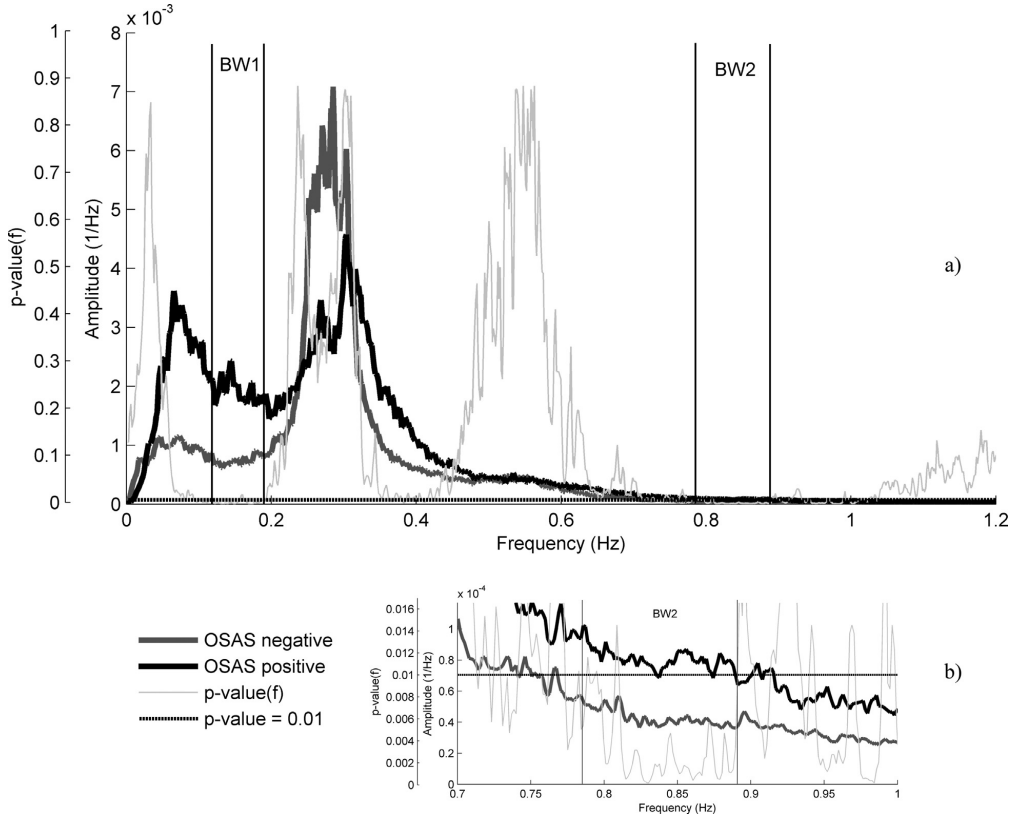

Fig. 1. Median values of the PSDns from OSAS-positive (black line) and OSAS-negative (gray line) samples, and *p*-values at each frequency (light gray line). Significance level *p*-value = 0.01 (black dashed line). In (a), the spectral bands of interest BW1 and BW2 are delimited outside the normal respiratory rate. In (b), BW2 has been expanded for a better viewing.

Table 2
Feature values for OSAS-positive and OSAS-negative groups.

	BW1 Median [IQR]		p	BW2 Median [IQR]		p
	OSAS-positive	OSAS-negative		OSAS-positive	OSAS-negative	
MA (10^{-3})	3.0 [1.3, 4.4]	1.3 [0.8, 2.4]	<0.01	0.11 [0.08, 0.28]	0.06 [0.05, 0.12]	<0.01
mA (10^{-4})	13.0 [7.0, 18.0]	5.1 [4.0, 6.7]	$\ll$0.01	0.5 [0.4, 1]	0.3 [0.2, 0.5]	$\ll$0.01
$Mf1$ (10^{-3})	1.9 [0.9, 2.7]	0.9 [0.5, 1.4]	<0.01	0.08 [0.06, 0.14]	0.04 [0.03, 0.07]	<0.01
$Mf2$ (10^{-4})	3.5 [1.4, 6.8]	1.5 [0.8, 3.9]	0.045	0.16 [0.11, 0.31]	0.09 [0.06, 0.19]	0.029
$Mf3$ (10^{-1})	5.9 [2.9, 8.7]	5.4 [3.5, 9.1]	0.993	3.9 [1.8, 6.6]	4.3 [1.9, 9.3]	0.541
$Mf4$ (10^{0})	2.7 [2.3, 3.3]	2.7 [2.5, 3.3]	0.749	2.7 [2.4, 3.3]	2.5 [2.3, 3.1]	0.356

p: p-value of the Mann–Whitney U test; IQR: interquartile range.

the probability of a subject belonging to OSAS-positive or OSAS-negative group, is computed through the logistic function:

$$p(C_k|\mathbf{x}_i) = \frac{e^{\beta_0+\boldsymbol{\beta}^T\mathbf{x}_i}}{1+e^{\beta_0+\boldsymbol{\beta}^T\mathbf{x}_i}}, \quad (1)$$

where β_0 and $\boldsymbol{\beta}$ are obtained by the weighted least squares minimization procedure [23]. Thus, an instance $\mathbf{x}_i$ is assigned to the class with larger posterior probability.

First, we used LR to automatically select those relevant and non-redundant features. This was performed through the stepwise LR method (SLR), proposed by Hosmer and Lemeshow [23]. Specifically, we applied the well-known forward-selection backward-elimination algorithm. Then, LR was also used to assess the joint potentiality of the selected features from BW1 and BW2 to predict OSAS in children.

2.4. Statistical analysis

Data did not pass the Lilliefors normality test. Hence, the non-parametric Mann–Whitney U test was used to evaluate statistical differences in the obtained features from OSAS-positive and OSAS-negative groups. Sensitivity (Se, percentage of OSAS-positive subjects rightly classified), specificity (Sp, percentage of OSAS-negative subjects rightly classified), accuracy (Acc, overall percentage of subjects rightly classified), positive predictive value (PPV, proportion of positive test results which are true positives), and negative predictive value (NPV, proportion of negative test results which are true negatives) were used to measure the diagnostic ability of both single features and LR models. In the assessment of single features, a receiver operating-characteristics (ROC) analysis was conducted. Thus, the area under the ROC (AROC) was also computed for each case.

2.5. Results validation: bootstrap 0.632

We used the bootstrap 0.632 algorithm to validate our results since it is particularly useful to estimate statistics in small-size samples [24]. Given a sample of N instances, this method proposes building B new samples (bootstrap samples) of size N by resampling with replacement from the original one [24]. A uniform probability is used to randomly select the instances for each B. Thus, the instances can be selected several times for a particular sample B_i, which acts as a training group and, most probably, will contain repeated instances [24]. Consequently, for each new sample, a number of instances from the original are not selected. These instances act as the test group.

The number of subjects in our database is $N = 50$ and the number of bootstrap samples chosen was $B = 1000$, since it ensures a proper estimation of the 95% confidence interval [25] (CI). Thus, 1000 new groups of size 50 were built, acting as training groups. As stated above, the instances not included in each case acted as the corresponding test groups. Following bootstrap 0.632, a statistic s obtained from a test set would be a downward estimation of the true one [25]. Hence, both the training and the test groups are used to compute s by weighting their corresponding estimations as follows [24]:

$$s = 0.632 \cdot s_{\text{test}} + 0.368 \cdot s_{\text{training}}. \quad (2)$$

Finally, the B estimations of s are averaged to show a global performance.

3. Results

3.1. Descriptive analysis and feature selection

Table 2 summarizes the values (median and IQR) of each spectral feature. Consistent with Fig. 1, the values of MA, mA, and Mf_1 in BW1 and BW2 were significantly higher in OSAS-positive than in OSAS-negative subjects (p-value <0.01). Near to significant differences were found in Mf_2 from both bands, and there were no differences in Mf_3 and Mf_4. As expected, ODI3 also showed statistical differences between groups (OSAS-negative: 0.87 e/h IQR [0.44, 1.9], OSAS-positive: 5.9 e/h IQR [1.8, 9.1], p-value <0.01).

SLR was used twice to select relevant and non-redundant features. First, we applied SLR to the 12 spectral features previously obtained. Thus, mA from BW1, and Mf_3 and Mf_4 from BW2 were automatically selected by SLR to form the corresponding model (SLR_{Spec}). Second, the selection process was repeated with the 13 features, i.e., including ODI3. In this case, ODI3, mA and Mf_4 from BW1, as well as Mf_3 from BW2 were selected for the model ($\text{SLR}_{\text{Spec-ODI3}}$).

3.2. Diagnostic performance

Table 3 shows the diagnostic performance of the spectral features and ODI3 after the bootstrap 0.632 procedure. Se, Sp, Acc, PPV, and NPV values (mean and 95% CI) were obtained by weighting the training and test estimations according to bootstrap 0.632, and averaging the results from the 1000 training-test group pairs. The best single feature in terms of Acc and AROC was the spectral mA from BW1 (76.3% [65.7–84.2]; 0.743 [0.584–0.871], respectively), outperforming ODI3 from oximeter (75.3% [67.0–83.4]; 0.676 [0.513–0.829]).

Table 4 includes the diagnostic performance of SLR_{Spec} and $\text{SLR}_{\text{Spec-ODI3}}$. SLR_{Spec}, which only used spectral information from AF, outperformed all the single features in terms of Acc and AROC (79.1% [68.6–87.9]; 0.875 [0.723–1]). The $\text{SLR}_{\text{Spec-ODI3}}$ model, which combines spectral information from AF with ODI3 from SpO_2, obtained the highest results at each statistic (85.9% Se [64.5–98.7]; 87.4% Sp [70.2–98.6]; 86.3% Acc [74.9–95.4]; 0.947 AROC [0.826–1]; 88.4% PPV [72.3–98.5]; 85.8% NPV [65.8–98.5]).

Table 3
Diagnostic performance of the single features.

	Se (%) [CI]	Sp (%) [CI]	Acc (%) [CI]	PPV (%) [CI]	NPV (%) [CI]	AROC [CI]
BW1						
MA	60.2 [37.0,87.8]	71.7 [34.5,95.8]	65.6 [55.4,74.2]	71.7 [53.6,94.9]	62.9 [49.1,80.3]	0.651 [0.450,0.805]
mA	71.9 [47.1,87.4]	81.1 [65.3,100.0]	76.3 [65.7,84.2]	80.9 [62.3,100.0]	73.2 [58.7,87.3]	0.743 [0.584,0.871]
Mf_1	66.4 [40.5,85.9]	72.3 [50.2,95.1]	69.1 [54.7,77.3]	72.9 [57.9,91.6]	67.3 [53.2,83.6]	0.684 [0.515,0.825]
Mf_2	59.4 [32.0,85.3]	67.0 [31.9,91.6]	62.9 [52.4,72.7]	67.1 [51.4,88.7]	60.8 [46.0,77.2]	0.603 [0.416,0.769]
Mf_3	53.8 [29.2,78.7]	57.0 [30.8,79.2]	54.0 [43.4,67.7]	57.5 [40.1,75.5]	53.6 [38.5,72.1]	0.542 [0.422,0.675]
Mf_4	52.1 [27.3,77.9]	56.3 [29.3,80.8]	60.2 [42.2,67.9]	57.8 [40.5,77.9]	50.7 [25.8,71.6]	0.539 [0.372,0.698]
BW2						
MA	74.6 [53.1,95.6]	64.4 [43.8,82.7]	70.6 [60.5,79.2]	70.1 [57.6,83.2]	73.2 [55.2,93.9]	0.670 [0.498,0.809]
mA	83.4 [48.6,98.7]	65.6 [51.2,85.8]	74.8 [61.1,83.6]	72.3 [60.5,84.4]	80.8 [56.7,98.0]	0.730 [0.576,0.859]
Mf_1	79.7 [40.4,96.4]	65.2 [47.8,84.4]	72.7 [59.6,81.3]	71.2 [58.8,83.5]	76.6 [55.2,94.7]	0.698 [0.527,0.837]
Mf_2	74.6 [42.1,92.2]	64.2 [41.7,83.6]	69.5 [57.6,79.1]	69.3 [56.1,82.8]	71.1 [53.0,89.9]	0.627 [0.449,0.784]
Mf_3	52.4 [27.6,78.2]	52.1 [25.8,78.7]	52.4 [39.1,67.1]	53.4 [30.2,73.2]	50.9 [32.7,71.8]	0.557 [0.422,0.706]
Mf_4	60.3 [33.5,83.4]	58.9 [34.6,80.1]	59.6 [48.1,70.4]	61.5 [47.4,78.0]	58.1 [42.0,74.2]	0.574 [0.436,0.712]
ODI3	70.9 [49.5,94.6]	80.3 [46.5,100.0]	75.3 [67.0,83.4]	81.9 [62.9,100.0]	72.8 [37.0,87.8]	0.676 [0.513,0.829]

CI: 95% confidence interval.

4. Discussion

In this paper, an alternative diagnostic methodology for OSAS in children was developed by combining spectral information from AF with the classic ODI3 from SpO_2. Our proposal was assessed by answering three research questions.

i. How does OSAS modify the spectral information of airflow recordings from children?

We found that the spectral power of AF was significantly higher in OSAS-positive subjects at novel frequency bands below (BW1) and above (BW2) the typical respiratory range in children reported in previous studies (0.220–0.430 Hz) [3,26,27]. The relationship of BW1 with apneas and hypopneas can be explained on the basis of the definition of these apneic events in children. As stated in section 2.1, apneas and hypopneas require at least 2 missed breaths of length in order to be scored [17]. Missing 2 cycles means that the recurrence of these apneic events is every 2 normal breaths, at most. Therefore, their frequency has to be located below the half of the normal respiratory frequency range, modifying the spectrum of AF in such band. Since BW1 is located below the half of the normal respiratory band, it is consistent with the occurrence of apneas and hypopneas. On the other hand, differences in the high frequency band, BW2 (0.784–0.890 Hz.), may be explained as the typical respiratory overexertion after an apneic event, which increases the respiratory rate [9]. Moreover, the greater variability in the PSDn of OSAS-positive children in the range 0.35–0.43 Hz., which is shown in Fig. 1, is consistent with the decrease of the deep sleep stage time of these patients reported in other works [28]. During deep sleeping, respiration becomes more regular [29], which leads to a condensed normal breathing band in the PSDn. OSAS interrupts the sleep cycle by the recurrence of arousals [1], causing respiratory instabilities [29] and, consequently, a more variable normal breathing rate.

ii. Are these changes useful to distinguish OSAS in children from at-home recordings?

Seven out of the 13 extracted features were significantly different in OSAS-positive than in OSAS-negative subjects, (6 out of 12 from AF, and ODI3). In the diagnostic ability assessment, *mA* from BW1 outperformed ODI3, whereas *mA* from BW2 performed similarly. Both SLR_{Spec} and $SLR_{Spec\text{-}ODI3}$ outperformed all the single features. Particularly high was the diagnostic ability of $SLR_{Spec\text{-}ODI3}$, which widely improved the performance of an in-lab 6-channel RP (74.2% Se, 81.8% Sp, 77.4% Acc, and 0.852 AROC) [9], only requiring information from 2 channels (thermistor and oximeter) recorded at patients' home. Additionally, SLR_{Spec} (information from single-channel AF only) also outperformed this 6-channel RP.

iii. Is the airflow spectral information complementary to the classic oxygen desaturation index in pediatric OSAS detection?

Table 4
Diagnostic performance of the logistic regression models.

	Se (%) [CI]	Sp (%) [CI]	Acc (%) [CI]	PPV (%) [CI]	NPV (%) [CI]	AROC [CI]
SLR_{Spec}	79.2 [59.1,96.6]	79.4 [59.3,95.8]	79.1 [69.6,87.9]	81.2 [65.2,94.5]	78.8 [60.4,95.4]	0.875 [0.723,1]
$SLR_{Spec\text{-}ODI3}$	85.9 [64.5,98.7]	87.4 [70.2,98.6]	86.3 [74.9,95.4]	88.4 [72.3,98.5]	85.8 [65.8,98.5]	0.947 [0.826,1]

The study showed complementarity between features in two cases: first, between features from the two novel AF bands, since SLR automatically selected features from both of them to build the SLR_{Spec} and the $SLR_{Spec\text{-}ODI3}$ models; second, between features from the two spectral bands BW1-BW2 and the ODI3, since the latter was also selected for the $SLR_{Spec\text{-}ODI3}$ model.

Other recent studies also analyzed physiological signals to help in pediatric OSAS diagnosis. Shouldice et al. used 50 ECG recordings, and reached 85.7% Se, 81.8% Sp, and 84% Acc in a test set (AHI $\geq$ 1), by applying a quadratic linear discriminant to 23 features [15]. Gil et al. investigated the diagnostic usefulness of the information contained in 21 PPG time series, reporting 75.0% Se, 85.7% Sp, and 80.0% Acc after a leave-one-out cross-validation procedure (AHI $\geq$ 5) [11]. The relationship of high frequency inspiratory sounds (HFIS) to OSAS in children has been evaluated as well [30,31]. Rembold and Suratt reported data to estimate that 10 HFIS events per hour can be useful to discriminate OSAS in children both for AHI $\geq$ 1 (70% Se, 100% Sp, and 76.9% Acc) and AHI $\geq$ 3 (61.5% Se, 100% Sp., and 80.8% Acc) [30]. Questionnaires and common symptoms have been also involved in screening tools for OSAS and sleep-disordered breathing. Spruyt and Gozal proposed a severity scale based on the answers of 1133 children from general population to 6 sleep-related questions [32]. They used a predictive score which reached 59.0% Se, 82.9% Sp, 0.79 AROC, 35.4% PPV, and 92.7% NPV (AHI $\geq$ 3). Kadmon et al. validated this 6-item questionnaire in a sample of 85 children referred to a pediatric sleep clinic [33], reaching 83.0% Se, 64.0% Sp, 0.65 AROC, 28.0% PPV, and 96% NPV (AHI $\geq$ 5). Finally, Chang et al. combined symptoms (observable apnea, restless sleep, and mouth breathing) with ODI from 141 children to assess both a logistic regression model and a new discriminative score [10]. The former reached 76.6% of diagnostic accuracy whereas the latter achieved 60.0% Se, 86.0% Sp, 71.6% Acc, 84.0% PPV, and 64.0% NPV (AHI $\geq$ 5). Our $SLR_{Spec\text{-}ODI3}$ outperformed the reported diagnostic ability in these studies, even though we used recordings obtained from an unsupervised environment. However, Shouldice et al. used a more restrictive AHI threshold to differentiate patients from control subjects and Gil et al., as well as Rembold and Suratt, worked with one single channel.

