Thomas Penzel • Roberto Hornero
Editors

Advances in the Diagnosis and Treatment of Sleep Apnea

Filling the Gap Between Physicians and Engineers

Springer

Editors

Thomas Penzel
Sleep Medicine Center
Charite Universitätsmedizin Berlin
Berlin, Berlin, Germany

Roberto Hornero
Biomedical Engineering Group
University of Valladolid
Valladolid, Spain

Centro de Investigación Biomédica en
Red Bioingeniería
Biomateriales y Nanomedicina
(CIBER-BBN)
Valladolid, Spain

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-3-031-06412-8 ISBN 978-3-031-06413-5 (eBook)
https://doi.org/10.1007/978-3-031-06413-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Sleep apnea is a sleep disorder with a very high prevalence and many health consequences. As such it is a major health burden (Benjafield et al., 2019). Sleep apnea has been systematically explored only a little more than 40 years now (Guilleminault & Dement, 1978). Major impacts of sleep apnea are sleepiness and associated risks for accidents (Bonsignore et al., 2021). Major health impacts are cardiovascular risk and pathophysiological traits, even if this is currently much debated when focusing on the apnea-hypopnea index as the measure for sleep apnea severity (Arnaud et al., 2020). Sleep apnea is a disorder which is a chronic condition and can be treated successfully.

The disorders of sleep-disordered breathing have largely supported the growth of sleep medicine in general from a small specialty field to a major spectrum of disorders in the arena of medical specialties. This activity helped to convert the niche field of sleep research into sleep medicine, a clinical discipline with its own departments, its own center certification, physician certification, dedicated conferences, journals, and research activities. The recognition and importance have grown so much that the new International Classification of Disorders by WHO in its 11th version, being launched in 2022, has added a new section on sleep and wake disorders with its own range of codes. This worldwide recognition will enable the growth of medical education on sleep physiology, sleep pathology, and specific sleep disorders.

The diagnostic field for sleep disorders, and for sleep apnea specifically, is strongly linked to the development of new and recent methods, which allow long-term recording and analysis of physiological functions during sleep. Sleep and sleep apnea are not just identified by taking a single blood sample or by a single measurement by a physician at a visit, but sleep recording requires the continuous recording of biosignals. This is comparable to monitoring of vital functions during anesthesia or intensive care. Because of this methodological challenge, biomedical engineering as well as new sensor and analysis technologies are closely linked to the development of sleep apnea diagnosis. New technologies helped to a large extent develop new diagnostic and treatment modalities for sleep-disordered breathing. Sleep apnea diagnostic research is now linked to the development of new wearables, nearables, and smartphone apps, and profits much from the ubiquitous development of photoplethysmography recording everywhere.

Artificial intelligence is playing a very important role in analyzing sleep recordings and, particularly, in automatizing several of the stages of sleep apnea diagnosis. Since the generalization of computerized analysis in the 1990s, automated processing of cardiorespiratory and neuromuscular signals from polysomnographic studies provided a number of indices able to assist sleep experts in the characterization of the disease (Shokoueinejad et al., 2017). Parameterization of the influence of apneic events on biological system dynamics has relied on widely known techniques from the engineering field, such as spectral and nonlinear analysis. Currently, there is a demand for novel alternative metrics able to overcome the limitations of the standard apnea-hypopnea index concerning its low association with patient symptoms and outcomes (Malhotra et al., 2021). In this regard, signal processing and pattern recognition are going to play a key role. In addition, machine learning has also shown its usefulness in the last decades (Uddin et al., 2018) and, like many other areas in our society, sleep apnea diagnosis is rapidly entering the deep learning era (Mostafa et al., 2019) and big data. These new analytical techniques, along with the advances in health device development, are the main hope for reaching a reliable diagnostic paradigm shift. One that finally could cope with the disease prevalence, personalized interventions, and runaway spending.