Some limitations have to be addressed in this study. The sample size should be larger to empower the generalization ability of our results. Although the bootstrap 0.632 algorithm is known to provide good estimates from small datasets [24], the assessment of our methodology in a larger sample is a very interesting future target. Additionally, a larger sample would let us define the AF bands of interest through an independent set of subjects. Nonetheless, our bands were consistent with the pathophysiology of the apneic events. A wide sample of subjects would be also useful to optimize the set of selected features. Moreover, since our methodology relies on a classification problem, it only provides information about the presence of OSAS and not about its severity. In this sense, future work focused on estimating the AHI or assessing different AHI thresholds could complement our findings. The only use of a thermistor to record AF is another limitation of the study since the AASM recommends the use of a thermistor to score apneas and a nasal pressure transducer to score hypopneas [17]. However, our approach does not rely on event scoring and, in spite of using thermistor alone, results showed high diagnostic ability. Recent studies have shown high performance when using automatic analysis of single-channel AF from thermistor in adults [18,19]. Finally, the application of different spectral or non-linear measures, as well as the training of more complex classification models, may be also useful to enhance our methodology.

5. Conclusion

To the best of our knowledge, this is the first time that the spectral information of AF recordings from children is analyzed in the context of OSAS. We showed that OSAS in children significantly modifies the PSDn of AF at two abnormal respiratory bands. Diagnostic ability of single features from these novel bands is similar to that of classic ODI3. Additionally, the information contained in the two bands showed complementarity both between them and with ODI3. Our optimum LR model, built with information from thermistor and oximeter at-home recordings, outperformed the diagnostic ability reported in previous in-lab studies focused on finding new alternatives to standard PSG. These results suggest that the spectral information contained in AF recordings is useful to help in pediatric OSAS and that its combination with ODI3 could be beneficial to diagnose OSAS in children at home.

Acknowledgments

This study was supported by the Proyecto Cero 2011 on Aging from Fundación General CSIC, the project TEC2011-22987 from Ministerio de Economía y Competitividad, the project VA059U13 from the Consejería de Educación de la Junta de Castilla y León, FEDER and SEPAR. G. C. Gutiérrez-Tobal was in receipt of a PIRTU grant from the Consejería de Educación de la Junta de Castilla y León and the European Social Fund.

References

[1] S.P. Patil, H. Schneider, A.R. Schwartz, P.L. Smith, Adult obstructive apnea: pathophysiology and diagnosis, Chest 132 (1) (2007) 325–337, http://dx.doi.org/10.1378/chest.07-0040.
[2] C.L. Marcus, L.J. Brooks, S. Davidson, et al., Diagnosis and management of childhood obstructive sleep apnea syndrome, Pediatrics 130 (3) (2012) e714–e755, http://dx.doi.org/10.1542/peds.2012-1672.
[3] C. Guilleminault, J.H. Lee, A. Chan, Pediatric obstructive sleep apnea syndrome, Arch. Pediatr. Adolesc. Med. 159 (2005) 775–785, http://dx.doi.org/10.1001/archpedi.159.8.775.
[4] J. Kim, R. Bhattacharjee, L. Kheirandish-Gozal, K. Spruyt, D. Gozal, Circulating microparticles in children with sleep disordered breathing, Chest 140 (2) (2011) 408–417, http://dx.doi.org/10.1378/chest. 10-2161.
[5] A.J. Lipton, D. Gozal, Treatment of obstructive sleep apnea in children: do we really know how? Sleep Med. Rev. 7 (1) (2003) 61–80, http://dx.doi.org/10.1053/smrv.2001.0256.
[6] H. Reuveni, T. Simon, A. Tal, A. Elhayany, A. Tarasiuk, Health care services utilization in children with obstructive sleep apnea syndrome, Pediatrics 110 (1) (2002) 68–72.
[7] J.A. Bennet, W.J.M. Kinnear, Sleep on the cheap: the role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome, Thorax 54 (1999) 958–959, http://dx.doi.org/10.1136/thx.54.11.958.
[8] E.S. Katz, R.B. Mitchell, C.M. D'Ambrosio, Obstructive sleep apnea in infants, Am. J. Respir. Crit. Care Med. 185 (8) (2012) 805–816, http://dx.doi.org/10.1164/rccm.201108-1455CI.
[9] M.L. Alonso-Álvarez, J. Terán-Santos, J.A. Cordero-Guevara, et al., Reliability of respiratory polygraphy for the diagnosis of sleep apnea-hypopnea syndrome in children, Arch. Bronconeumol. 44 (2008) 318–323, http://dx.doi.org/10.1016/S1579-2129(08)60052-X.
[10] L. Chang, J. Wu, L. Cao, Combination of symptoms and oxygen desaturation index in predicting childhood obstructive sleep apnea, Int. J. Pediatr. Otorhinolaryngol. 77 (3) (2013) 365–371, http://dx.doi.org/10.1016/j.ijporl.2012.11.028.
[11] E. Gil, R. Bailón, J.M. Vergara, P. Laguna, PTT variability for discrimination of sleep apnea related decreases in the amplitude fluctuations of PPG signal in children, IEEE Trans. Biomed. Eng. 57 (5) (2010) 1079–1088, http://dx.doi.org/10.1109/TBME.2009.2037734.
[12] J.V. Marcos, R. Hornero, D. Álvarez, F. del Campo, M. Aboy, Automated detection of obstructive sleep apnea syndrome from oxygen saturation recordings using linear discriminant analysis, Med. Biol. Eng. Comput. 49 (9) (2010) 895–902, http://dx.doi.org/10.1007/s11517-010-0646-6.
[13] D. Álvarez, R. Hornero, J.V. Marcos, 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. 23 (5) (2013) 1–18, http://dx.doi.org/10.1142/S0129065713500202.
[14] T. Penzel, J. McNames, P. De Chazal, B. Raymond, A. Murray, G. Moody, Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings, Med. Biol. Eng. Comput. 40 (4) (2002) 402–407, http://dx.doi.org/10.1007/BF02345072.

[15] R.E. Shouldice, L.M. O'Brien, C. O'Brien, P. De Chazal, D. Gozal, C. Heneghan, Detection of obstructive sleep apnea in pediatric subjects using surface lead electrocardiogram features, Sleep 27 (4) (2004) 784–792.
[16] J. Vavrina, Computer assisted pulse oximetry for detecting children with obstructive sleep apnea syndrome, Int. J. Pediatr. Otorhinolaryngol. 33 (3) (1995) 239–248, http://dx.doi.org/10.1016/0165-5876(95)01217-6.
[17] C. Iber, S. Ancoli-Israel, A.L. Chesson, S.F. Quan, The AASM Manual for the Scoring of Sleep and Associated Events, Manual, American Academy of Sleep Medicine, 2007.
[18] G.C. Gutiérrez-Tobal, R. Hornero, D. Álvarez, J.V. Marcos, F. del Campo, Linear and nonlinear analysis of airflow recordings to help in sleep apnoea–hypopnoea syndrome diagnosis, Physiol. Meas. 33 (7) (2012) 1261–1275, http://dx.doi.org/10.1088/0967-3334/33/7/1261.
[19] G.C. Gutiérrez-Tobal, D. Álvarez, J.V. Marcos, F. del Campo, R. Hornero, Pattern recognition in airflow recordings to assist in the sleep apnoea–hypopnoea syndrome diagnosis, Med. Biol. Eng. Comput. 51 (2013) 1367–1380, http://dx.doi.org/10.1007/s11517-013-1109-7.
[20] M.L. Alonso-Álvarez, T. Canet, M. Cubel-Alarco, E. Estivill, E. Fernandez-Julian, D. Gozal, M.J. Jurado-Luqué, A. Lluch-Roselló, F. Martínez-Pérez, M. Merino-Andreu, G. Pin-Arboledas, N. Roure, F. Sanmartí, O. Sans-Capdevila, F. Segarra-Isern, T. Tomás-Vila, J. Terán-Santos, Consensus document on sleep apnea-hypopnea syndrome in children, Arch. Bronconeumol. 47 (5) (2011) 1–18, http://dx.doi.org/10.1016/S0300-2896(11)70026-6.
[21] 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 Audio Electroacoust. AU-15 (1967) 70–73.
[22] L. Sörnmo, P. Laguna, Bioelectrical Signal Processing in Cardiac and Neurological Applications, Elsevier/Academic, London, U.K./New York, 2005.
[23] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, John Wiley and Sons, New York, NY, USA, 2000.
[24] I.H. Witten, E. Frank, M.A. Hall, Data Mining Practical Machine Learning Tools and Techniques, Morgan Kaufmann/Elsevier, Burlington, MA, USA, 2011.
[25] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, New York, NY, USA, 1994.
[26] E. Tabachnik, N. Muller, B. Toye, H. Levison, Measurement of ventilation in children using the respiratory inductive plethysmograph, J. Pediatr. 99 (6) (1981) 895–899, http://dx.doi.org/10.1016/S0022-3476(81)80012-1.
[27] S. Fleming, M. Thompson, R. Stevens, et al., Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies, Lancet 377 (9770) (2011) 1011–1018, http://dx.doi.org/10.1016/S0140-6736(10)62226-X.
[28] C. Guilleminault, R. Korobkin, R. Winkle, A review of 50 children with obstructive sleep apnea syndrome, Lung 159 (1) (1981) 275–287, http://dx.doi.org/10.1007/BF02713925.
[29] S.A. Shea, Behavioural and arousal-related influences on breathing in humans, Exp. Physiol. 81 (1) (1996) 1–26.
[30] C.M. Rembold, P.M. Suratt, Children with obstructive sleep-disordered breathing generate high-frequency inspiratory sounds during sleep, Sleep 27 (6) (2004) 1154–1161.
[31] C.M. Rembold, P.M. Suratt, An upper airway resonator model of high-frequency inspiratory sounds in children with sleep-disordered breathing, J. Appl. Physiol. 98 (5) (1985) 1855–1861, http://dx.doi.org/10.1152/japplphysiol.01231.2004.
[32] K. Spruyt, D. Gozal, Screening of pediatric sleep-disordered breathing pediatric sleep-disordered breathing complaints. A proposed unbiased discriminative set of questions using clinical severity scales, Chest 142 (6) (2012) 1508–1515, http://dx.doi.org/10.1378/chest.11-3164.
[33] G. Kadmon, C.M. Shapiro, S.A. Chung, D. Gozal, Validation of a pediatric obstructive sleep apnea screening tool, Int. J. Pediatr. Otorhinolaryngol. 77 (9) (2013) 1461–1464, http://dx.doi.org/10.1016/j.ijporl.2013.06.009.

Utility of AdaBoost to Detect Sleep Apnea-Hypopnea Syndrome from Single-Channel Airflow

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

***Abstract*—*Goal:* The purpose of this study is to evaluate the usefulness of the boosting algorithm *AdaBoost* (AB) in the context of the sleep apnea-hypopnea syndrome (SAHS) diagnosis. *Methods:* We characterize SAHS in single-channel airflow (AF) signals from 317 subjects by the extraction of spectral and non-linear features. Relevancy and redundancy analyses are conducted through the fast correlation-based filter (FCBF) to derive the optimum set of features among them. These are used to feed classifiers based on linear discriminant analysis (LDA) and classification and regression trees (CART). LDA and CART models are sequentially obtained through AB, which combines their performances to reach higher diagnostic ability than each of them separately. *Results:* Our AB-LDA and AB-CART approaches showed high diagnostic performance when determining SAHS and its severity. The assessment of different apnea-hypopnea index cutoffs using an independent test set derived into high accuracy: 86.5% (5 events/h), 86.5% (10 events/h), 81.0% (15 events/h), and 83.3% (30 events/h). These results widely outperformed those from logistic regression and a conventional event-detection algorithm applied to the same database. *Conclusion:* Our results suggest that AB applied to data from single-channel AF can be useful to determine SAHS and its severity. *Significance*: SAHS detection might be simplified through the only use of single-channel AF data.**

***Index Terms*—AdaBoost, airflow, sleep apnea-hypopnea syndrome, spectral analysis, nonlinear analysis**

I. Introduction

In recent years, the Sleep Apnea-Hypopnea Syndrome (SAHS) has become a major concern due to the high prevalence and severe consequences for the patients' health and quality of life [1], [2]. People suffering from SAHS experiment recurrent episodes of complete (apnea) or partial (hypopnea) collapse of the upper airway during sleep, which lead to cessation or significant reduction of airflow (AF) [3]. These apneic events cause oxygen desaturations and arousals [3], preventing patients from resting while sleeping [2]. Unsuccessful rest derives in daytime symptoms such as hypersomnolence, cognitive impairment, and depression [1], some of which have been related to motor-vehicle collisions and occupational accidents [4], [5]. Moreover, SAHS has been associated with cardiac and vascular illnesses [2], as well as with an increase in the cancer incidence [6].

The standard test to diagnose SAHS is overnight in-lab polysomnography (PSG) [3]. Although its effectiveness is well-known, PSG implies monitoring and recording multiple physiological signals, including electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG), oxygen saturation of blood (SpO_2), and AF [3]. This makes PSG a complex test which requires expensive equipment and technical expertise [7], [8]. Moreover, the specialists need an offline inspection of the recordings to derive the apnea-hypopnea index (AHI), which is the parameter used to establish SAHS and its severity [9]. Thus PSG is also time-consuming, leading to a delayed diagnostic process and increased waiting lists [8], [10].

One widespread approach to reduce complexity, cost, and time delay is the study of a limited set of signals among those involved in PSG [8]. The analysis of a single one has been often adopted. Thus, the oxygen desaturation index (ODI) from SpO_2, the apneic-related events from ECG, and the respiratory disturbance index (RDI) from AF have been already assessed to help in SAHS diagnosis [10]-[13]. These works followed a common methodology: detecting the effects caused by each apnea and hypopnea in the signals under study, scoring them as apneic-related events, and deriving the corresponding diagnostic index. However, our research group has lately adopted a different approach based on an exhaustive analysis of a signal through the extraction of global features [14]-[18].

In this paper, we propose such a global analysis in single-channel AF. AF is a straightforward choice to look for simpler alternatives to PSG, since apneas and hypopneas are defined on the basis of its amplitude oscillations [9]. The American Academy of Sleep Medicine (AASM) recommends the use of two AF channels: one acquired through an oronasal thermal sensor and the second one acquired by means of a nasal prong pressure sensor (NPP) [9]. The former is suitable for a proper scoring of apneas whereas the latter is used to score hypopneas [9]. However, previous studies have shown that it is possible to reach high diagnostic ability following an automatic global analysis of the single-channel AF from a thermal sensor [17], [18]. In this paper, one major goal is to assess whether it is also possible to reach a high performance when using data from single-channel AF obtained by NPP.

Our proposal starts with the extraction of spectral (frequency domain) and non-linear (time domain) features from NPP AF. The analysis in frequency domain is justified due to the overnight recurrence of these events. Thereby, common spectral features have already shown their utility to characterize SAHS as well as other disorders [15]-[19]. On the other hand, non-linear measures of variability, complexity,

and irregularity in time series have been also used to extract useful information from biomedical signals [14], [17]-[19]. This exhaustive characterization of AF, however, may lead to obtain features with a high degree of shared information, i.e., redundant features. In order to avoid this issue, a second step is included in our methodology: an automatic feature selection stage based on the fast correlation-based filter (FCBF) [20]. The FCBF algorithm selects optimum sets of features on the basis of their relevancy and redundancy. It has been also assessed in biomedical applications [17], [21]. Finally, a classification approach is used to distinguish SAHS and its severity. Thus, we evaluate two different cases: a binary classification task, in which the objective is to determine the presence (SAHS-positive) or absence (SAHS-negative) of SAHS, and a multiclass task, in which the aim is to assess the AHI cutoffs which establish the four severity levels of SAHS (no-SAHS, mild-SAHS, moderate-SAHS, and severe-SAHS). We propose the *AdaBoost* (AB) algorithm for both classification tasks. AB is a boosting algorithm commonly used to take advantage of the performance of several weak classifiers of the same type [22]. It is known to be able to reach high yields when it is applied to new data [22], i.e., the AB algorithm produces generalized models. Moreover, it relies on a simple sequential procedure [22], which barely increases the complexity of the methodology. These characteristics make it a suitable algorithm to be used in diagnostic aid contexts. Actually, it has been already assessed in the context of SAHS under a classic event-detection approach [23], [24]. As weak classifiers we propose two well-known machine learning algorithms based on *i*) linear discriminant analysis (LDA) and *ii*) classification and regression trees (CART). Both of them have been already assessed in the context of SAHS [16], [23]. Since classifiers favor the right sorting of classes with more subjects, one major issue in the present work is how to deal with imbalanced classes. The high prevalence of SAHS leads to prioritize diagnosis in at-risk population [25]. Consequently, data from SAHS patients is more available than from no SAHS subjects. Thus, to compensate for this imbalance, we use the synthetic minority oversampling technique (SMOTE) [26], which creates new synthetic data from the minority classes on the basis of the real data.

Our hypothesis is that the information obtained from AF and the generalization ability of AB can be useful to automatically detect SAHS and establish its severity. Thus, the main objective of the present work is to evaluate the diagnostic usefulness of AB when the only source of SAHS-related information is single-channel AF from NPP. In order to achieve this goal, we evaluate whether our proposal outperforms the diagnostic ability of a typical classification algorithms such as logistic regression (LR), which is based on one single classifier. We also apply to our AF recordings an algorithm focused on the classical event-detection approach, which has been previously assessed in other databases [17], [27]. Finally, our results are also compared with other recent studies focused on SAHS detection from single-channel AF.

II. Population And Signal Under Study

In this study, AF recordings from 317 adults were involved. Before undergoing PSG, all of the subjects suffered from common symptoms such as daytime sleepiness, loud snoring, nocturnal choking and awakenings, and/or referred apneic events. PSG was conducted in the sleep unit of the Hospital Universitario Río Hortega in Valladolid, Spain. Physicians scored apneas and hypopneas according to the American Academy of Sleep Medicine (AASM) rules [9]. Consequently, an apnea was defined as a 90% or more reduction in the pre-event baseline of the AF amplitude, measured through an oronasal thermal sensor. In contrast, a hypopnea was scored after 30% or more reduction in the pre-event baseline of the AF amplitude, measured through a nasal pressure sensor, and accompanied by a drop of 3% in SpO_2 and/or an EEG arousal. In both cases, duration of 10 seconds or more was required to annotate the event [9]. All the subjects gave their informed consent and the Ethics Committee of the Hospital Universitario Rio Hortega (Spain) accepted the protocol.

Common AHI cutoffs to determine SAHS and its severity are 5, 10, 15, and 30 e/h [9], [10], [13], [17]. Particularly, SAHS severity levels are: no-SAHS (5<AHI), mild-SAHS (5≤AHI<15), moderate-SAHS (15≤AHI<30), and severe-SAHS (AHI>30) [28]. Alternatively, AHI=10 e/h has been widely used as cutoff to determine the presence or absence of SAHS [10], [13], [17], [18], [29]. Consequently, for the binary classification task, we chose AHI=10 e/h to distinguish SAHS-negative and SAHS-positive subjects, whereas for the multi-classification task we divided our database according to the four SAHS severity levels. Tables I and II show clinical and demographical data of the subjects under study when they are divided for the binary or the multiclass tasks, respectively. No statistically significant differences were found (p-value>0.01) between SAHS-positive and SAHS-negative (Mann-Whitney U test), or among the four severity levels (Kruskal-Wallis test), in body mass index (BMI) and age.