Beyond the widespread application of machine learning methods to automate polysomnography scoring and to provide sleep experts with tools for automated diagnosis, artificial intelligence has also the potential to significantly improve the management of sleep apnea treatment. Recent advances in the framework of big data together with remote monitoring capability of novel treatment devices are able to promote conventional sleep medicine towards a real personalized medicine. Identification of refined clinical phenotypes of patients will allow the development of precision interventions, enabling the quick identification of the treatment option that best fits the particular characteristics of a patient (Watson & Fernández., 2021). Similarly, machine learning is able to accurately model patient's adherence from usage data (pressure setting, residual respiratory events, mask leaks) derived from portable treatment devices, improving the efficacy of available therapies (Goldstein et al., 2020). Thus, artificial intelligence is going to significantly change the management of sleep apnea treatment in the short term.

This volume gives a basis of current knowledge on sleep research, sleep medicine, and sleep apnea, with a strong focus on new challenges and new research directions in the diagnosis of sleep apnea and its treatment. The volume contains three sections: the first one is on physiology and pathophysiology, the second one is on diagnostic advances, and the third one is on treatment advances. Each chapter author was asked to not only describe the state of the art but also develop visions for future research as seen from their special angle and viewpoint.

As editors, we think that the volume can serve as an introduction to the field of sleep-disordered breathing, can serve as a basis for educating in sleep-disordered breathing, and can immediately stimulate and trigger new research in physiology, clinical trials, and biomedical engineering for sensors and analysis methodologies.

Berlin, Berlin, Germany Thomas Penzel
Valladolid, Spain Roberto Hornero

References

Arnaud, C., Bochaton, T., Pépin, J. L., & Belaidi, E. (2020). Obstructive sleep apnoea and cardiovascular consequences: Pathophysiological mechanisms. *Archives of Cardiovascular Disease, 113*:350—358

Benjafield, A. V., Ayas, N. T., Eastwood, P. R., Heinzer, R., Ip, M. S. M., Morrell, M. J., Nunez, C. M., Patel, S. R., Penzel, T., Pepin, J. L. D., Peppard, P. E., Sinha, S., Tufik, S., Valentine, K., & Malhotra, A. (2019). Estimating the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respiratory Medicine, 7*:687–698. 10.1016/S2213-2600(19)30198-5

Bonsignore, M.R., Randerath, W., Schiza, S., Verbraecken, J., Elliott, M. W., Riha, R., Barbe, F., Bouloukaki, I., Castrogiovanni, A., Deleanu, O., Goncalves, M., Leger, D., Marrone, O., Penzel, T., Ryan, S., Smyth, D., Teran-Santos, J., Turino, C., McNicholas, W. T. (2021). European Respiratory Society statement on sleep apnoea, sleepiness and driving risk. *European Respiratory Journal, 57*: 2001272 doi: 10.1183/13993003.01272-2020

Goldstein, C. A., Berry, R. B., Kent, D. T., Kristo, D. A., Seixas, A. A., Redline, S., Westover, M. B., Abbasi-Feinberg, F., Aurora, R. N., Carden, K. A., Kirsch, D. B., Malhotra, R. K., Martin, J. L., Olson, E. J., Ramar, K., Rosen, C. L., Rowley, J. A., Shelgikar, A. V. (2020). Artificial intelligence in sleep medicine: An American Academy of Sleep Medicine position statement. *Journal of Clinical Sleep Medicine, 16*(4):605-607. 10.5664/jcsm.8288

Guilleminault, C., & Dement, W. C. (eds) (1978). *Sleep apnea syndromes*. New York: Alan R. Liss Inc.

Malhotra, A., Ayappa, I., Ayas, N., Collop, N., Kirsch, D., Mcardle, N., Mehra, R, Pack, A. I., Punjabi, N., White, D. P., & Gottlieb, D. J. (2021). Metrics of sleep apnea severity: beyond the apnea-hypopnea index. *Sleep, 44*(7):1-16. 10.1093/sleep/zsab030

Mostafa, S. S., et al. (2019). A systematic review of detecting sleep apnea using deep learning. *Sensors, 19.22*: 4934. 10.3390/s19224934

Shokoueinejad, M., Fernandez, C., Carroll, E., Wang, F., Levin, J., Rusk, S., Glattard, N., Mulchrone, A., Zhang, X., Xie, A., Teodorescu, M., Dempsey, J., & Webster, J. (2017). Sleep apnea: a review of diagnostic sensors, algorithms, and therapies. *Physiol Meas, 38*:R204–R252 doi: 10.1088/1361-6579/aa6ec6