The AF recordings were obtained during overnight PSG, which was performed through a polysomnograph (E-series, Compumedics). A NPP sensor was used to acquire AF (sample rate=128 Hz). The recording length was 7.4 ± 0.3 hours (mean ± standard deviation). An anti-aliasing filter was applied to the AF recordings to satisfy the Nyquist-Shannon theorem. We also applied an infinite impulse response Butterworth low-pass filter (cutoff = 1.2 Hz) to reduce noise for a prospective non-linear analysis in time domain.

We divided our recordings into a training set (60%) and a test set (40%). A uniformly random selection was conducted to assign the AF recordings to each one. However, for the sake of the balance of the classes in the training set, we fixed the size of each class as follows: 29 no-SAHS subjects, 54 mild-SAHS, 54 moderate-SAHS, and 54 severe-SAHS. This distribution in the multiclass problem leads to 75 SAHS-negative and 116 SAHS-positive for the binary classification task. The SMOTE algorithm was used to compensate the remaining imbalance in classes of the training set (section *F*).

TABLE I
DEMOGRAPHIC AND CLINICAL DATA FOR THE TWO-CLASS DIVISION

	All	SAHS-negative	SAHS-positive
# Subjects	317	110	207
Age (years)	49.9 ± 12.0	47.6 ± 12.9	51.1 ± 11.4
Men (%)	226 (71.3)	68 (61.8)	158 (76.3)
BMI (kg/m^2)	28.1 ± 5.2	26.5 ± 5.0	29.0 ± 5.1
AHI (e/h)	28.1 ± 26.5	6.0 ± 2.6	39.9 ± 25.9

TABLE II
DEMOGRAPHIC AND CLINICAL DATA FOR THE FOUR-CLASS DIVISION

	no-SAHS	mild	moderate	severe
# Subjects	39	92	70	116
Age (years)	43.9 ± 12.5	50.3 ± 12.4	49.9 ± 11.3	51.6 ± 11.5
Men (%)	19 (48.7)	58 (63.0)	56 (80.0)	93 (80.2)
BMI(kg/m^2)	26.0 ± 5.5	27.0 ± 4.6	28.5 ± 3.9	29.5 ± 5.8
AHI (e/h)	3.0 ± 1.3	8.6 ± 2.4	22.2 ±4.1	55.7 ± 24.7

The recordings not selected for the training set were assigned to the test set.

III. METHODS

Our methodology consists of three steps. First, a feature extraction stage is implemented, in which spectral and nonlinear analyses are conducted over the AF recordings. Then, an automatic feature selection is performed to obtain an optimum set of the extracted features. Finally, a boosting classification approach is adopted to determine SAHS (binary classification) and its severity (multiclass task). Fig. 1 depicts a block diagram with the entire methodology followed during the study, which is widely explained in next sections.

A. *Feature extraction*

1) Spectral analysis

Apneas and hypopneas recurrently modify AF throughout the night. This behavior supports its study in the frequency domain. Hence, the power spectral density (PSD) of each AF recording was estimated. Welch's method was applied for this purpose since it is suitable for non-stationary signals [30]. A Hamming window of 2^{15} points (50% overlap), along with a discrete Fourier transform of 2^{16} points, were used to compute PSD. To avoid the influence of factors not related to the pathophysiology of SAHS, each PSD was normalized (PSDn) dividing the amplitude value at each frequency by their corresponding total power [31]. Fig. 2a shows the averaged PSDn for the four SAHS severity groups in the training set.

A spectral band of interest (BW) was defined between 0.025 Hz. and 0.050 Hz. (Fig. 2b). This corresponds to events lasting from 20 to 40 seconds, which has been reported as the typical range of the apneic events duration [32]. Moreover, BW is consistent with the bands found through statistical approaches [17], [18]. Thus, to characterize SAHS, 9 spectral features were extracted from the 0.025-0.050 Hz. band of each PSDn: minimum amplitude (*mA*), maximum amplitude (*MA*), first to fourth statistical moments (M_{f1}- M_{f4}), median frequency (*MF*), spectral entropy (*SpecEn*), and Wootters distance (*WD*).

mA and *MA* were computed as the lowest and the highest PSDn values in BW. Since PSD is normalized, the amplitude values of the original AF time-series do not affect the power at each frequency component. Hence, as BW is related to apneic events, *mA* and *MA* estimate the minimum and the maximum occurrence of them. Mean (M_{f1}), standard deviation (M_{f2}), skewness (M_{f3}), and kurtosis (M_{f4}) of BW were also obtained. They are common statistics which quantify central tendency, dispersion, asymmetry, and peakedness of data, respectively. According to Fig. 2b, *mA* and *MA* should be higher as SAHS worsens. Similarly, the mean (M_{f1}) and the standard deviation (M_{f2}) should be also higher. Finally, both the skewness (M_{f3}) and the peakedness (M_{f4}) seem to be higher in the BW spectral data of moderate and severe groups.

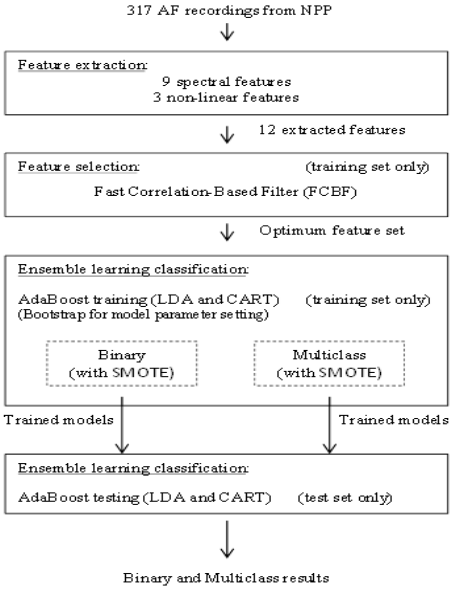

Fig. 1 Block diagram of the signal processing methodology followed during the study.

MF is defined as the frequency component which separates the spectrum into two parts with 50% of the power each of them [33]. Thus, the lower the *MF* value, the more comprised is the spectrum into small frequencies. As seen in Fig 2b, the spectrum of BW for the no-SAHS and mild-SAHS groups is flat, i. e., the power is equally distributed. Conversely, a fewer amount of power is observed in higher frequencies of BW for moderate-SAHS and severe-SAHS groups. As a consequence, a *MF* value closer to 0.0375 (the half of the band) is expected for the lowest severity degrees.

SpecEn quantifies the flatness of the PSD content, which indirectly measures the irregularity of time series [33]. Thereby, high values of *SpecEn* are related to a flat PSD (similar to white noise) and, consequently, it is associated with more irregularity in time domain. By contrast, low values imply a spectrum condensed into a narrow frequency band, which is related to less irregularity in time domain (like in a sum of sinusoids) [33]. A flatter spectrum is observed in BW as SAHS severity decreases. Therefore, higher values of *SpecEn* are expected in no-SAHS and mild-SAHS groups.

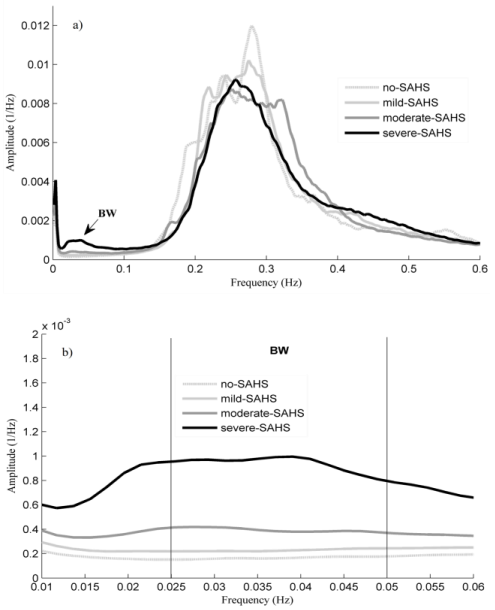

Fig. 2 a) Averaged PSDn for the four SAHS severity groups in the training set and b) detail of the band of interest BW.

WD is a disequilibrium measure which assigns values close to 1 to those distributions with higher statistical distance to the uniform distribution; whereas values close to zero are assigned as the distance becomes smaller [34], [35]. In BW, the averaged spectrum of the no-SAHS and mild-SAHS groups is similar to a normal distribution (Fig. 2b). Hence, smaller values of *WD* are expected than in the case of the moderate-SAHS and severe-SAHS groups.

2) Non-linear analysis

Alterations caused by SAHS in AF could modify the variability, the complexity, and the irregularity of the signal. Hence, to complement the spectral analysis, three global non-linear features were also obtained from each recording in time domain: central tendency measure (*CTM*), Lempel-Ziv complexity (*LZC*), and sample entropy (*SampEn*). Similarly to PSD, each AF time series was normalized before obtaining *CTM*, *LZC*, and *SampEn*. Thereby, measuring effects caused by factors not related to the pathophysiology of SAHS are avoided. We firstly eliminated the spurious values of the signal. Then the time-series were divided by the remaining maximum absolute value in order to constrain each recording into the range -1, 1.

CTM quantifies the variability of a given series $x[n]$ on the basis of its first-order differences [36]. These are plotted following $x[n+2]$-$x[n+1]$ vs. $x[n+1]$-$x[n]$ [37]. The value of *CTM* is computed as the proportion of points in the plot which fall within a radius ρ [36], which acts as a design parameter. Thus, *CTM* ranges between 0 and 1, with higher values corresponding to points more concentrated around the center of the plot, i.e., corresponding to less degree of variability. People suffering from SAHS experiment continuous changes in the respiratory pattern (apneic events, snoring, choking, respiratory overexertion after apneas and hypopneas), which may add variability to the AF signal. Consequently, it is expected that *CTM* decreases in the presence of SAHS.

LZC estimates the complexity of a finite sequence of symbols [38]. Hence, the first step of the algorithm is to convert a time-series $x[n]$ into such a sequence [37]. Usually, a binary transformation is performed, with the median of each $x[n]$ being used as threshold [37]. Then the sequence is scanned, and a counter $c(n)$ is increased with every new subsequence of consecutive symbols. Finally, $c(n)$ is normalized in order to make the method independent of the sequence length. The higher the value of *LZC*, the higher the complexity of the corresponding time-series is [37]. Abnormalities in the AF pattern may introduce new subsequences of symbols. Hence, more complexity is expected in the AF of SAHS patients.

SampEn is a measure of the irregularity in time-series [39]. It was developed by Richman and Moorman to reduce the bias caused by self-matching in the estimation of the approximate entropy [40]. *SampEn* divides a time-series into consecutive vectors of length m. It assesses whether the maximum absolute distance between the corresponding components of each pair of vectors is less than or equal to a tolerance r, i.e., if the vectors match each other within r. If so, the vectors are considered as similar. Then the same process is repeated for vectors of length m+1 and the conditional probability that similar vectors of length m remain similar when the length is m+1 is computed. The final *SampEn* value is obtained as the negative logarithm of such a conditional probability [39], [40]. Thus, higher values of *SampEn* indicate less self-similarity in the times-series and, consequently, more irregularity [39]. SAHS is reflected in the AF signal by the addition of not regular events. As a consequence, it is expected that *SampEn* present higher values in SAHS patients.

B. Feature selection: fast correlation-based filter

The exhaustive characterization of the AF signal may lead to the extraction of several features which provide similar information about SAHS, i.e., which are redundant. Hence, a feature selection stage is included to discard those features (X_i) which share more information with the others than with a SAHS-related dependent variable, Y. The FCBF has shown its utility in previous studies involving SAHS [17], as well as other biomedical applications [21]. In our case, Y is a vector whose components are the AHI value of each subject.

FCBF relies on symmetric uncertainty (*SU*), which is a normalized quantification of the information gain (*IG*) between two variables [20]. It consists of two steps. In the first one, a relevance analysis of the features (X_i) is done. Thus, *SU* between each feature X_i and Y is computed as follows:

$$SU(X_i,Y) = 2\left[\frac{IG(X_i \mid Y)}{H(X_i)+H(Y)}\right] \quad i = 1,2,...,F, \qquad (1)$$

where $IG(X_i \mid Y) = H(X_i) - H(X_i \mid Y)$, H is the well-known

Shannon's entropy, and F is the number of features extracted (F = 12). SU is constrained to 0-1. A 0 value indicates that the two variables are independent, whereas SU = 1 indicates that knowing one feature it is possible to completely predict the other [20]. Thus, the higher the value of SU, the more information shares the corresponding feature with the AHI and, consequently, the more relevant is. Then a ranking of features is done based on their $SU(X_i, Y)$ values, i.e., from most relevant to least relevant. The second step is a redundancy analysis in which the SU between each pair of features ($SU(X_i, X_j)$) is sequentially estimated beginning from the first-ranked ones. If $SU(X_i, X_j) \geq SU(X_i, Y)$, with X_i being more highly ranked than X_j, the feature j is discarded due to redundancy and is not considered in next comparisons [20]. The optimum features are those not discarded when the algorithm ends.

C. Classification approach: boosting

After the feature selection procedure, each subject from our database is associated with a vector $\mathbf{x}_k$ (k = 1, 2,…, N, where N is the size of our sample), whose components are the values of the features included in the optimum set. The purpose is to build models with the ability to determine SAHS and its severity on the basis of the information contained in the vectors $\mathbf{x}_k$. Boosting procedures are known to achieve good generalization ability [22]. Thus, 60% of the instances are used as training set ($N_{training}$ = 191) to feed the boosting method *AdaBoost* (AB), which we use along with LDA and CART as weak classifiers (AB-LDA and AB-CART). The remaining 40% (N_{test} = 126) is used as test group to validate the models. For comparison purposes, we also train a classic logistic regression (LR) classifier.

1) AdaBoost algorithm

Boosting procedures are iterative algorithms designed to combine models that complement one another [22]. Such a combination is conducted on the basis of weighted votes from classifiers of the same type [22], [41]. AB is a widely used boosting algorithm, originally developed by Freund and Schapire [42], which can be used along with any classifier [22]. However, if AB is applied to complex classifiers, the prediction ability on new data may be significantly decreased [22], i.e., its generalization ability may be lost. Thus, simpler procedures known as weak classifiers are preferable [22]. In our case, we chose the well-known LDA and CART algorithms to act as weak classifiers.

At each m iteration, the AB algorithm assigns a weight, w_k^m, to every instance (or vector) $\mathbf{x}_k$ in the training set. Thus, the mth weak classifier is trained using the corresponding weighted instances. Then its performance is assessed through an error ε_m. This error is used to determine the weighted vote, α_m, of this mth classifier [22]. Thereby, those classifiers with smaller ε_m contribute more to the final decision. At the end of the iteration the weights of the misclassified instances are updated (w_k^{m+1}) [22]. Then, the weights of all instances are normalized in order to maintain the original distribution [42].

Two versions of AB have been implemented in this study: AB.M1, for binary classification, and AB.M2 for the multiclass task. Both of them rely on reweighting those instances which have been misclassified after each iteration. Thus, the weak classifier trained during the next iteration gives more importance to these instances [42], being more likely to classify them rightly [22]. The main difference between AB.M1 and AB.M2 is how the error ε is defined. For AB.M1 ε_m is the sum of the weights of the misclassified instances in a given iteration m, divided by the sum of the total weights of all instances at that iteration:

$$\varepsilon_m = \frac{\sum_{k=1}^{N_{training}} w_k^m(\text{miss.})}{\sum_{k=1}^{N_{training}} w_k^m}. \quad (2)$$

By contrast, a weighted pseudo-loss is defined in the case of AB.M2, for which ε_m is as follows [42]:

$$\varepsilon_m = \frac{1}{2}\cdot \sum_{k=1}^{N_{training}} \sum_{c\neq c_{true}} w_{k,c}^m \cdot (1 - h_m(\mathbf{x}_k, c_{true}) + h_m(\mathbf{x}_k, c)), \quad (3)$$

where c is a categorical variable representing the multiple classes, c_{true} refers to the actual class of $\mathbf{x}_k$, and h_m is the confidence of the prediction of the weak learner for an instance $\mathbf{x}_k$ and a class from c.

AB.M1 and AB.M2 perform the final classification task by returning the class with the highest sum of the votes from all classifiers, taking into account the weight of their corresponding predictions α_m computed as follows [42]:

$$\alpha_m = \ln(\beta_m), \quad (4)$$

where β_m is defined as (1 - ε_m)/ ε_m. Additionally, the shrinkage regularization technique has been proposed to minimize overfitting [43]. It is based on adding a learning rate υ to the iterative process by redefining β_m as $(\beta_m)^{\upsilon}$, where υ ranges 0–1 and has to be experimentally estimated.

Two criteria were used to stop the AB.M1 algorithm: *i*) ε_m does not belong to the interval (0, 0.5) [22] or *ii*) the number of weak learners is not higher than 400 (to minimize the overfitting chances). In the case of AB.M2 only the second criterion was applied since the first one is considered too restrictive for multiclass approaches [42].

D. Logistic regression and conventional approach algorithm

We also implemented LR models and a conventional event-detection algorithm to evaluate them using our own database.

LR is a widely-used supervised learning algorithm which has become a standard for binary classification tasks [44]. It estimates the posterior probability that a given instance (or vector) $\mathbf{x}_k$ belongs to one of two classes. First, the LR algorithm uses the maximum likelihood estimation of the coefficients of a linear transformation where the dependent variables are the components of each $\mathbf{x}_k$ [44], in our case, features extracted from the signals. Then the well-known logit function is applied to this linear transformation in order to obtain the above mentioned probability [44]. Vector $\mathbf{x}_k$ is then assigned to the class with the highest posterior probability.

We also implemented a conventional scoring algorithm in order to apply it to our AF recordings database. Thus, a peak detection algorithm was used to locate inspiratory onsets and endings in AF time series [45]. The difference between AF values in consecutive onsets and endings locations determined the amplitude of every inspiration. According to the rules of the AASM, the algorithm scored those respiratory events which meet with *i*) a drop of 30% or more from the AF pre-event baseline and *ii*) the drop lasts 10 seconds or more [9]. The baseline was computed as the mean amplitude of the *s* previous inspirations [27]. Hence, *s* was a design parameter. Once all events are scored, the total amount of them is divided by the sleep time to obtain an AHI estimation. To choose an optimum *s* value we computed the AHI estimations of the subjects in the training group, with *s* ranging 1-10. For each *s*, the Spearman's correlation was computed between the corresponding AHI estimations and the actual AHI from the subjects. The highest correlation was obtained for $s = 6$, which was established as the optimum value.

E. Statistical analysis

The extracted features did not pass the Lilliefors normality test. Hence, the non-parametric Kruskal-Wallis test was used to establish significant statistical differences between the four groups of SAHS severity (p-value<0.01). Bonferroni correction was applied to deal with multiple comparisons. Diagnostic ability of the AB and LR models was assessed in terms of sensitivity (Se, percentage of positive subjects rightly classified), specificity (Sp, percentage of negative subjects rightly classified), accuracy (Acc, overall percentage of subjects rightly classified), and Cohen's kappa (κ). κ measures the agreement between predicted and observed classes, avoiding the part of agreement by chance [22].

The bootstrap 0.632 algorithm [22], which was only applied to the training group, was used to find an optimum learning rate υ for the AB models. Thus, B new bootstrap training groups ($B_{training}$), with the same size as the original one, were built by resampling with replacement from this [46]. We chose $B = 500$ since it suffices for a proper estimation of the error, while let the variance remain low [46]. A uniform probability was used to select from the original instances in the training group. Consequently, some of these instances were repeated for each new $B_{training}$, whereas the same number remained unemployed. The latter were used as the corresponding bootstrap test groups (B_{test}). We evaluated υ in the range (0, 1] (step = 0.1). At each step, we computed κ^n ($n = 1, 2, \ldots, B$) as follows [22]:

$$\kappa^n = 0.369 \cdot \kappa^n_{Btraining} + 0.632 \cdot \kappa^n_{Btest}, \qquad (5)$$

where $\kappa^n_{Btraining}$ and κ^n_{Btest} are the Cohen's kappa values for each $B_{training}$ and B_{test}, respectively. Then, the 500 κ^n statistics were averaged in each step, and υ was chosen according to the highest κ averaged value.

F. Balancing the classes: SMOTE

Before training the classifiers, we applied SMOTE to compensate the imbalance among classes. SMOTE creates new synthetic instances on the basis of the available minority class real ones [26]. In our case, the real instances are the vectors of features associated to each subject in this minority class. According to the number of new instances (vectors) required for the compensation of the classes, the algorithm selects the K-nearest neighbors of each of the real ones [26]. Thus, if it is required to double the minority class vectors, K should be 1, and so on. Then, the difference between each vector and its K-nearest neighbors is computed. These differences, multiplied by a random number in the range 0 to 1, are subsequently added to the original vector again, to form new synthetic ones whose components are between the vector considered and its corresponding K-nearest neighbors [26].

As it can be derived from Table II, our instances of features, $\mathbf{x_k}$, come from: 39 no-SAHS, 92 mild-SAHS, 70 moderate-SAHS, and 116 severe-SAHS. These were divided into a training (60%) and a test set (40%). Since the training set plays the key role to avoid the bias towards majority classes [26], we adjusted its configuration to balance the classes as much as possible. Hence, although the inclusion of instances into the training set was uniformly random per class, we forced to include 29 no-SAHS, 54 mild-SAHS, 54 moderate-SAHS, and 54 severe SAHS. Then we applied SMOTE (K=1) to the instances of the no-SAHS class to create 29 additional synthetic ones. Consequently, the balanced training set was finally composed of 58 no-SAHS, 54 mild-SAHS, 54 moderate-SAHS, and 54 severe SAHS. Accordingly, the test set was composed of 10 no-SAHS, 38 mild-SAHS, 16 moderate-SAHS, and 62 severe-SAHS.

This instance distribution, carried out for the four classes, also resulted in a balanced training set for the binary classification task. Thus, it was composed of 104 SAHS-negative instances (75 real and 29 synthetic) and 116 SAHS-positive instances (all real). The test set was composed of 35 SAHS-negative instances and 91 SAHS-positive instances.

IV. RESULTS

A. Feature extraction and selection

The optimum values for ρ (*CTM*), as well as m and r (*SampEn*), were obtained by evaluating the ranges ρ [0.001, 0.1] (step=0.001), and m=1, 2 and r [0.10*SD, 0.25*SD] (step=0.05*SD), where SD is the standard deviation of the time series. In the case of ρ, the range was chosen according to the character of data [36]. Thus, values of $\rho < 0.001$ were discarded since they led to a *CTM* value ≈ 0 regardless the SAHS severity group of the subjects. Similarly, values of $\rho >$ 0.1 were also not considered since they led to CTM values = 1 for every subject. The ranges of m and r were suggested by Pincus (2001) as those which experimentally produced good entropy estimation in time series longer than 60 samples [47]. We chose those configurations ($\rho = 0.05$ for *CTM* and $m = 2$ and r = 0.1*SD for *SampEn*) for which the corresponding *CTM* and *SampEn* values showed the highest Spearman's correlation with the variable composed of the AHI measures from the subjects. We only used training data for this purpose. Table III shows the values of the extracted features for the

SAHS severity levels in the training set (mean ± SD only from the real instances), along with the corresponding p-values. Four out of the 9 spectral features (*MA*, *mA*, M_{f1}, and M_{f2}), as well as *CTM*, showed statistical significant differences among classes after the Bonferroni correction (p-value < 0.01). These spectral features showed higher values as the SAHS severity increased. An opposite tendency was shown by *CTM* values. Thus, the variability also increased with the severity of SAHS.

The FCBF was also applied to the training set (only real instances). According to FCBF, the ranking of the 12 extracted features, from higher to lower *SU* values, was: M_{f1}, *MA*, *CTM*, *mA*, M_{f2}, *WD*, *SpecEn*, *MF*, M_{f4}, *LZC*, M_{f3}, and *SampEn*. Then, *WD* was found redundant with M_{f2}; and M_{f3} with *MF*. Hence, the final FCBF optimum set was composed of 10 features, 7 from BW (M_{f1}, *MA*, *mA*, M_{f2}, *SpecEn*, *MF*, and M_{f4}) and 3 from the non-linear analysis (*CTM*, LZC, and *SampEn*).

B. Classification

1) Model selection and training

The AB binary models (AB-LDA$_2$ and AB-CART$_2$) were selected according to the optimum υ value. Fig. 3 displays the corresponding averaged κ values for each υ after the bootstrap 0.632 algorithm. As mentioned above, this procedure was only applied to the training set. The maximum values of κ for AB-LDA$_2$ and AB-CART$_2$ (0.602 and 0.713, respectively) were reached at υ = 0.1 and υ = 0.6. Then the whole original training set was used along with these υ values to train the AB-LDA$_2$ and AB-CART$_2$ models. AB-LDA$_2$ ended after 53 iterations ($\varepsilon_{54} \geq 0.5$). Hence, 53 LDA models were taken into account for the final classification task. AB-CART$_2$ reached the limit of learners established. Therefore, it was assessed in the bootstraps sets with more weak learners (500 to 1000). No improvement in κ was reached. Consequently, the weighted votes of 400 CART models were used for the classification.

For the case of the AB multiclass models (AB-LDA$_4$ and AB-CART$_4$), we optimized both the learning rate and the number of learners (up to 400) during the bootstrap procedure. Hence, for each value of υ between 0 and 1 (step=0.1) we varied the number of weak learners from 1 to 400 (step=10) in order to compute κ. Fig. 4 displays the values of κ as a function of υ and the number of weak learners. For AB-LDA$_4$ the optimum values were υ = 1 along with 110 weak learners, whereas for AB-CART$_4$, were υ = 0.8 and 160 weak learners.

2) Performance of the models

Table IV shows the diagnostic ability of the binary models (test set). The highest values for Acc and κ are shown in bold. AB-CART$_2$ outperformed the other models in Se, Acc, and κ, as well as reached the highest Sp along with LR. These results show its higher diagnostic performance. AB-LDA$_2$ also improved the results from the classic event-detection algorithm and LR. However, the latter was more specific. Additionally, AB-LDA$_2$ and AB-CART$_2$ widely improved the performance of single models based on LDA and CART (LDA$_2$ and CART$_2$). The lowest performance was reached by the algorithm based on the event-detection approach.

In the multiclass task, Table V displays the confusion matrices of each model, i.e., the model class estimation for each subject vs. their actual SAHS severity group. Notice that, since it is a binary classifier, LR was evaluated following the one vs. all strategy [41]. The overall accuracy (main diagonal) of the models and the event-detection algorithm was low in test set: event-detection 39.7%, LR 57.4%, AB-LDA$_4$ 60.3% (47.6 % in the case of a single LDA$_4$ model), and AB-CART$_4$ 57.4% (54.8 % in the case of a single CART$_4$ model). Classification of mild and moderate subjects was particularly poor for all the models. In contrast to the overall accuracy, the diagnostic performance increases when assessing the predictions of the models in each of the AHI severity cutoffs (5 e/h, 15 e/h, and 30 e/h). Table VI displays such performance for the multiclass models and the event-detection algorithm. Consistent with the overall accuracy, κ values are low. However, high diagnostic accuracies are reached by AB-LDA$_4$ and AB-CART$_4$. They outperformed LR and the event-detection algorithm in terms of Acc and κ when assessing the three AHI cutoffs. Finally, AB-LDA$_4$ widely improved the overall performance of single

TABLE III
FEATURE VALUES FOR THE SEVERITY GROUPS (MEAN ± SD)

Feat.	no-SAHS	mild	moderate	severe	p-value
MA (10^{-4})	2.012±1.091	2.854±1.460	5.148±3.134	13.736±11.360	<<0.01
mA (10^{-4})	1.359±0.729	1.849±0.930	2.903±1.294	6.225±4.498	<<0.01
$\boldsymbol{M_{f1}}$ (10^{-4})	1.670±0.912	2.296±1.131	3.900±1.886	9.400±7.295	<<0.01
$\boldsymbol{M_{f2}}$ (10^{-5})	2.140±1.424	3.193±2.428	7.418±8.268	24.864±27.774	<<0.01
$\boldsymbol{M_{f3}}$	0.190±0.540	0.259±0.512	0.149±0.619	0.429±0.689	0.19[+]
$\boldsymbol{M_{f4}}$	2.154±0.590	2.269±0.569	2.298±0.637	2.608±1.115	0.41[+]
WD	0.046±0.019	0.052±0.029	0.063±0.041	0.086±0.056	0.003[+]
MF	0.038±0.001	0.038±0.002	0.037±0.002	0.036±0.002	0.004[+]
SpecEn (10^{-1})	9.963±0.032	9.958±0.046	9.924±0.168	9.882±0.134	0.024[+]
CTM (10^{-1})	9.993±0.007	9.988±0.015	9.987±0.009	9.963±0.023	<<0.01
LZC	0.057±0.009	0.057±0.007	0.057±0.006	0.058±0.007	0.71[+]
SampEn	0.059±0.012	0.063±0.014	0.062±0.016	0.058±0.014	0.18[+]

[+]Not lower than Bonferroni correction (p-value=0.01/6).

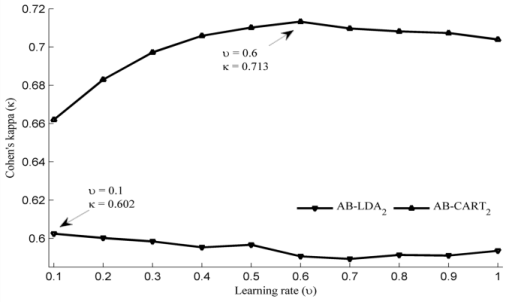

Fig. 3 Optimum υ configuration for AB-LDA$_2$ and AB-CART$_2$ (obtained after bootstrap process).

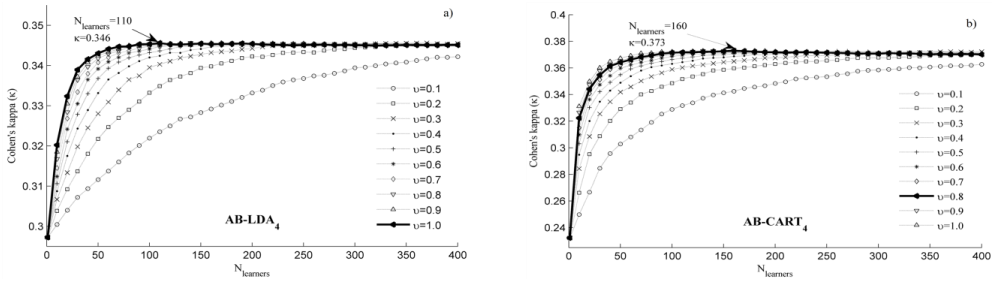

Fig. 4 Optimum υ and number of weak learners for a) AB-LDA$_4$ and b) AB-CART$_4$ (obtained after the bootstrap process).

LDA$_4$, as well as the Acc for each AHI cutoff. AB-CART$_4$ also improved the overall performance of CART$_4$, as well as the Acc for 5 e/h and 30 e/h. However, CART$_4$ outperformed the Acc of AB-CART$_4$ when considering 15 e/h as the AHI cutoff.

V. Discussion and Conclusions

In this paper, new methodologies to help in SAHS diagnosis have been proposed. Binary and multiclass AB models, composed of LDA and CART classifiers, have been evaluated to distinguish SAHS and its severity. Their performances were compared with a conventional approach (event-detection algorithm) and the classic LR classifier, both of them applied to our own database. AB outperformed these, showing high diagnostic ability.

Spectral and non-linear data, extracted from single-channel AF from NPP, were the only source of SAHS-related information used to feed the models. The spectral analysis showed significantly higher spectral power (M_{f1}) and power spectral density (*MA* and *mA*) in the 0.025-0.050 Hz. frequency band as SAHS severity increased. Since we normalized the PSD values, these measures are related to a higher occurrence of the apneic events, and not with their amplitude. This supports these features as SAHS severity dependents. Dispersion (M_{f2}) in the PSDn values at BW was also significantly higher as SAHS worsened, suggesting a more heterogeneous occurrence of apneic events throughout the frequencies within BW. Finally, the non-linear analysis showed significantly higher variability (lower *CTM* values) when SAHS severity increased. This is consistent with our initial assumption that the more severe SAHS the more changes in the respiratory pattern and, consequently, the higher variability in AF. These five features were selected by FCBF. Although M_{f2}, *SpecEn*, *MF*, *LZC*, and *SampEn* did not show discriminative power to distinguish SAHS severity, they were also automatically chosen, suggesting their usefulness by providing complementary information. Moreover, spectral and non-linear features were included in the 10-feature FCBF optimum set, which indicates that one analysis complement the other, as suggested in previous studies involving AF from thermistor [17], [18].

TABLE IV
Diagnostic Ability of the Binary Models in the Test Set

Models	Se (%)	Sp (%)	Acc (%)	κ
Event-detec.	75.8	54.3	69.0	0.286
LR	83.5	80.0	82.5	0.593
LDA$_2$	72.5	74.3	73.0	0.410
CART$_2$	85.7	68.6	81.0	0.593
AB-LDA$_2$	86.8	77.1	84.1	0.618
AB-CART$_2$	89.0	80.0	**86.5**	**0.672**

AB-CART$_2$ achieved the highest Acc and κ values for the binary (AHI cutoff = 10 e/h) classification task (86.5% Acc,

TABLE V. Confusion Matrices for each Model in the Test Set. Results from LDA and CART single models in parentheses.

		Event-detection				LR (one vs. all)				AB-LDA$_4$ *(LDA$_4$)*				AB-CART$_4$ *(CART$_4$)*			
	Estimated →	**no**	**mild**	**mod.**	**severe**	**no**	**mild**	**mod.**	**sever**	**no**	**mild**	**mod.**	**severe**	**no**	**mild**	**mod.**	**severe**
Actual	**no-SAHS**	2	4	3	1	8	0	2	0	8 *(8)*	0 *(0)*	2 *(2)*	0 *(0)*	8 *(7)*	1 *(2)*	1 *(1)*	0 *(0)*
	mild	12	16	5	5	14	8	10	6	11 *(13)*	16 *(7)*	8 *(13)*	3 *(5)*	14 *(16)*	8 *(11)*	12 *(9)*	4 *(2)*
	moderate	1	5	5	5	3	3	4	6	3 *(5)*	4 *(2)*	6 *(6)*	3 *(3)*	3 *(4)*	2 *(3)*	6 *(6)*	5 *(3)*
	severe	3	17	15	27	2	1	7	52	1 *(4)*	3 *(5)*	12 *(14)*	46 *(39)*	0 *(3)*	3 *(0)*	9 *(14)*	50 *(45)*

TABLE VI. Diagnostic Ability of the Multiclass Models in the Test Set. Results from LDA and CART single models in parentheses.

	Event-detection			LR (one vs. all)			AB-LDA$_4$ *(LDA$_4$)*			AB-CART$_4$ *(CART$_4$)*		
	5	15	30	5	15	30	5	15	30	5	15	30
Se (%)	86.2	66.7	43.5	83.6	88.5	83.9	87.1 *(81.0)*	85.9 *(79.5)*	74.2 *(62.9)*	85.3 *(82.8)*	89.7 *(87.2)*	80.6 *(72.6)*
Sp (%)	20.0	70.8	82.8	80.0	62.5	81.3	80.0 *(80.0)*	72.9 *(58.3)*	90.6 *(87.5)*	80.0 *(70.0)*	64.6 *(75.0)*	85.9 *(92.2)*
Acc (%)	81.0	68.3	63.5	83.3	78.6	82.5	**86.5** *(81.0)*	81.0 *(71.4)*	82.5 *(75.4)*	84.9 *(81.7)*	80.2 ***(82.5)***	**83.3** *(82.5)*
κ		0.152			0.370			**0.432** *(0.281)*			0.381 *(0.369)*	

0.672 κ). In the multiclass classification, AB-LDA$_4$ obtained 86.5%, 81.0%, 82.5% Acc for 5 e/h, 15 e/h, and 30 e/h, respectively, as well as κ = 0.432. It is worth noting that both AB-LDA$_4$ and AB-CART$_4$ reached high statistics when evaluating 5 e/h and 30 e/h. They outperformed the LR models, the single-model LDA and CART classifiers, as well as the event-detection algorithm. These cutoffs are particularly important. AHI = 5 e/h draws the line for the lower degree of SAHS. Furthermore, AHI = 30 e/h, which establish the boundary for the highest SAHS severity, has been associated with mortality [49], as well as suffices to recommend a treatment even in the absence of other symptoms [49]. In this regard, and according to Table V, 46 out of the 52 subjects (88.5 %) that the AB-LDA$_4$ ensemble predicted as severe-SAHS were rightly classified, whereas the remaining 6 (11.5%) were mild- or moderate-SAHS, at least. Similarly, 50 out of the 59 subjects (84.7%) that the AB-CART$_4$ ensemble predicted as severe were indeed severe, with 0 subjects from the no-SAHS group falling within this class.

Table VII summarizes performances from previous works focused on the use of single-channel AF from NPP to help in SAHS diagnosis [10], [13], [50]-[52]. All studies, except the present one, adopted an event detection approach. When assessing AHI = 10 e/h, only Wong *et al* achieved higher diagnostic performance than AB-CART$_2$ [10], [51]. However, a small sample size was used to evaluate their proposals. Nakano *et al* detected apneic events in AF with the help of spectral analysis [50]. They reported higher Se (97.0%) but lower Sp (76.0%). Unfortunately, some data about the population under study, required to complete the comparison, were not reported by the authors. None of the studies, outperformed our AB-LDA$_4$ model (86.5% Acc) in the assessment of AHI = 5 e/h. However, Nakano *et al* reported significantly higher Se (97.0%) [50]. Additionally, BaHammam *et al* and Nigro *et al* exhibited higher diagnostic ability when assessing AHI = 30 e/h [13], [52]. Nonetheless, their databases were composed of 95 and 90 subjects, respectively, in contrast to the 317 subjects involved in our study. Finally, all the studies performed similarly to our AB-LDA$_4$ model (81.0% Acc) when evaluating AHI = 15 e/h.