Uddin, M. B., Chow, C. M., & S. W. Su. (2018). Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: A systematic review. *Physiological Measurement, 39.3*: 03TR01. 10.1088/1361-6579/aaafb8

Watson, N. F., & Fernandez, C. R. (2021). Artificial intelligence and sleep: Advancing sleep medicine. *Sleep Med Rev, 59*:101512. 10.1016/j.smrv.2021.101512

Contents

Part I Physiology

Part II Diagnostic Innovations

Part III Therapeutic Innovations

Conventional Machine Learning Methods Applied to the Automatic Diagnosis of Sleep Apnea

8

Gonzalo C. Gutiérrez-Tobal, Daniel Álvarez, Fernando Vaquerizo-Villar, Verónica Barroso-García, Javier Gómez-Pilar, Félix del Campo, and Roberto Hornero

Abstract

The overnight polysomnography shows a range of drawbacks to diagnose obstructive sleep apnea (OSA) that have led to the search for artificial intelligence-based alternatives. Many classic machine learning methods have been already evaluated for this purpose. In this chapter, we show the main approaches found in the scientific literature along with the most used data to develop the models, useful and large easily available databases, and suitable methods to assess performances. In addition, a range of results from selected studies are presented as examples of these methods. Very high diagnostic performances are reported in these results regardless of the approaches taken. This leads us to conclude that conventional machine learning methods are useful techniques to develop new OSA diagnosis simplification proposals and to act as benchmark for other more recent methods such as deep learning.

Keywords

Sleep apnea · Machine learning · Sleep Heart Health Study · Childhood Adenotonsillectomy Trial · Classification · Regression · Biomedical signal processing · Airflow · Blood oxygen saturation · Electrocardiogram

8.1 Introduction

The technical complexity, costs, and logistic-associated problems in the diagnosis of obstructive sleep apnea (OSA) have driven the scientific community to search for new simpler and automatic alternatives to standard polysomnography (PSG) (Ghegan et al., 2006). One of the most common and ambitious approximations to achieve this goal has been the implementation of systems or algorithms based on the study of a reduced set of information from the PSG. Usually,

G. C. Gutiérrez-Tobal (✉) · D. Álvarez · F. Vaquerizo-Villar · V. Barroso-García · J. Gómez-Pilar · R. Hornero
Centro de Investigación Biomédica en Red, Bioingeniería, Biomateriales, Nanomedicina, Madrid, Spain

Biomedical Engineering Group, University of Valladolid, Valladolid, Spain
e-mail: gonzalo.gutierrez@gib.tel.uva.es

F. del Campo
Centro de Investigación Biomédica en Red, Bioingeniería, Biomateriales, Nanomedicina, Madrid, Spain

Biomedical Engineering Group, University of Valladolid, Valladolid, Spain

Sleep Unit, Pneumology Service, Hospital Universitario Rio Hortega, Valladolid, Spain

T. Penzel, R. Hornero (eds.), *Advances in the Diagnosis and Treatment of Sleep Apnea*, Advances in Experimental Medicine and Biology 1384, https://doi.org/10.1007/978-3-031-06413-5_8

the automatic analysis of only a very small set of signals – out of a maximum of 32 recorded during PSG – has been conducted, with the investigation on a single one being a very frequent approach (Uddin et al., 2018; Mendonça et al., 2019; Gonzalo C Gutiérrez-Tobal et al., 2021c). In this regard, overnight blood oxygen saturation (SpO_2), airflow (AF), and electrocardiogram (ECG) are among the most analyzed signals (Uddin et al., 2018; Mendonça et al., 2019; Gutiérrez-Tobal et al., 2021c).

Since the beginning of the twenty-first century, machine learning techniques have gained an increasing role when improving the performance of automatic health-related diagnostic tools, and OSA has not been an exception. A three-step methodology, the so-called feature engineering approach, has been traditionally applied to the problem (Vaquerizo-Villar et al., 2021). This strategy begins with the "feature extraction" stage, in which the data – usually an overnight signal – are analyzed following one or several complementary analytical techniques, such as spectral, non-linear, or time-frequency methods. The purpose of this step is to characterize the signal or signals under study, so that the original raw or pre-processed data becomes information of interest for the problem. Then, an optional but useful automatic "feature selection" stage is conducted to ensure that all the information extracted in the previous step is as relevant for your problem and as complementary to each other as possible (Guyon & Elisseeff, 2003). The third stage is what involves machine learning. It could be termed simply "machine learning" stage or, depending on the context, "classification" stage, "regression" stage, or, in a more general way, "pattern recognition" stage (Bishop, 2006).