Despite we have shown the utility of our proposal, some limitations need to be addressed. Although our sample is large (317 subjects), analyzing more recordings would enhance the statistical power of our results. Particularly, a more balanced proportion of the classes would be desirable for the sake of the model training. Nonetheless, our sample reflects a realistic proportion among the people who undergo the PSG test. Additionally, we applied the SMOTE technique to our data in order to compensate the imbalance. The single use of NPP to acquire AF may be another limitation. The AASM recommends using both NPP and thermistor for a proper quantification of the number of apneas and hypopneas [9]. However, our proposal does not rely on a classic event-detection approach. In this regard, previous studies of our research group showed high diagnostic ability when evaluating data from single-channel AF acquired through a thermistor [17], [18]. Our current proposal has shown that using AF data from NPP is also possible in order to reach a high diagnostic performance. Another limitation arises regarding the redundant information removed by the FCBF algorithm. The features discarded share more information with the selected ones than with the AHI. However, the features selected might still share information with the others to some extent. The training time of the AdaBoost models is another limitation if we compare it with simpler methodologies such as logistic regression. However, once the models are trained, the runtime after they are applied to new data is trivial. Finally, since we propose an automatic procedure with potential to reach diagnosis in few minutes after data collection, it would be of great interest if future works could address the assessment of our methodology embedded in a diagnostic test at patient's home. It would be also interesting the implementation and assessment of a multiclass logistic-regression based AdaBoost algorithm.

TABLE VII
COMPARISON WITH THE STATE OF THE ART OF SINGLE-CHANNEL AF FROM NPP

Studies	Subjects	AHI cutoff	Se (%)	Sp (%)	Acc (%)
[a]De Almeida *et al* [10]	30	10	85.7	87.5	nd
[a]Nakano *et al* [50]	217	5	97.0	77.0	nd
		10	97.0	76.0	nd
		15	97.0	73.0	nd
[a]Wong *et al* [51]	33	10	92.0	86.0	90.9*
		30	91.0	75.0	81.5*
[a]BaHammam *et al* [52]	95	5	79.0	68.0	76.8*
		10	70.0	89.0	77.9*
		15	65.0	94.0	81.8*
		30	63.0	98.0	83.2*
[a]Nigro *et al* [13]	90	5	89.3	60.0	84.4*
		10	80.4	82.3	nd
		15	76.7	83.0	80.0*
		30	88.5	95.3	93.3*
[b]AB-CART$_2$	317	10	89.0	80.0	86.5
[b]AB-LDA$_4$	317	5	87.1	80.0	86.5
		15	85.9	72.9	81.0
		30	74.2	90.6	82.5

[a]Event detection approach; [b]Direct subject classification approach; *Computed from reported data; nd: Not enough data to estimate.

To the best of our knowledge, this is the first time that the AB algorithm is used along with spectral and nonlinear features from single-channel AF to help in SAHS diagnosis. Our AB proposals for binary and multiclass classification outperformed the classic LR as well as a conventional event-detection algorithm, both of them applied to our own database. The new AB-CART$_2$ and AB-LDA$_4$ models achieved high diagnostic ability compared with the state of the art. Additionally, we showed that it is possible to achieve high diagnostic ability by the use of spectral and nonlinear data from NPP AF. These results highlight the usefulness of our proposal when detecting SAHS and its severity.

REFERENCES

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

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

[3] S. P. Patil, et al, "Adult Obstructive Apnea," *Chest*, vol. 132, pp. 325-337, 2007.
[4] A. Sassani, et al, "Reducing Motor-Vehicle Colissions, Costs, and Fatalities by Treating Obstructive Sleep Apnea Syndrome," *Sleep*, vol. 27, pp. 453-458, 2003.
[5] E. Lindberg et al, "Role of Snoring and Daytime Sleepiness in Occupational Accidents," *Am. J. Respir. Crit. Care Med.*, vol. 164, pp. 2031-2035, 2001.
[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. 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. 958-959, 1999.
[8] W. W. Flemons et al, "Home Diagnosis of Sleep Apnea: A Systematic Review of the Literature," *Chest*, vol. 124, pp. 1543-1579, 2003.
[9] 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.
[10] F. R. de Almeida et al, "Nasal pressure recordings to detect obstructive sleep apnea," *Sleep Breath*, vol. 10, pp. 62-69, 2006.
[11] U. J. Magalang et al, "Prediction of the apnea-hypopnea index from overnight pulse oximetry," *Chest*, vol. 124, pp. 1694-1701, 2003.
[12] T. Penzel et al, "Systematic Comparison of Different Algorithms for Apnoea Detection Based on Electrocardiogram Recordings," *Med. Biol. Eng. Comput.*, vol. 40, pp. 402-407, 2002.
[13] C. A. Nigro et al, "Comparison of the automatic analysis versus the manual scoring from ApneaLink™ device for the diagnosis of obstructive sleep apnoea syndrome," *Sleep Breath* vol. 15, pp. 679-686, 2011.
[14] C. Gómez et al, "Complexity analysis of the magnetoencephalogram background activity in Alzheimer's disease patients," *Med. Eng. Phys.*, vol. 28, pp. 851-59, 2006.
[15] 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. 2816–2824, 2010.
[16] 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.
[17] G. C. Gutiérrez-Tobal et al., "Pattern recognition in airflow recordings to assist in the sleep apnoea–hypopnoea syndrome diagnosis," *Med. Biol. Eng. Comput.*, vol. 51, pp. 1367-80, 2013.
[18] G. C. Gutiérrez-Tobal et al, "Linear and nonlinear analysis of airflow recordings to help in sleep apnoea-hypopnoea syndrome diagnosis," *Physiol. Meas.*, vol. 33, pp. 1261-75, 2012.
[19] R. Hornero et al, "Spectral and nonlinear analyses of MEG background activity in patients with Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 55, pp. 1658-1665, 2008.
[20] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205-1224, 2004.
[21] A. Aarabi et al, "Automated neonatal seizure detection: A multistage classification system through feature selection based on relevancy and redundancy analysis," *Clin. Neurophysiol.*, vol. 117, pp. 328-340, 2006.
[22] I. H. Witten, E. Frank and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann/Elsevier, 2011.
[23] C. Morgenstern et al, "Assessment of changes in upper airway obstruction by automatic identification of inspiratory flow limitation during sleep," *IEEE Trans. Biomed. Eng.*, vol. 56, pp. 2006-2015, 2009.
[24] B. Xie and H. Minn, "Real-time sleep apnea detection by classifier combination. Information Technology in Biomedicine," *IEEE Trans. Biomed. Eng.*, vol. 16, pp. 469-477, 2012.
[25] 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.
[26] N. V. Chawla et al, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16(1), pp. 321-357, 2002.
[27] J. Han et al, "Detection of apnoeic events from single channel nasal airflow using 2nd derivative method," *Comput. Meth. Prog. Bio.*, vol. 98, pp. 199-207, 2008.
[28] A. Qureshi et al, "Obstructive sleep apnea," *J. Allergy Clin. Immunol.*, vol. 112, pp. 643-651, 2003.
[29] A. S. Karunajeewa et al, "Multi-feature snore sound analysis in obstructive sleep apnea–hypopnea syndrome," *Physiol. Meas.*, vol. 32(1), pp. 83, 2011.
[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] L. Sörnmo and P. Laguna, *Bioelectrical signal processing in cardiac and neurological applications*. London, U.K./New York: Elsevier/Academic; 2005.
[32] D. J. Eckert and A. Malhotra, "Pathophysiology of adult obstructive sleep apnea," *Proc. Am. Thorac. Soc.*, vol. 5, pp. 144-153, 2008.
[33] J. Poza et al, "Extraction of spectral based measures from MEG background oscillations in Alzheimer's disease," *Med. Eng. Phys.*, vol. 29, pp. 1073-1083, 2007.
[34] W. K. Wootters, "Statistical distance and Hilbert space," *Physical Review D*, vol. 23(2), pp. 357-362, 1981.
[35] M. T. Martin et al, "Statistical complexity and disequilibrium," *Physics Letters A*, vol. 311(2), pp. 126-132, 2003.
[36] M. E. Cohen et al, "Applying continuous chaotic modeling to cardiac signal analysis," *IEEE Eng. Med. Biol. Mag.*, vol. 15, pp. 97-102, 1996.
[37] D. Abásolo et al, "Analysis of EEG background activity in Alzheimer's disease patients with Lempel–Ziv complexity and central tendency measure," *Med. Eng. Phys.*, vol. 28(4), pp. 315-322, 2006.
[38] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. 24, pp. 530-536, 1978.
[39] D. Abásolo et al, "Entropy analysis of the EEG background activity in Alzheimer's disease patients." *Phys.Mmeas.*, vol. 27, pp. 241-253, 2006.
[40] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Am. J. of Physiol-Heart C.*, vol. 278, pp. H2039-H2049.
[41] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.
[42] 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.
[43] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, 1189-1232, 2001.
[44] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York, NY: John Wiley and Sons, 2000.
[45] J. B. Korten and G. G. Haddad, "Respiratory waveform pattern recognition using digital techniques," *Comput. Biol. Med.*, vol. 19, pp. 207-217, 1989.
[46] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.
[47] S. M. Pincus, "Assessing serial irregularity and its implications for health," *Ann. NY. Acad. Sci.*, vol. 954, pp. 245–267, 2001.
[48] N. M. Punjabi et al, "Sleep-disordered breathing and mortality: a prospective cohort study," *PLoS medicine*, vol. 6(8), pp. e1000132, 2009.
[49] P. Lloberes et al, "Diagnosis and treatment of sleep apnea-hypopnea syndrome," *Arch. Bronconeumol.* ((English Edition)), vol. 47(3), pp. 143-156, 2011.
[50] H. Nakano et al, "Automatic Detection of Sleep-disordered breathing from a single-channel airflow record," *Eur. Respir. J.*, vol. 29, pp. 728-736, 2007.
[51] K. K. Wong et al, "Diagnostic test evaluation of a nasal flow monitor for obstructive sleep apnea detection in sleep apnea research," *Behavior research methods*, vol. 40(1), pp. 360-366, 2008.
[52] A. BaHammam et al, "Evaluation of the accuracy of manual and automatic scoring of a single airflow channel in patients with a high probability of obstructive sleep apnea," *Med. Sci. Mon.*, vol. 17, pp. MT13-MT19, 2011.

Appendix B: scientific production during the study

Papers indexed in the Journal Citation Reports

1. **Gutiérrez-Tobal, G. C.**, Álvarez, D., del Campo, F., & Hornero, R. (2015). Utility of AdaBoost to Detect Sleep Apnea-Hypopnea Syndrome from Single-Channel Airflow. *IEEE Transactions on Biomedical Engineering*, *In Press*. Accepted August 2015. Impact Factor: 2.347.

2. Álvarez, D., **Gutiérrez-Tobal, G. C.**, del Campo, F., & Hornero, R. (2015). Positive airway pressure and electrical stimulation methods for obstructive sleep apnea treatment: a patent review (2005-2014). *Expert opinion on therapeutic patents*, 25 (9), 971-989. Impact Factor: 4.297.

3. **Gutiérrez-Tobal, G. C.**, Alonso-Álvarez, M. L., Álvarez, D., del Campo, F., Terán-Santos, J., & Hornero, R. (2015). Diagnosis of pediatric obstructive sleep apnea: Preliminary findings using automatic analysis of airflow and oximetry recordings obtained at patients' home. *Biomedical Signal Processing and Control*, 18, 401-407. Impact Factor: 1.419.

4. **Gutiérrez-Tobal, G. C.**, Álvarez, D., Gomez-Pilar, J., del Campo, F., & Hornero, R. (2015). Assessment of Time and Frequency Domain Entropies to Detect Sleep Apnoea in Heart Rate Variability Recordings from Men and Women. *Entropy*, 17(1), 123-141. Impact Factor: 1.502.

5. **Gutiérrez-Tobal, G. C.**, Álvarez, D., Marcos, J. V., Del Campo, F., & Hornero, R. (2013). Pattern recognition in airflow recordings to assist in the sleep apnoea-hypopnoea syndrome diagnosis. *Medical & biological engineering & computing*, 51(12), 1367-1380. Impact Factor: 1.500.

6. **Gutiérrez-Tobal, G. C.**, Hornero, R., Álvarez, D., Marcos, J. V., & del Campo, F. (2012). Linear and nonlinear analysis of airflow recordings to help in sleep apnoea-hypopnoea syndrome diagnosis. *Physiological measurement*, 33(7), 1261. Impact Factor: 1.496.

Book chapters

1. Marcos, J. V., Hornero, R., Álvarez, D., **Gutiérrez-Tobal, G. C.**, del Campo, F. (2013). Automatic estimation of the apnea-hypopnea index from oxygen saturation recordings using non-linear regression analysis. *Nonlinear analysis research in biomedical engineering*. pp. 43-68. Nova Science Publishers Inc.

International conferences

1. Crespo, A., del Campo, F., Hornero, R., Álvarez, D., Terán-Santos, J., **Gutiérrez-Tobal, G. C.**, Alonso-Álvarez, M. L. Usefulness of linear discriminant analysis of overnight oximetry from respiratory polygraphyat home to assist in the diagnosis of pediatric obstructive sleep apnea hypopnea syndrome, *European Respiratory Society International Congress 2015*, September 26-30, Amsterdam, The Netherlands.

2. **Gutiérrez-Tobal, G. C.**, Kheirandish-Gozal, L., Álvarez, D., Crespo, A., Philby, M. F., Mohammadi, M., del Campo, F., Gozal, D., Hornero, R. Analysis and Classification of Oximetry Recordings to Predict Obstructive

Sleep Apnea Severity in Children. *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference*, EMBC 2015, August 25-28, Milan, Italy.

3. Álvarez, D., Kheirandish-Gozal, L., **Gutiérrez-Tobal, G. C.**, Crespo, A., Philby, M. F., Mohammadi, M., del Campo, F., Gozal, D., Hornero, R. Automated Analysis of Nocturnal Oximetry as Screening Tool for Childhood Obstructive Sleep Apnea-Hypopnea Syndrome. *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference*, EMBC 2015, August 25-28, Milan, Italy.

4. **Gutiérrez-Tobal, G. C.**, Álvarez, D., Alonso-Álvarez, M. L., Terán-Santos, J., del Campo, F., & Hornero, R. (2015). Exploring the Spectral Information of Airflow Recordings to Help in Pediatric Obstructive Sleep Apnea-Hypopnea Syndrome Diagnosis. *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference*, EMBC 2014, August 26-30, Chicago, USA.

5. Álvarez, D., **Gutiérrez-Tobal, G. C.**, Alonso-Álvarez, M. L., Terán-Santos, J., del Campo, F., Hornero, R. (2015). Statistical and Nonlinear Analysis of Oximetry from Respiratory Polygraphy to Assist in the Diagnosis of Sleep Apnea in Children. *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference*, EMBC 2014, August 26-30, Chicago, USA.

6. **Gutiérrez-Tobal, G. C.**, Álvarez, D., Gomez-Pilar, J., del Campo, F., Hornero, R. AdaBoost Classification to Detect Sleep Apnea from Airflow Recordings. *IFMBE Proceedings of the XIII Mediterranean Conference on Biomedical and Biological Engineering and Computing*, MEDICON 2013, September 25-28, Sevilla, Spain.

7. Álvarez, D., **Gutiérrez-Tobal, G. C.**, Gomez-Pilar, J., del Campo, F., Hornero, R. Applying variable ranking to oximetric recordings in sleep apnea diagnosis. *IFMBE Proceedings of the XIII Mediterranean Conference on Biomedical and Biological Engineering and Computing*, MEDICON 2013, September 25-28, Sevilla, Spain.

8. Gomez-Pilar, J., **Gutiérrez-Tobal, G. C.**, Álvarez, D., del Campo, F., Hornero, R. Classification Methods from Heart Rate Variability to Assist in SAHS Diagnosis. *IFMBE Proceedings of the XIII Mediterranean Conference on Biomedical and Biological Engineering and Computing*, MEDICON 2013, September 25-28, Sevilla, Spain.

9. **Gutiérrez-Tobal, G. C.**, Álvarez, D., Gomez-Pilar, J., del Campo, F., Hornero, R. Assessment of Spectral Bands of Interest in Airflow Signal to Assist in Sleep Apnea-Hypopnea Syndrome Diagnosis. *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference*, EMBC 2013, July 3-7, Osaka, Japan.

10. **Gutiérrez-Tobal, G. C.**, Hornero, R., Álvarez, D., Marcos, J. V., Gómez, C., del Campo, F. Apnea-hypopnea index estimation from spectral analysis of airflow recordings. *Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference*, EMBC 2012, August 28 - September 1, San Diego, USA.

11. Álvarez, D., **Gutiérrez-Tobal, G. C.**, Marcos, J. V., Campo, F. D., Hornero, R. Spectral analysis of single-channel airflow and oxygen saturation recordings in obstructive sleep apnea detection. *Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference*, EMBC 2011, August 30 - September 3, Buenos Aires, Argentina.

National conferences

1. Crespo, A., del Campo, F., Juez, L., López, G., Álvaro, T., de Frutos, J., Arroyo, A., Ruiz, T., Álvarez, D., **Gutiérrez-Tobal, G. C.**, Hornero, R. Eficacia diagnóstica de la pulsioximetría nocturna en el ámbito domiciliario frente al hospitalario en pacientes con sospecha de síndrome de apnea hipopnea del sueño. *Actas del XXXIV Congreso de la Sociedad Castellano-leonesa y Cántabra de Patología Respiratoria*, SOCALPAR 2015, May 15 - 16, Salamanca, Spain.

2. Crespo, A., Alonso-Álvarez, M. L., Hornero, R., Álvarez, D., Terán-Santos, J., **Gutiérrez-Tobal, G. C.**, del Campo, F. Análisis automático de la señal de oximetría en el domicilio del paciente como método de ayuda en el diagnóstico del síndrome de apnea obstructiva del sueño en niños. *Actas de la XXIII Reunión Anual de la Sociedad Española del Sueño*, March 12 - 14, Lérida, Spain.

3. Álvarez, D., **Gutiérrez-Tobal, G. C.**, Alonso-Álvarez, M. L., Terán-Santos, J., del Campo, F., & Hornero, R. Análisis espectral y no lineal de la señal de oximetría domiciliaria en la ayuda al diagnóstico de la apnea infantil. *Actas del XXXII Congreso Anual de la Sociedad Española de Ingneiería Biomédica*, CASEIB 2014, November 26 - 28, Barcelona, Spain.

4. Crespo, A., del Campo, F., Álvarez, D., de Frutos, J., Arroyo, A., Ruiz, T., **Gutiérrez-Tobal, G. C.**, Hornero, R. Estudio comparativo entre la oximetría domiciliaria versus hospitalaria en pacientes con trastornos del sueño. *Actas del XXXIIII Congreso de la Sociedad Castellano-leonesa y Cántabra de Patología Respiratoria*, SOCALPAR 2014, May 9 - 10, Palencia, Spain.

5. Juez, L., Crespo, A., Arroyo, A., Hornero, R., Álvarez, D., **Gutiérrez-Tobal, G. C.**, Ruiz, T., López, G., de Frutos, J., Sánchez, A., Andrés, A., Moche, J. A., Tijero, B., del Campo, F. Trastornos respiratorios del sueño y obesidad. *Actas del XXXIIII Congreso de la Sociedad Castellano-leonesa y Cántabra de Patología Respiratoria*, SOCALPAR 2014, May 9 - 10, Palencia, Spain.

6. Crespo, A., del Campo, F., Álvarez, D., de Frutos, J., Arroyo, A., Ruiz, T., **Gutiérrez-Tobal, G. C.**, Hornero, R. Estudio del análisis espectral de la señal de flujo aéreo como método de ayuda en el diagnóstico del síndrome de apnea hipopnea del sueño. *Actas de la XXII Reunión Anual de la Sociedad Española del Sueño*, April 2 - 5, San Sebastián, Spain.

7. del Campo, F., **Gutiérrez-Tobal, G. C.**, Álvarez, D., Crespo, A., Hornero, R. Estudio comparativo entre la oximetría domiciliaria versus hospitalaria en pacientes con sospecha de síndrome de apnea obstructiva del sueño. *Actas*

del 46 Congreso Nacional de la Sociedad Española de Neumología y Cirugía Torácica, SEPAR 2013, June 14 - 17, Barcelona, Spain.

8. del Campo, F., Hornero, R., Gomez-Pilar, J., Álvarez, D., **Gutiérrez-Tobal, G. C.**. Utilidad diagnóstica de la variabilidad de la frecuencia cardiaca en pacientes con sospecha de SAHS. *Actas del 46 Congreso Nacional de la Sociedad Española de Neumología y Cirugía Torácica*, SEPAR 2013, June 14 - 17, Barcelona, Spain.

9. **Gutiérrez-Tobal, G. C.**, Gomez-Pilar, J., Álvarez, D., del Campo, F., Hornero, R. Evaluación de bandas espectrales de interés en la señal de flujo aéreo para ayudar en el diagnóstico del síndrome de la apnea hipopnea del sueño. *Actas del XXX Congreso Anual de la Sociedad Española de Ingneiería Biomédica*, CASEIB 2012, November 19 - 21, San Sebastián, Spain.

10. Gomez-Pilar, J., **Gutiérrez-Tobal, G. C.**, Álvarez, D., del Campo, F., Hornero, R. Extracción y selección de características de la señal de variabilidad del ritmo cardiaco para la ayuda al diagnóstico del síndrome de la apnea hipopnea del sueño. *Actas del XXX Congreso Anual de la Sociedad Española de Ingneiería Biomédica*, CASEIB 2012, November 19 - 21, San Sebastián, Spain.

11. **Gutiérrez-Tobal, G. C.**, Hornero, R., Álvarez, D., Marcos, J. V., del Campo, F. Selección de características espectrales procedentes de la señal de flujo aéreo en la ayuda al diagnóstico del síndrome de la apnea hipopnea del sueño. *Actas del XXIX Congreso Anual de la Sociedad Española de Ingneiería Biomédica*, CASEIB 2011, November 16 - 18, Cáceres, Spain.

12. **Gutiérrez-Tobal, G. C.**, Hornero, R., Álvarez, D., del Campo, F. Análisis espectral de la señal de flujo aéreo para la ayuda en el diagnóstico del síndrome de la apnea obstructiva del sueño. *Actas del XXVIII Congreso Anual de la Sociedad Española de Ingneiería Biomédica*, CASEIB 2010, November 24 - 26, Madrid, Spain.

13. Álvarez, D., Hornero, R., **Gutiérrez-Tobal, G. C.**, Marcos, J. V., del Campo, F. Análisis de las variaciones en la saturación de oxígeno y flujo aéreo en la ayuda al diagnóstico de la apnea del sueño. *Actas del XXVIII Congreso Anual de la Sociedad Española de Ingneiería Biomédica*, CASEIB 2010, November 24 - 26, Madrid, Spain.

Appendix C: resumen en castellano

C.1. Introducción: problemática del síndrome de la apnea-hipopnea del sueño

El síndrome de la apnea-hipopnea del sueño (SAHS) es una enfermedad que se caracteriza por la aparición de eventos de cese total (apneas) y disminución significativa (hipopneas) de la respiración durante el sueño. La recurrencia de apneas e hipopneas conduce a una ventilación deficiente caracterizada por hipoxia e hipercapnia, que a su vez producen descensos en la saturación de oxígeno en sangre (desaturaciones), microdespertares periódicos y, por tanto, fragmentación del sueño. Como consecuencia, los pacientes de SAHS no son capaces de conseguir un sueño reparador, lo que afecta enormemente a su calidad de vida. Hipersomnolencia diurna, dificultad para la concentración, disminución de la memoria a corto plazo y depresión son algunos de los sintomas diurnos presentes en los pacientes de SAHS. Además, el SAHS está asociado con graves patologías cardiovasculares y metabólicas como el fallo cardiaco, accidentes cerebrovasculares, la muerte súbita y la diabetes. Recientemente, el SAHS también ha sido asociado con un aumento en la incidencia del cáncer. Es por ello que un diagnóstico rápido resulta fundamental para la mejora de la salud y la calidad de vida de los pacientes.

El SAHS es una enfermedad muy prevalente, estimándose que la padecen entre el 2 % y el 7 % de la población adulta, y hasta un 6 % de los niños. Además, se considera una afección muy infradiagnosticada, con una incidencia creciente asociada a la epidemia de obesidad presente en los paises desarrollados. La polisomnografía nocturna (PSG), realizada en una unidad del sueño especializada, es la prueba estándar para diagnosticar el SAHS. Sin embargo, esta prueba resulta compleja técnicamente, debido al alto número de señales fisiológicas que deben registrarse, costosa económicamente, por el gasto derivado de la hospitalización de los pacientes, y requiere mucho tiempo de análisis posterior para alcanzar un diagnóstico. Éste se obtiene mediante el cálculo del índice de apnea-hipopnea (*apnea-hypopnea index*, AHI), tras inspeccionar los registros polisomnográficos. Además, la PSG priva al paciente de dormir en su entorno de sueño habitual.

C.2. Alternativas a la polisomnografía

Las limitaciones de la PSG, junto con la alta prevalencia de la enfermedad y la insuficiente disponibilidad de instalaciones especializadas, han llevado a la búsqueda de formas de simplificar el proceso de diagnóstico. Reducir su complejidad es el factor clave para disminuir el coste asociado, la incomodidad de los pacientes, y el tiempo necesario para alcanzar el diagnóstico. Además, permitiría desarrollar estrategias que incluyeran la utilización de dispositivos portátiles en el domicilio de los pacientes. Un enfoque habitual es el análisis de un grupo reducido de señales de entre aquellas involucradas en la PSG. De acuerdo con el número y tipo de señales analizadas los dispositivos utilizados en los estudios del sueño se clasifican del Tipo 1 al 4, siendo el primero la PSG convencional, y el 4 un dispositivo que registra solamente 1 o 2 canales, frecuentemente el flujo aéreo (FA) y/o la desaturación de oxígeno en sangre (SpO_2). Por otro lado, un enfoque muy frecuente de los estudios dirigidos a la simplificación de la prueba de diagnóstico es la detección de cada uno de los eventos apneicos. En esta Tesis Doctoral se plantea el análisis automático de la señal del FA monocanal como alternativa simple y fiable a la PSG. Además, se

propone el reconocimiento de patrones como principal técnica para el diagnóstico automático del SAHS, incluyendo tanto su detección (clasificación binaria) como la determinación de su severidad (clasificación multiclase y estimación del AHI mediante regresión).

C.3. Hipótesis y objetivos

En la presente Tesis Doctoral se trabaja bajo la **hipótesis** general de que *es posible reducir la complejidad de la prueba diagnóstica del SAHS mediante el reconocimiento de patrones automático en el FA*. De acuerdo con dicha hipótesis, el **objetivo** principal de la tesis es el análisis exhaustivo de la señal de FA monocanal y la posterior evaluación de su capacidad diagnóstica. Para llevarlo a cabo se han definido varios **objetivos específicos**:

I. Construir una base de datos con señales FA procedentes de sujetos sospechosos de padecer SAHS.

II. Revisar el estado del arte de métodos de procesado automático de señales fisiológicas, especialmente aquellos relacionados con la extracción y selección de características, así como con el reconocimiento de patrones.

III. Seleccionar e implementar aquellos métodos que, de acuerdo con el estado del arte, son más apropiados para su utilización en la ayuda al diagnóstico del SAHS.

IV. Procesar las señales mediante los métodos implementados anteriormente.

V. Realizar análisis estadísticos de los resultados obtenidos para evaluar la idoneidad de cada metodología aplicada a los registros, así como llevar a cabo una evaluación del rendimiento general de la propuesta realizada.

VI. Comparar y discutir los resultados para extraer las conclusiones apropiadas. Este objetivo incluye la comparación con los estudios del estado del arte, la implementación de otros métodos clásicos en nuestra base de datos, así como la aplicación de nuestra metodología a otras señales de referencia ampliamente estudiadas.

VII. Publicar los principales resultados y conclusiones obtenidos en revistas de impacto indexadas en el *Journal Citation Reports*.

C.4. Materiales

Durante la investigación se han analizado 4 bases de datos (3 de sujetos adultos y 1 de sujetos pediátricos), todas ellas provenientes de sujetos sospechosos de padecer SAHS. La primera base de datos la forman 148 registros de FA adquiridos con un termistor. La segunda base de datos está formada por 317 registros de FA obtenidos con una sonda de presión. Además, una tercera base de datos la forman 188 señales de variabilidad de la frecuencia cardiaca (*heart rate variability*, HRV). Por último, la última base de datos la forman registros de FA (termistor) y SpO_2 provenientes de 50 niños.

Los registros de sujetos adultos fueron adquiridos en la unidad del sueño del Hospital Universitario Rio Hortega de Valladolid (HURH), mientras que

los correspondientes a sujetos pediátricos se obtuvieron en el domicilio de los pacientes como parte de las investigaciones de la unidad de desórdenes respiratorios del sueño del Hospital Universitario de Burgos (HUBU). Los especialistas médicos establecieron el diagnóstico de cada paciente en base al AHI obtenido de la PSG. Tanto en el caso de los niños como en el de los adultos, se siguieron las reglas de la Academia Americana de Medicina del Sueño (*American Academy of Sleep Medicine*, AASM) para detectar los eventos apneicos [18]. Para determinar tanto el SAHS como su severidad se utilizan umbrales de AHI comunes como 5, 10, 15 y 30 eventos/hora (e/h) [18, 35, 91]. El umbral de 10 e/h ha sido ampliamente utilizado para determinar la presencia o ausencia del SAHS [35, 74, 91]. Además, los grados de severidad del SAHS se definen como: no SAHS (AHI$<$5), SAHS leve (5$\leq$AHI$<$15), SAHS moderado (15$\leq$AHI$<$30) y SAHS severo (AHI$\geq$30) [101]. En sujetos pediátricos, AHI = 3 e/h es un umbral común para establecer la presencia del SAHS [6]. Todos los sujetos adultos, así como los responsables legales de los sujetos pediátricos, dieron su consentimiento informado para participar en el estudio. Los comités éticos del HURH y del HUBU aceptaron los correspondientes protocolos para llevarlo a cabo. Las siguientes tablas muestran los datos demográficos y clínicos de los sujetos involucrados en las 4 bases de datos, incluyendo edad, índice de masa corporal (IMC) y porcentaje de sujetos masculinos. Los datos se presentan divididos en SAHS-negativo (sujetos sin SAHS) y SAHS-positivo (sujetos con SAHS) de acuerdo con los umbrales típicos en adultos (AHI = 10 e/h) y niños (AHI=3 e/h).

C.5. Métodos

La metodología empleada para llevar a cabo estos objetivos se basa en tres etapas fundamentales. La primera de ellas es la **extracción de características**, utilizada para obtener información sobre el SAHS en los registros de FA. Las señales fisiológicas se caracterizan por tener tanto comportamientos deterministas como caóticos. Por ello, se han utilizado métodos de extracción de características procedentes de distintos ámbitos como el análisis espectral y el no lineal. El objetivo es que estos enfoques permitan una óptima caracterización del SAHS mediante la obtención de información que se complemente entre sí. La segunda etapa es la **selección automática de características**. El exhaustivo análisis realizado en la etapa anterior puede llevar a la extracción de características no útiles para el diagnóstico del SAHS o que comparten la misma información que las demás. Por ello, se ha implementado una etapa de selección

Tabla C1: Datos demográficos y clínicos de los sujetos de la base de datos de adultos de **FA** (señales obtenidas mediante **termistor**), divididos en los grupos SAHS-negativo y SAHS-positivo (media $\pm$ desviación típica). IMC: índice de masa corporal. AHI: índice de apnea hipopnea.

	Todos	**SAHS-negativo**	**SAHS-positivo**
Sujetos (n)	148	48	100
Hombres (n)	117(79,0 %)	32(66,7 %)	85(85,0 %)
Edad (años)	$50{,}9 \pm 11{,}7$	$48{,}8 \pm 12{,}1$	$51{,}9 \pm 11{,}4$
IMC (Kg/m^2)	$29{,}1 \pm 4{,}6$	$27{,}6 \pm 4{,}9$	$29{,}9 \pm 4{,}7$
AHI (e/h)	–	$4{,}0 \pm 2{,}4$	$32{,}9 \pm 24{,}3$

Tabla C2: Datos demográficos y clínicos de los sujetos de la base de datos de adultos de **FA** (señales obtenidas mediante **sonda de presión**), divididos en los grupos SAHS-negativo y SAHS-positivo (media ± desviación típica). IMC: índice de masa corporal. AHI: índice de apnea hipopnea.

	Todos	**SAHS-negativo**	**SAHS-positivo**
Sujetos (n)	317	110	207
Hombres (n)	226(71,3 %)	68(61,8 %)	158(76,3 %)
Edad (años)	$49{,}9 \pm 12{,}0$	$47{,}6 \pm 12{,}9$	$51{,}1 \pm 11{,}4$
IMC (Kg/m^2)	$28{,}1 \pm 5{,}2$	$26{,}5 \pm 5{,}0$	$29{,}0 \pm 5{,}1$
AHI (e/h)	–	$6{,}0 \pm 2{,}6$	$39{,}9 \pm 25{,}9$

Tabla C3: Datos demográficos y clínicos de los sujetos de la base de datos de niños de **FA** (señales obtenidas mediante **termistor**), divididos en los grupos SAHS-negativo y SAHS-positivo (media ± desviación típica). IMC: índice de masa corporal. AHI: índice de apnea hipopnea.

	Todos	**SAHS-negativo**	**SAHS-positivo**
Sujetos (n)	50	24	26
Hombres (n)	27(54,0 %)	11(45,8 %)	16(61,5 %)
Edad (años)	$5{,}3 \pm 2{,}5$	$5{,}2 \pm 2{,}4$	$5{,}4 \pm 2{,}7$
IMC (Kg/m^2)	$16{,}5 \pm 2{,}5$	$16{,}1 \pm 1{,}7$	$16{,}9 \pm 3{,}0$
AHI (e/h)	–	$1{,}3 \pm 0{,}8$	$17{,}9 \pm 15{,}4$

Tabla C4: Datos demográficos y clínicos de los sujetos de la base de datos de adultos de **HRV** divididos en los grupos SAHS-negativo y SAHS-positivo (media ± desviación típica). IMC: índice de masa corporal. AHI: índice de apnea hipopnea.

	Todos	**SAHS-negativo**	**SAHS-positivo**
Sujetos (n)	188	69	119
Hombres (n)	134(71,3 %)	41(59,4 %)	93(78,2 %)
Edad (años)	$50{,}7 \pm 12{,}0$	$47{,}3 \pm 11{,}5$	$52{,}7 \pm 12{,}3$
IMC(Kg/m^2)	$28{,}7 \pm 4{,}7$	$28{,}0 \pm 6{,}1$	$29{,}1 \pm 3{,}7$
AHI (e/h)	–	$3{,}8 \pm 2{,}4$	$33{,}0 \pm 22{,}9$

de características que tiene como objetivo eliminar aquellas que son no relevantes o redundantes. Para ello se han empleado dos enfoques distintos. El primero es el conocido algoritmo *forward-selection backward-elimination* (SLR-FSBE), que está íntimamente relacionado con el clasificador regresión logística. Se trata por tanto de un método *wrapper*. El segundo es independiente del método de reconocimiento de patrones aplicado posteriormente. Es por tanto un método de filtrado (fast correlation-based filter, FCBF). Finalmente, la tercera etapa es la de **reconocimiento de patrones**. En esta Tesis Dcotoral se ha utilizado para obtener un diagnóstico automático mediante la aplicación de diferentes métodos de clasificación y regresión a los datos obtenidos y seleccionados en etapas anteriores. El objetivo de esta estapa ha sido la determinación de la presencia o ausencia del SAHS (clasificación binaria), la clasificación de los sujetos en uno de los cuatro grados de severidad de la enfermedad (clasificación multiclase) y la estimación del AHI (regresión). Este enfoque contrasta con el seguido de forma común en el estado del arte, cuyos principales estudios están centrados en la detección de cada uno de los eventos apneicos de los registros.

C.6. Resultados y discusión

Tras la aplicación al FA monocanal de nuestra metodología de análisis automático, se ha mejorado el rendimiento diagnóstico de un algoritmo de detección de eventos clásico aplicado a nuestras bases de datos. Así, en clasificación binaria, un modelo basado en la técnica de *ensemble learning* AdaBoost, construido con árboles de decisión, obtuvo 89.0 % de sensibilidad (S), 80.0 % de especificidad (E), 86.5 % de precisión (P), 0.950 de área bajo la curva *receiver-operating characteristics* (AROC) y 0.672 de la κ de Cohen, frente a 75.8 % S, 54.3 % E, 64.0 % P, 0.635 AROC y 0.286 κ de dicho algoritmo clásico. En cuanto a la clasificación multiclase, otro modelo AdaBoost, construido con clasificadores basados en análisis discriminante lineal, obtuvo precisiones diagnósticas del 86.5 %, 81.0 % y 82.5 % al ser evaluado en cada uno de los cortes del AHI que establencen los 4 grados de severidad del SAHS (AHI = 5 eventos/hora, 15 e/h y 30 e/h). El algoritmo clásico alcanzó peor rendimiento diagnóstico para cada umbral: 81.0 %, 68.3 % y 63.5 %, respectivamente. Por último, en cuanto a la estimación del AHI mediante regresión, un modelo de red neuronal artificial basado en el método multi-layer perceptron (MLP) obtuvo un coeficiente de correlación intra-clase (ICC) de 0.849, precisiones diagnósticas de 79.7 %, 91.5 %, 79.7 % y 88.1 % para los AHI = 5 e/h, 10 e/h, 15 e/h y 30 e/h, respectivamente, cada uno de ellos asociados además a un AROC de 0.903, 0.956, 0.904 y 0.973. Por el contrario, el algoritmo de detección de eventos alcanzó 0.840 ICC, y unas precisiones diagnósticas correspondientes de 79.7 % (0.823 AROC), 78.0 % (0.833 AROC), 66.1 % (0.867 AROC) y 74.6 % (0.982 AROC).