Certainly, the latest deep learning methods are able to avoid the two first stages in the above-described feature engineering approach (Ian et al., 2016). However, many traditional methods are still used nowadays in the context of OSA diagnosis and constitute a valid and very useful benchmark to compare the results obtained with any new approximation to the problem. Accordingly, this chapter aims at exposing the interested readers to a set of conventional machine learning tools that have proven their usefulness to help in the automatic OSA diagnosis. As shown in the next sections, it is not a minor challenge to outperform some of these methods, so any new algorithm must demonstrate that demanding an extra effort, on either data or computation, is justified.

This chapter continues with a data section, which is dedicated to briefly present the information traditionally analyzed in the OSA diagnosis simplification. Then, a methods section explains the two main machine learning approaches (classification and regression) that have shown usefulness in this problem. It includes introducing some specific examples used in OSA context along with a brief explanation on their rationale, as well as appropriate references where the readers will be able to gain insight into these methods. Next, a results section shows some of the highest performances achieved using them. Finally, the "Discussion and Conclusions" section analyzes the most important information included in this chapter.

8.2 Data Analyzed in the Simplification of Sleep Apnea Diagnosis

Several chapters of this book are specifically devoted to describing useful sources of information in the context of sleep apnea. Therefore, this section is only a short introduction on those that have been more frequently used along with machine learning approaches. These include some overnight biomedical signals recorded during PSG and other clinical and demographic data. In addition, we present some popular public databases that have been used in dozens of different studies to gain insight into sleep apnea in both adults and children.

8.2.1 Typical Overnight Biomedical Signals

AF, SpO_2, and ECG (including the ECG-derived heart rate variability or HRV) have been extensively analyzed in the context of simplifying OSA diagnosis in the last decades (Uddin et al., 2018; Mendonça et al., 2019; Gutiérrez-Tobal et al., 2021c). Usually, the recordings are acquired during the night with the same equipment used in the PSG, but there exist a substantial number of scientific studies using devices specifically dedicated to acquiring each signal alone. In addition, the most common approach has focused on the analysis of single-channel signals, but some studies also analyzed the usefulness of automatically combining the information from two or more of them.

8.2.1.1 Airflow (AF)

As explained in dedicated chapters of this book, one of the most important indicators of the presence and severity of OSA is the apnea-hypopnea index (AHI) (Iber et al., 2007; Berry et al., 2012, 2017). AHI accounts for the number of apnea – complete cessation of the respiratory cycle – and hypopnea events, significant reduction of the respiratory amplitude, per hour of sleep (Berry et al., 2012). These qualitative definitions of apneas and hypopneas are detailed in the rules for scoring respiratory events published and updated by the American Academy of Sleep Medicine (Iber et al., 2007; Berry et al., 2012, 2017). A reduction of 90% in AF is mandatory to annotate an apnea event (Berry et al., 2012), showing a duration of a minimum of two respiratory cycles in pediatric patients and 10 seconds in adults. In the case of hypopneas, a 30% drop in AF suffices, but the event needs to be accompanied by either a 3% drop in the SpO_2 signal or an arousal (Berry et al., 2012). The minimum duration requirement of the AF drop is also two respiratory cycles for children and 10 seconds for adults. According to these definitions, in which AF plays a key role, the study of this signal is a natural choice to search for simpler OSA diagnostic alternatives.

When scoring these respiratory events, it is necessary to consider that apneas must be counted using an oronasal thermal sensor, whereas hypopneas are annotated using a nasal pressure sensor (Berry et al., 2012). This is because of the complementary performances of these two kinds of probes when detecting each of the event types (Bahammam, 2004). This also needs to be considered when using machine learning techniques that only focus on detecting apneas and hypopneas. However, in machine learning approaches not conducting event detection, but full characterization of the overnight AF signal, recent studies have shown similar performances using single-channel AF approaches regardless if thermal or nasal pressure sensors were used (Gutiérrez-Tobal et al., 2013; Gutierrez-Tobal et al., 2016).