La Tabla C.5 muestra los resultados obtenidos en los principales estudios del estado de la técnica. Además, la Tabla C.6 resume los mejores resultados mostrados en esta Tesis Doctoral para el caso de sujetos adultos. Como puede observarse, nuestro enfoque se encuentra entre los que mayor rendimiento diagnóstico obtienen, comparado con los estudios que implementan un enfoque de detección de eventos en el FA, así como con los estudios centrados en las señales HRV y SpO_2.

Por otro lado, nuestra metodología aplicada a registros FA domiciliarios de pacientes pediátricos mostraron un mejor rendimiento que el índice de desaturación de oxígeno (*oxygen desaturation index*, ODI), común en la práctica clínica. Además, la combinación de la información espectral de dichos registros con el ODI mediante regresión logística obtuvo 85.9 % S, 87.4 % E, 86.3 % P, y 0.947 AROC. Como puede verse en la Tabla C.7 estos resultados mejoran los obtenidos por los estudios del estado de la técnica centrados en sujetos pediátricos.

C.7. Conclusiones

De acuerdo con todo lo anteriormente expuesto, en esta Tesis Doctoral se han alcanzado las siguientes conclusiones principales:

1. Las técnicas de reconocimiento de patrones aplicadas sobre la señal FA monocanal son útiles para mejorar el proceso automático del diagnóstico del SAHS.

2. Se puede alcanzar un alto rendimiento diagnóstico a través del análisis automático de la señal FA monocanal independientemente de que ésta haya sido adquirida mediante un termistor o una sonda de presión.

Tabla C5: Resumen de la capacidad diagnóstica mostrada por los principales estudios del estado de la técnica. S: sensibilidad (%); E: especificidad (%); P (%): precisión; AROC: area bajo la curva ROC; Ter: termistor; SP: sonda de presión. PSG: polisomnografía. *: estimado a partir de los datos del estudio; -: sin datos suficientes para estimar; H-O: validación *hold-out* (entrenamiento y test); loo: validación cruzada dejando uno fuera (*leave-one-out cross-validation*); *k*-fold: validación cruzada *k*-fold (*k-fold cross-validation*). SVM: máquinas de vector soporte (support vector machines); MLP: perceptron multicapa (*multi-layer perceptron*); LDA: análisis discriminante lineal (*linear discriminant analysis*); QDA: análisis discriminante cuadrático (*quadratic discriminant analysis*); KNN: K vecinos más cercanos (*K-nearest neighbors*). E-D: detección de eventos.

Estudio	**Mét.**	**Señal**	***n***	**umbral AHI**	**Valid.**	**S**	**E**	**P**	**AROC**
Shochat et al [116]	E-D	AF(Ter)	288	10	PSG	86.0	57.0	-	-
Gergely et al [54]	E-D	AF(Ter)	83	15	PSG	71.9	73.1	72,3*	-
Nakano et al [89]	E-D	AF(Ter)	216	5	H-O	88.0	80.0	-	0.950
				10		92.0	90.0	-	0.960
				15		86.0	90.0	-	0.950
Nakano et al [89]	E-D	AF(SP)	217	5	H-O	97.0	77.0	-	0.950
				10		97.0	76.0	-	**0.970**
				15		97.0	73.0	-	**0.980**
De Almeida et al [35]	E-D	AF(SP)	30	5	PSG	86.4	75.0	83,3*	0.886
				10		85.7	87.5	86,7*	0.915
				15		83.3	83.3	83,3*	0.898
Erman et al [41]	E-D	AF(SP)	59	5	PSG	85.4	50.0	74,6*	0.863
				10		82.1	83.9	83,1*	0.862
				15		90.9	94.6	**93,2***	0.977
Chen et al [28]	E-D	AF(SP)	50	5	PSG	97.7	66.7	94,0*	**0.951**
				15		87.5	88.9	88,0*	0.944
				30		88.2	93.9	**90,0***	0.955
Rofail et al [111]	E-D	AF(SP)	200	5	PSG	94.0	62.0	87,0*	0.840
				30		90.0	89.0	89,5*	**0.960**
BaHammam et al [14]	E-D	AF(SP)	95	5	PSG	79.0	68.0	77,9*	0.854
				10		70.0	89.0	75,8*	0.856
				15		65.0	94.0	75,8*	0.805
				30		63.0	98.0	83,2*	0.878
Roche et al [107]	*Tree*	HRV	147	10	*k*-fold	64,2*	75,6*	69,3*	-
Al-Angari et al [5]	*SVM*	HRV	100	5	-	79.6	78.4	79.0	-
Ravelo-García et al [103]	*LR*	HRV	97	10	*k*-fold	88.7	82.9	86,6*	0.941
Marcos et al [85]	*MLP*	SpO_2	187	10	H-O	89.8	79.4	85.5	0.900
Marcos et al [84]	*LDA*	SpO_2	187	10	H-O	86.6	80.4	84.1	0.925
	QDA					91.1	78.3	85.8	0.913
	KNN					88.1	84.8	86.7	0.822
	LR					85.1	87.0	85.8	0.930
Álvarez et al [9]	*LR*	SpO_2	148	10	loo	92.0	85.4	**89.7**	0.967
Marcos et al [83]	*MLP*	SpO_2	240	5	H-O	91.8	58.8	84.0	-
				10		89.6	81.3	86.8	-
				15		94.9	90.9	93.1	-
Álvarez et al [10]	*SVM*	SpO_2	320	10	H-O	95.2	80.0	84.5	-
Al-Angari et al [5]	*SVM*	SpO_2	100	5	-	91.8	98.0	**95.0**	-

Tabla C6: Resumen de los métodos de este estudio que alcanzaron el mayor rendimiento diagnóstico en las bases de datos de adultos para clasificación binaria, classificación multiclase y regresión. S: sensibilidad (%); E: especificidad (%); P (%): precisión; AROC: area bajo la curva ROC; Ter: termistor; SP: sonda de presión. w: mujeres. -: sin datos suficientes para estimar; H-O: validación *hold-out* (entrenamiento y test); loo: validación cruzada dejando uno fuera (*leave-one-out cross-validation*)

Método	**Señal**	***n***	**umbral AHI**	**Valid.**	**S**	**E**	**P**	**AROC**
$AB-CART$ (binario)[62]	AF(NP)	317	10	H-O	89.0	80.0	86.5	0.935
LR^w [63]	HRV	54	10	loo	80.8	89.3	85.2	0.951
$AB-LDA$ (multi)[62]	AF(NP)	317	5	H-O	87.1	80.0	86.5	-
			15		85.9	72.9	81.0	-
			30		74.2	90.6	82.5	-
MLP [64]	AF, RRV(Th)	148	5	H-O	91.7	27.3	79.7	0.903
			10		92.5	89.5	**91.5**	0.956
			15		83.9	75.0	79.7	0.904
			30		88.9	88.0	88.1	**0.973**

Tabla C7: Resumen de la capacidad diagnóstica mostrada por los principales estudios del estado de la técnica centrados en SAHS pediátrico. S: sensibilidad (%); E: especificidad (%); P (%): precisión; AROC: area bajo la curva ROC; PSG: polisomnografía; PPG: fotopletismografía (*photopletysmography*); PRV: variabilidad del ritmo de pulso (*pulse rate variability*). *: estimado a partir de los datos del estudio; -: sin datos suficientes para estimar; loo: validación cruzada dejando uno fuera (*leave-one-out cross-validation*).

Estudio	**Señal**	***n***	**umbral AHI**	**Valid.**	**S**	**E**	**P**	**AROC**
Shouldice et al. [117]	HRV	50	1	loo	85.7	81.8	84.0	0.830
Rembold and Suratt [104]	Sonidos	26	3	-	61.5*	100*	80.8*	-
Gil et al. [56]	PRV	21	5	loo	75.0	85.7	80.0	-
Spruyt and Gozal [119]	-	1133	3	PSG	59.0	82.9	-	0.790
Kadmon et al. [73]	-	85	5	PSG	83.0	64.0	70.6*	0.650
Chang et al. [25]	SpO_2	141	5	PSG	60.0	86.0	71.6*	-
Garde et al. [53]	SpO_2+PRV	146	5	k-fold	83.6	88.4	84.9	0.860
LR_{AF+ODI}	AF+ SpO_2	50	3	loo	**85.9**	87.4	**86.3**	**0.947**

3. Nuestra propuesta, basada en el análisis exhaustivo y automático de la señal FA monocanal, mejora el rendimiento diagnóstico de un algoritmo clásico de detección de eventos aplicado sobre nuestras propias bases de datos. Además, nuestra propuesta mostró un alto rendimiento diagnóstico en comparación con estudios del estado de la técnica que también aplicaron el enfocque de detección de eventos.

4. El método de *ensemble learning AdaBoost*, mejora los modelos LDA, CART, y LR, tanto en la clasificación binaria (AB-CART) como en la clasificación multiclase (AB-LDA). La red neuronal MLP alcanza el mayor rendimiento diagnóstico en la estimación del AHI, mejorando los métodos MLR y RBF.

5. Los enfoques lineal y no lineal, implementados como análisis en el dominio de la frecuencia y en el del tiempo, ofrecen información complementaria para caracterizar el SAHS.

6. La señal FA monocanal muestra un mayor potencial diagnóstico que la señal HRV, tanto en el caso de la aplicación de la misma metodología (nuestro enfoque aplicado sobre la señal HRV) como en el caso de estudios del estado de la técnica también centrados en dicha señal.

7. La información espectral de los registros FA obtenidos en el domicilio de los pacientes es de utilidad para el diagnóstico del SAHS en niños.

8. La combinación de la información espectral del FA y el ODI es de utilidad para diagnosticar de manera precisa el SAHS en niños en el domicilio de los pacientes.

En resumen, se ha caracterizado el SAHS a partir de la información extraída del FA monocanal. Ésta fue útil para construir modelos de reconocimiento de patrones con capacidad para alcanzar un alto rendimiento diagnóstico. Nuestra propuesta superó el enfoque de detección de eventos y mostró un alto rendimiento diagnóstico en comparación con los principales estudios del estado de la técnica. Estos resultados sugieren que la prueba diagnóstica del SAHS puede ser simplificada de manera fiable mediante el uso del análisis automático del FA monocanal.

Bibliography

[1] A. Aarabi, F. Wallois, and R. Grebe. Automated neonatal seizure detection: A multistage classification system through feature selection based on relevance and redundancy analysis. *Clinical Neurophysiology*, 117(2):328–340, 2006.

[2] D. Abásolo, R. Hornero, P. Espino, D. Álvarez, and J. Poza. Entropy analysis of the eeg background activity in alzheimer's disease patients. *Physiological measurement*, 27(3):241, 2006.

[3] D. Abásolo, R. Hornero, C. Gómez, M. García, and M. López. Analysis of eeg background activity in alzheimer's disease patients with lempel–ziv complexity and central tendency measure. *Medical engineering & physics*, 28(4):315–322, 2006.

[4] U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri. Heart rate variability: a review. *Medical and biological engineering and computing*, 44(12):1031–1051, 2006.

[5] H. M. Al-Angari and A. V. Sahakian. Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier. *Information Technology in Biomedicine, IEEE Transactions on*, 16(3):463–468, 2012.

[6] M. Alonso-Álvarez, T. Canet, M. Cubel-Alarco, E. Estivill, E. Fernandez-Julian, D. Gozal, M. Jurado-Luqué, A. Lluch-Roselló, F. Martínez-Pérez, M. Merino-Andreu, et al. Consensus document on sleep apnea-hypopnea syndrome in children. *Arch Bronconeumol*, 47(Suppl 5):1–18, 2011.

[7] D. Álvarez, R. Hornero, D. Abásolo, F. Del Campo, and C. Zamarrón. Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection. *Physiological Measurement*, 27(4):399, 2006.

[8] D. Álvarez, R. Hornero, M. García, F. del Campo, and C. Zamarrón. Improving diagnostic ability of blood oxygen saturation from overnight pulse oximetry in obstructive sleep apnea detection by means of central tendency measure. *Artificial intelligence in medicine*, 41(1):13–24, 2007.

[9] D. Álvarez, R. Hornero, J. V. Marcos, and F. del Campo. Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. *Biomedical Engineering, IEEE Transactions on*, 57(12):2816–2824, 2010.

[10] D. Álvarez, R. Hornero, J. V. Marcos, N. Wessel, T. Penzel, M. Glos, and F. Del Campo. Assessment of feature selection and classification approaches to enhance information from overnight oximetry in the context of apnea diagnosis. *International journal of neural systems*, 23(05), 2013.

[11] M. L. A. Álvarez, J. T. Santos, J. A. C. Guevara, A. I. N. Egüia, E. O. Carbajo, J. F. M. Jiménez, and R. Pelayo. Reliability of respiratory polygraphy for the diagnosis of sleep apnea-hypopnea syndrome in children. *Archivos de Bronconeumología ((English Edition))*, 44(6):318–323, 2008.

[12] D. Álvarez-Estévez and V. Moret-Bonillo. Fuzzy reasoning used to detect apneic events in the sleep apnea-hypopnea syndrome. *Expert Systems with Applications*, 36(4):7778–7785, 2009.

[13] A. BaHammam. Comparison of nasal prong pressure and thermistor measurements for detecting respiratory events during sleep. *Respiration*, 71(4):385–390, 2004.

[14] A. BaHammam, M. Sharif, D. E. Gacuan, and S. George. Evaluation of the accuracy of manual and automatic scoring of a single airflow channel in patients with a high probability of obstructive sleep apnea. *Medical science monitor: international medical journal of experimental and clinical research*, 17(2):MT13, 2011.

[15] G. Baselli, S. Cerutti, S. Civardi, F. Lombardi, A. Malliani, M. Merri, M. Pagani, and G. Rizzo. Heart rate variability signal processing: a quantitative approach as an aid to diagnosis in cardiovascular pathologies. *International journal of bio-medical computing*, 20(1):51–70, 1987.

[16] D. Benitez, P. Gaydecki, A. Zaidi, and A. Fitzpatrick. The use of the hilbert transform in ecg signal analysis. *Computers in biology and medicine*, 31(5):399–406, 2001.

[17] J. Bennett and W. Kinnear. Sleep on the cheap: the role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome. *Thorax*, 54(11):958–959, 1999.

[18] 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(5):597–619, 2012.

[19] C. M. Bishop et al. *Neural networks for pattern recognition*. Clarendon press Oxford, 1995.

[20] C. M. Bishop et al. *Pattern recognition and machine learning*. springer New York, 2006.

[21] M. R. Bonsignore, S. Romano, O. Marrone, M. Chiodi, and G. Bonsignore. Different heart rate patterns in obstructive apneas during nrem sleep. *Sleep*, 20(12):1167–1174, 1997.

[22] J. D. Bronzino. *Medical devices and systems*. CRC Press, 2006.

[23] P. Calderón de la Barca. *La vida es sueño*. 1635.

[24] F. Campos-Rodriguez, M. A. Martinez-Garcia, M. Martinez, J. Duran-Cantolla, M. d. l. Peña, M. J. Masdeu, M. Gonzalez, F. d. Campo, I. Gallego, J. M. Marin, et al. Association between obstructive sleep apnea and cancer incidence in a large multicenter spanish cohort. *American journal of respiratory and critical care medicine*, 187(1):99–105, 2013.

[25] L. Chang, J. Wu, and L. Cao. Combination of symptoms and oxygen desaturation index in predicting childhood obstructive sleep apnea. *International journal of pediatric otorhinolaryngology*, 77(3):365–371, 2013.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.

[27] C.-C. Chen and H. X. Barnhart. Comparison of icc and ccc for assessing agreement for data without and with replications. *Computational Statistics & Data Analysis*, 53(2):554–564, 2008.

[28] H. Chen, A. A. Lowe, Y. Bai, P. Hamilton, J. A. Fleetham, and F. R. Almeida. Evaluation of a portable recording device (apnealink™) for case selection of obstructive sleep apnea. *Sleep and Breathing*, 13(3):213–219, 2009.

[29] E. Chiner, J. Signes-Costa, J. M. Arriero, J. Marco, I. Fuentes, and A. Sergado. Nocturnal oximetry for the diagnosis of the sleep apnoea hypopnoea syndrome: a method to reduce the number of polysomnographies? *Thorax*, 54(11):968–971, 1999.

[30] J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[31] M. E. Cohen, D. L. Hudson, and P. C. Deedwania. Applying continuous chaotic modeling to cardiac signal analysis. *Engineering in Medicine and Biology Magazine, IEEE*, 15(5):97–102, 1996.

[32] M. Costa, A. L. Goldberger, and C.-K. Peng. Multiscale entropy analysis of complex physiologic time series. *Physical review letters*, 89(6):068102, 2002.

[33] M. Costa, A. L. Goldberger, and C.-K. Peng. Multiscale entropy analysis of biological signals. *Physical review E*, 71(2):021906, 2005.

[34] D. Cysarz, R. Zerm, H. Bettermann, M. Frühwirth, M. Moser, and M. Kröz. Comparison of respiratory rates derived from heart rate variability, ecg amplitude, and nasal/oral airflow. *Annals of biomedical engineering*, 36(12):2085–2094, 2008.

[35] F. R. de Almeida, N. T. Ayas, R. Otsuka, H. Ueda, P. Hamilton, F. C. Ryan, and A. A. Lowe. Nasal pressure recordings to detect obstructive sleep apnea. *Sleep and Breathing*, 10(2):62–69, 2006.

[36] P. De Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'Malley. Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea. *Biomedical Engineering, IEEE Transactions on*, 50(6):686–696, 2003.

[37] J. Durán, S. Esnaola, R. Rubio, and Á. Iztueta. Obstructive sleep apnea–hypopnea and related clinical features in a population-based sample of subjects aged 30 to 70 yr. *American journal of respiratory and critical care medicine*, 163(3):685–689, 2001.

[38] J. Durán-Cantolla, C. Puertas, A. Pin, and C. Santa María. Documento de consenso nacional sobre el síndrome de apneas-hipopneas del sueño (sahs) grupo español de sueño (ges). *Arch Bronconeumol*, 41:1–110, 2005.

[39] D. J. Eckert and A. Malhotra. Pathophysiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, 5(2):144–153, 2008.

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

[41] M. K. Erman, D. Stewart, D. Einhorn, N. Gordon, and E. Casal. Validation of the apnealink™for the screening of sleep apnea: a novel and simple single-channel recording device. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, 3(4):387, 2007.

[42] J. Escudero, D. Abásolo, R. Hornero, P. Espino, and M. López. Analysis of electroencephalograms in alzheimer's disease patients with multiscale entropy. *Physiological measurement*, 27(11):1091, 2006.

[43] R. Farre, J. Montserrat, M. Rotger, E. Ballester, and D. Navajas. Accuracy of thermistors and thermocouples as flow-measuring devices for detecting hypopnoeas. *European Respiratory Journal*, 11(1):179–182, 1998.

[44] J. A. Fiz, R. Jane, J. Solà-Soler, J. Abad, M. García, and J. Morera. Continuous analysis and monitoring of snores and their relationship to the apnea-hypopnea index. *The Laryngoscope*, 120(4):854–862, 2010.

[45] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. Maconochie, L. Tarassenko, and D. Mant. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. *The Lancet*, 377(9770):1011–1018, 2011.

[46] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley. Access to diagnosis and treatment of patients with suspected sleep apnea. *American journal of respiratory and critical care medicine*, 169(6):668–672, 2004.

[47] W. W. Flemons and M. R. Littner. Measuring agreement between diagnostic devices. *CHEST Journal*, 124(4):1535–1542, 2003.

[48] W. W. Flemons, M. R. Littner, J. A. Rowley, P. Gay, W. M. Anderson, D. W. Hudgel, R. D. McEvoy, and D. I. Loube. Home diagnosis of sleep apnea: a systematic review of the literature: an evidence review cosponsored by the american academy of sleep medicine, the american college of chest physicians, and the american thoracic society. *CHEST Journal*, 124(4):1543–1579, 2003.

[49] A. O. S. A. T. Force, A. A. of Sleep Medicine, et al. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, 5(3):263, 2009.

[50] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[51] J. H. Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[52] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[53] A. Garde, P. Dehkordi, W. Karlen, D. Wensley, J. M. Ansermino, and G. A. Dumont. Development of a screening tool for sleep disordered breathing in children using the phone oximeter™. *PLoS ONE*, 2014.

[54] V. Gergely, H. Pallos, K. Mashima, S. Miyazaki, T. Tanaka, M. Okawa, and N. Yamada. Evaluation of the usefulness of the sleepstrip for screening obstructive sleep apnea-hypopnea syndrome in japan. *Sleep and Biological Rhythms*, 7(1):43–51, 2009.

[55] M. D. Ghegan, P. C. Angelos, A. C. Stonebraker, and M. B. Gillespie. Laboratory versus portable sleep studies: A meta-analysis. *The Laryngoscope*, 116(6):859–864, 2006.

[56] E. Gil, R. Bailón, J. M. Vergara, and P. Laguna. Ptt variability for discrimination of sleep apnea related decreases in the amplitude fluctuations of ppg signal in children. *Biomedical Engineering, IEEE Transactions on*, 57(5):1079–1088, 2010.

[57] E. Gil, M. Mendez, J. M. Vergara, S. Cerutti, A. M. Bianchi, and P. Laguna. Discrimination of sleep-apnea-related decreases in the amplitude fluctuations of ppg signal in children by hrv analysis. *Biomedical Engineering, IEEE Transactions on*, 56(4):1005–1014, 2009.

[58] S. S. Grover and S. D. Pittman. Automated detection of sleep disordered breathing using a nasal pressure monitoring device. *Sleep and Breathing*, 12(4):339–345, 2008.

[59] C. Guilleminault, J. H. Lee, and A. Chan. Pediatric obstructive sleep apnea syndrome. *Archives of pediatrics & adolescent medicine*, 159(8):775–785, 2005.

[60] C. Guilleminault, R. Winkle, S. Connolly, K. Melvin, and A. Tilkian. Cyclical variation of the heart rate in sleep apnoea syndrome: Mechanisms, and usefulness of 24 h electrocardiography as a screening technique. *The Lancet*, 323(8369):126–131, 1984.

[61] G. C. Gutiérrez-Tobal, M. L. Alonso-Álvarez, D. Álvarez, F. del Campo, J. Terán-Santos, and R. Hornero. Diagnosis of pediatric obstructive sleep apnea: Preliminary findings using automatic analysis of airflow and oximetry recordings obtained at patients' home. *Biomedical Signal Processing and Control*, 18:401–407, 2015.

[62] G. C. Gutiérrez-Tobal, D. Álvarez, F. del Campo, and R. Hornero. Utility of adaboost to detect sleep apnea-hypopnea syndrome from single-channel airflow. *IEEE Transactions on Biomedical Engineering*, In Press, 2015.

[63] G. C. Gutiérrez-Tobal, D. Álvarez, J. Gomez-Pilar, F. del Campo, and R. Hornero. Assessment of time and frequency domain entropies to detect sleep apnoea in heart rate variability recordings from men and women. *Entropy*, 17(1):123–141, 2015.

[64] G. C. Gutiérrez-Tobal, D. Álvarez, J. V. Marcos, F. del Campo, and R. Hornero. Pattern recognition in airflow recordings to assist in the sleep apnoea–hypopnoea syndrome diagnosis. *Medical & biological engineering & computing*, 51(12):1367–1380, 2013.

[65] G. C. Gutiérrez-Tobal, R. Hornero, D. Álvarez, J. V. Marcos, and F. del Campo. Linear and nonlinear analysis of airflow recordings to help in sleep apnoea–hypopnoea syndrome diagnosis. *Physiological measurement*, 33(7):1261, 2012.

[66] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[67] J. Han, H.-B. Shin, D.-U. Jeong, and K. S. Park. Detection of apneic events from single channel nasal airflow using 2nd derivative method. *Computer methods and programs in biomedicine*, 91(3):199–207, 2008.

[68] R. Hornero, D. Álvarez, D. Abásolo, F. del Campo, and C. Zamarrón. Utility of approximate entropy from overnight pulse oximetry data in the diagnosis of the obstructive sleep apnea syndrome. *Biomedical Engineering, IEEE Transactions on*, 54(1):107–113, 2007.

[69] D. W. Hosmer Jr and S. Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2002.

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

[71] T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda, and M. Hirano. Quantification of eeg irregularity by use of the entropy of the power spectrum. *Electroencephalography and clinical neurophysiology*, 79(3):204–210, 1991.

[72] J. Jobson. *Applied multivariate data analysis: volume I: Regression and Experimental Design*. Springer Science & Business Media, 1991.

[73] G. Kadmon, C. M. Shapiro, S. A. Chung, and D. Gozal. Validation of a pediatric obstructive sleep apnea screening tool. *International journal of pediatric otorhinolaryngology*, 77(9):1461–1464, 2013.

[74] A. S. Karunajeewa, U. R. Abeyratne, and C. Hukins. Multi-feature snore sound analysis in obstructive sleep apnea–hypopnea syndrome. *Physiological measurement*, 32(1):83, 2011.

[75] E. S. Katz, R. B. Mitchell, and C. M. D'Ambrosio. Obstructive sleep apnea in infants. *American journal of respiratory and critical care medicine*, 185(8):805–816, 2012.

[76] B. L. Koley and D. Dey. Automatic detection of sleep apnea and hypopnea events from single channel measurement of respiration signal employing ensemble binary svm classifiers. *Measurement*, 46(7):2082–2092, 2013.

[77] J. Korten and G. Haddad. Respiratory waveform pattern recognition using digital techniques. *Computers in biology and medicine*, 19(4):207–217, 1989.

[78] P. Levy, J. L. Pepin, C. Deschaux-Blanc, B. Paramelle, and C. Brambilla. Accuracy of oximetry for detection of respiratory disturbances in sleep apnea syndrome. *Chest Journal*, 109(2):395–399, 1996.

[79] C.-L. Lin, C. Yeh, C.-W. Yen, W.-H. Hsu, and L.-W. Hang. Comparison of the indices of oxyhemoglobin saturation by pulse oximetry in obstructive sleep apnea hypopnea syndrome. *CHEST Journal*, 135(1):86–93, 2009.

[80] E. Lindberg, N. Carter, T. Gislason, and C. Janson. Role of snoring and daytime sleepiness in occupational accidents. *American journal of respiratory and critical care medicine*, 164(11):2031–2035, 2001.

[81] F. Lopez-Jimenez, F. H. S. Kuniyoshi, A. Gami, and V. K. Somers. Obstructive sleep apnea: implications for cardiac and vascular disease. *CHEST Journal*, 133(3):793–804, 2008.

[82] U. J. Magalang, J. Dmochowski, S. Veeramachaneni, A. Draw, M. J. Mador, A. El-Solh, and B. J. Grant. Prediction of the apnea-hypopnea index from overnight pulse oximetry. *CHEST Journal*, 124(5):1694–1701, 2003.

[83] J. V. Marcos, R. Hornero, D. Álvarez, M. Aboy, and F. Del Campo. Automated prediction of the apnea-hypopnea index from nocturnal oximetry recordings. *Biomedical Engineering, IEEE Transactions on*, 59(1):141–149, 2012.

[84] J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo, and C. Zamarrón. Assessment of four statistical pattern recognition techniques to assist in obstructive sleep apnoea diagnosis from nocturnal oximetry. *Medical engineering & physics*, 31(8):971–978, 2009.

[85] J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo, C. Zamarrón, and M. López. Utility of multilayer perceptron neural network classifiers in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry. *computer methods and programs in biomedicine*, 92(1):79–89, 2008.

[86] M. Martin, A. Plastino, and O. Rosso. Statistical complexity and disequilibrium. *Physics Letters A*, 311(2):126–132, 2003.

[87] D. S. Morillo and N. Gross. Probabilistic neural network approach for the detection of sahs from overnight pulse oximetry. *Medical & biological engineering & computing*, 51(3):305–315, 2013.

[88] I. Nabney. *NETLAB: algorithms for pattern recognition*. Springer Science & Business Media, 2002.

[89] H. Nakano, T. Tanigawa, T. Furukawa, and S. Nishima. Automatic detection of sleep-disordered breathing from a single-channel airflow record. *European Respiratory Journal*, 29(4):728–736, 2007.

[90] N. Netzer, A. H. Eliasson, C. Netzer, and D. A. Kristo. Overnight pulse oximetry for sleep-disordered breathing in adults: a review. *CHEST Journal*, 120(2):625–633, 2001.

[91] C. A. Nigro, E. Dibur, S. Aimaretti, S. González, and E. Rhodius. Comparison of the automatic analysis versus the manual scoring from apnealink™device for the diagnosis of obstructive sleep apnoea syndrome. *Sleep and Breathing*, 15(4):679–686, 2011.

[92] G. Parati, C. Lombardi, and K. Narkiewicz. Sleep apnea: epidemiology, pathophysiology, and relation to cardiovascular risk. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 293(4):R1671–R1683, 2007.

[93] S. P. Patil, H. Schneider, A. R. Schwartz, and P. L. Smith. Adult obstructive sleep apnea: pathophysiology and diagnosis. *Chest Journal*, 132(1):325–337, 2007.

[94] T. Penzel, J. W. Kantelhardt, L. Grote, J.-H. Peter, and A. Bunde. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *Biomedical Engineering, IEEE Transactions on*, 50(10):1143–1151, 2003.

[95] T. Penzel, J. McNames, P. De Chazal, B. Raymond, A. Murray, and G. Moody. Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. *Medical and Biological Engineering and Computing*, 40(4):402–407, 2002.

[96] T. Penzel, G. Moody, R. Mark, A. Goldberger, and J. Peter. The apnea-ecg database. In *Computers in Cardiology 2000*, pages 255–258. IEEE, 2000.

[97] S. M. Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[98] S. M. Pincus. Assessing serial irregularity and its implications for health. *Annals of the New York Academy of Sciences*, 954(1):245–267, 2001.

[99] J. Poza, R. Hornero, D. Abásolo, A. Fernández, and M. García. Extraction of spectral based measures from meg background oscillations in alzheimer's disease. *Medical engineering & physics*, 29(10):1073–1083, 2007.

[100] N. M. Punjabi. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, 5(2):136–143, 2008.

[101] A. Qureshi, R. D. Ballard, and H. S. Nelson. Obstructive sleep apnea. *Journal of Allergy and Clinical Immunology*, 112(4):643–651, 2003.

[102] S. I. Rathnayake, I. A. Wood, U. R. Abeyratne, and C. Hukins. Nonlinear features for single-channel diagnosis of sleep-disordered breathing diseases. *Biomedical Engineering, IEEE Transactions on*, 57(8):1973–1981, 2010.

[103] A. Ravelo-García, P. Saavedra-Santana, G. Juliá-Serdá, J. Navarro-Mesa, J. Navarro-Esteva, X. Álvarez-López, A. Gapelyuk, T. Penzel, and N. Wessel. Symbolic dynamics marker of heart rate variability combined with clinical variables enhance obstructive sleep apnea screening. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(2):024404, 2014.

[104] C. M. Rembold and P. M. Suratt. Children with obstructive sleep-disordered breathing generate high-frequency inspiratory sounds during sleep. *SLEEP-NEW YORK THEN WESTCHESTER-*, 27:1154–1162, 2004.

[105] J. S. Richman and J. R. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.

[106] M. Riedl, A. Müller, J. F. Kraemer, T. Penzel, J. Kurths, and N. Wessel. Cardio-respiratory coordination increases during sleep apnea. *PloS one*, 9(4), 2014.

[107] F. Roche, V. Pichot, E. Sforza, D. Duverney, F. Costes, M. Garet, J. Barthélémy, et al. Predicting sleep apnoea syndrome from heart period: a time-frequency wavelet analysis. *European Respiratory Journal*, 22(6):937–942, 2003.

[108] N. Roche, B. Herer, C. Roig, and G. Huchon. Prospective testing of two models based on clinical and oximetric variables for prediction of obstructive sleep apnea. *CHEST Journal*, 121(3):747–752, 2002.

[109] G.-M. J. Rodríguez, R. P. de Lucas, J. M. Sánchez, A. J. Izquierdo, A. R. Peraíta, and M. J. Cubillo. [usefulness of the visual analysis of night oximetry as a screening method in patients with suspected clinical obstructive sleep apnea syndrome]. *Archivos de bronconeumologia*, 32(9):437–441, 1996.

[110] L. M. Rofail, K. K. Wong, G. Unger, G. B. Marks, and R. R. Grunstein. Comparison between a single-channel nasal airflow device and oximetry for the diagnosis of obstructive sleep apnea. *Sleep*, 33(8):1106, 2010.

[111] L. M. Rofail, K. K. Wong, G. Unger, G. B. Marks, and R. R. Grunstein. The role of single-channel nasal airflow pressure transducer in the diagnosis of osa in the sleep laboratory. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, 6(4):349, 2010.

[112] L. M. Rofail, K. K. Wong, G. Unger, G. B. Marks, and R. R. Grunstein. The utility of single-channel nasal airflow pressure transducer in the diagnosis of osa at home. *Sleep*, 33(8):1097, 2010.

[113] D. Sánchez-Morillo, M. López-Gordo, and A. León. Novel multiclass classification for home-based diagnosis of sleep apnea hypopnea syndrome. *Expert Systems with Applications*, 41(4):1654–1662, 2014.

[114] A. Sassani, L. J. Findley, M. Kryger, E. Goldlust, C. George, and T. M. Davidson. Reducing motor-vehicle collisions, costs, and fatalities by treating obstructive sleep apnea syndrome. *SLEEP-NEW YORK THEN WESTCHESTER-*, 27(3):453–458, 2004.

[115] A. S. Shamsuzzaman, B. J. Gersh, and V. K. Somers. Obstructive sleep apnea: implications for cardiac and vascular disease. *Jama*, 290(14):1906–1914, 2003.

[116] T. Shochat, N. Hadas, M. Kerkhofs, A. Herchuelz, T. Penzel, J. Peter, and P. Lavie. The sleepstriptm: an apnoea screener for the early detection of sleep apnoea syndrome. *European Respiratory Journal*, 19(1):121–126, 2001.

[117] R. B. Shouldice, L. M. O'Brien, C. O'Brien, P. de Chazal, D. Gozal, and C. Heneghan. Detection of obstructive sleep apnea in pediatric subjects using surface lead electrocardiogram features. *Sleep*, 27(4):784, 2004.

[118] L. Sörnmo and P. Laguna. *Bioelectrical signal processing in cardiac and neurological applications*. Academic Press, 2005.

[119] K. Spruyt and D. Gozal. Screening of pediatric sleep-disordered breathing: a proposed unbiased discriminative set of questions using clinical severity scales. *CHEST Journal*, 142(6):1508–1515, 2012.

[120] J. Stradling. Sleep studies for sleep-related breathing disorders. *Journal of sleep research*, 1(4):265–273, 1992.

[121] E. Tabachnik, N. Muller, B. Toye, and H. Levinson. Measurement of ventilation in children using the respiratory inductive plethysmograph. *The Journal of pediatrics*, 99(6):895–899, 1981.

[122] P. Várady, T. Micsik, S. Benedek, and Z. Benyó. A novel method for the detection of apnea and hypopnea events in respiration signals. *Biomedical Engineering, IEEE Transactions on*, 49(9):936–942, 2002.

[123] J. Vavrina. Computer assisted pulse oximetry for detecting children with obstructive sleep apnea syndrome. *International journal of pediatric otorhinolaryngology*, 33(3):239–248, 1995.

[124] J. P. Weir. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *The Journal of Strength & Conditioning Research*, 19(1):231–240, 2005.

[125] 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 Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[126] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[127] K. K. Wong, D. Jankelson, A. Reid, G. Unger, G. Dungan, J. A. Hedner, and R. R. Grunstein. Diagnostic test evaluation of a nasal flow monitor for obstructive sleep apnea detection in sleep apnea research. *Behavior research methods*, 40(1):360–366, 2008.

[128] T. Young, P. E. Peppard, and D. J. Gottlieb. Epidemiology of obstructive sleep apnea: a population health perspective. *American journal of respiratory and critical care medicine*, 165(9):1217–1239, 2002.

[129] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

[130] C. Zamarron, P. Romero, J. Rodriguez, and F. Gude. Oximetry spectral analysis in the diagnosis of obstructive sleep apnoea. *Clinical Science*, 97:467–473, 1999.

[131] X.-S. Zhang, Y.-S. Zhu, N. V. Thakor, and Z.-Z. Wang. Detecting ventricular tachycardia and fibrillation by complexity measure. *Biomedical Engineering, IEEE Transactions on*, 46(5):548–555, 1999.

[132] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *Information Theory, IEEE Transactions on*, 24(5):530–536, 1978.

[133] G. Zonios, U. Shankar, and V. K. Iyer. Pulse oximetry theory and calibration for low saturations. *Biomedical Engineering, IEEE Transactions on*, 51(5):818–822, 2004.

[134] M. H. Zweig and G. Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.

Index