8.2.1.2 Blood Oxygen Saturation (SpO_2)

Blood oxygen drops – or desaturations – are typical effects caused by apneic events (Iber et al., 2007; Berry et al., 2012, 2017). Actually, we have already shown that 3% desaturations are directly involved in the hypopnea definition. Additional important advantages need to be considered that have led SpO_2 to be probably the most analyzed and successful signal when simplifying OSA diagnosis, in both adults and children. The first one is that it is easily acquired using a single-channel pulse oximetry placed on a finger (or a toe in babies). This very comfortable compared to all the channels required to conduct a full PSG. As a result, the associated portable technology is highly developed, which facilitates to move the diagnostic test to patients' homes. A second advantage is that the overnight blood oxygen saturation gathers not only the information regarding the apneic events but also the health prognosis associated with the condition. In this regard, 3% and 4% oxygen desaturation indices (ODI3 and ODI4), cumulative time under 90% of saturation (CT90), or, more recently, hypoxic burden have been linked to different negative health consequences in OSA presence (Azarbarzin et al., 2019; Karhu et al., 2021). Finally, as shown in the next sections, the results reached when applying machine learning methods to SpO_2 are among the highest in the related scientific literature.

8.2.1.3 Electrocardiogram and Heart Rate Variability (ECG/HRV)

The natural cardiorespiratory coupling is one of the main reasons behind the study of ECG to help simplify OSA diagnosis. This coordination has been found to increase during the night in the presence of sleep apnea (Riedl et al., 2014), being one of its expressions the occurrence of a clear bradycardia/tachycardia pattern following the apneic events (Penzel et al., 2003). Moreover, the ECG was one of the first biomedical signals studied, and it is still one of the most analyzed in different health contexts, which very often provides a comfortable scientific knowledge background on which to justify the interpretations of eventual results (Acharya et al., 2006). Similarly, OSA in adults is known to be significantly associated with cardiovascular morbidity (Newman et al., 2001). Together, these aspects have led to an intensive scientific activity regarding the simplification of OSA diagnosis based on ECG information (Penzel et al., 2002). Particularly common has been the investigations on HRV, which offers a nexus between OSA and the autonomic nervous system (Acharya et al., 2006). An additional advantage of the HRV information is that it can be surrogated in some contexts by the pulse rate variability signal (PRV) (Gil et al., 2010), which can be easily obtained from a pulse oximeter.

8.2.2 Other Sources of Information

The clinical analysis of PSG is the result of the examination of a range of up to 32 biomedical channels. Consequently, it is not surprising that several approaches explored the combination of the information from two or three of the abovementioned biomedical signals along with the use of machine learning techniques (Garde et al., 2014; Álvarez et al., 2020; Jiménez-García et al., 2020). In addition, other single- or combined-channel approaches have been evaluated. In this regard, the use of overnight snoring sounds (Solà-Soler et al., 2012), thoracic and/or abdominal movements (Lin et al., 2017), photoplethysmography (Gil et al., 2010; Lázaro et al., 2014), or the electroencephalography (Gonzalo C. Gutiérrez-Tobal, Gomez-Pilar, et al., 2021b), among others, have been also explored with promising results.

Moreover, machine learning has been also used with data other than those from PSG. Demographic, social, clinical, and anthropometric variables have been also used as source of information to train machine learning models with ability to diagnose OSA (El-Solh et al., 1999; Skotko et al., 2017; Gonzalo C Gutiérrez-Tobal et al., 2021c). These have been used most often combined within them and with the information obtained from the PSG, such as overnight biomedical signals.

8.2.3 Important Databases

Large and commonly used databases are very useful both to properly train and validate the machine learning models and to share a reference to which compare the performance from different methods. Unfortunately, freely available large databases are very uncommon in OSA context, if there exist. However, the National Research Sleep Resource offers several very large sleep-related databases with only minor requirements to be accomplished. Here, we briefly introduce two of them that have been used in dozens of OSA-related studies from adults and children, namely, the Sleep Heart Health Study (SHHS) database and the Childhood Adenotonsillectomy Trial (CHAT) database, respectively.

8.2.3.1 Sleep Heart Health Study (SHHS)

The SHHS was originally designed to evaluate whether OSA is an independent risk factor for the development of cardiovascular morbidity in adults (Newman et al., 2001). The database comprises at-home conducted PSGs from 5804 individuals older than 40 years who were recruited from several previous cohorts aimed at evaluating cardiovascular risks (Quan et al., 1997). It is divided into SHHS1, with a first round of sleep data and recordings from all the participants, and SHHS2, with a follow-up at-home PSG conducted on 2647 participants 5 years later. Accordingly, longitudinal studies are possible when using this database. In total, 8451 full PSGs are available to use it as source of information in machine learning-based studies, including annotations such as respiratory events or sleep stages, along with a wide range of clinical, social, and anthropometric variables (Quan et al., 1997; Newman et al., 2001).

8.2.3.2 Childhood Adenotonsillectomy Trial (CHAT)

The aim of the CHAT randomized study was to analyze the effects of a treatment based on the removing of tonsils and adenoids in a cohort of OSA-affected children (Marcus et al., 2013). To assess these effects, PSGs from 1447 children between 5 and 9 years were conducted, from 464 who were randomized to adenotonsillectomy treatment (206 children) or the alternative watchful waiting with supportive care (198 children) (Marcus et al., 2013). Accordingly, these participants underwent a baseline PSG and a follow-up PSG 7 months later, once completing the treatment or the alternative. A wide range of clinical, sociodemographic, cognitive, and anthropometric variables is also available (Marcus et al., 2013). As in the case of SHHS, the follow-up conducted on the children allows for longitudinal studies taking into account that there is a therapeutic intervention between the two PSGs. In addition to the randomized children, the PSGs from the non-randomizing are also available to develop the machine learning approaches. However, the set of additional variables is dramatically reduced compared to the randomized set.

8.3 Methods: Classic Machine Learning Approaches in Sleep Apnea Diagnosis

In accordance with the purpose of automatically diagnosing OSA, supervised learning is the most common strategy followed in the scientific literature. Particularly, both classification and regression approaches have been frequently implemented. OSA presence and severity are routinely categorized by using AHI thresholds in clinical practice, which leads to classification methods. Moreover, AHI can be also directly estimated, thus leading to regression approaches. In this section, we also introduce the ways in which the performance of the OSA-related machine learning methods should be assessed for both classification and regression.

8.3.1 Classification

There are two typical ways to implement classification approaches in OSA diagnosis context: binary classification and multiclass classification. In addition, these may have different purposes. On the one hand, classification may focus on directly assigning subjects into two (presence vs. absence of OSA) or more (presence and severity of OSA) categories. This should be the final goal of any automatic diagnostic approach. On the other hand, however, classification may also focus on detecting apneic events, and this can be also implemented as binary classification (apneic vs. normal signal segments) or multiclass classification (apneas/hypopneas/normal or obstructive apneas/central apneas/normal, etc.).

8.3.1.1 Binary Classification

Over the years, the clinicians have focused on AHI thresholds to assess whether a person suffers from OSA. Ten and 15 events per hour (e/h) have been commonly used in adults, and 1 e/h, 3 e/h, and 5 e/h in children, the exact cut-off evolving as the corresponding medical associations proposed new rules (Iber et al., 2007; Berry et al., 2012, 2017). In accordance with these thresholds, one of the machine learning approaches has focused on automatically detecting the presence of the illness, that is, classifying subjects into OSA positive (above or equal the AHI cut-off) or OSA negative (below the AHI cut-off). Different classic machine learning methods have been used to implement this approach. Linear discriminant analysis (LDA) is one of the most typical classification procedures (Bishop, 2006) and has been evaluated in both adults and children in OSA context. LDA assumes a linear relationship between the predictors (variables used as the data to predict OSA) and the target (the variable containing the OSA-positive and OSA-negative labels). Despite its relatively simplicity, LDA has reached promising results when discriminating OSA-positive and OSA-negative patients using information from SpO_2 (Marcos et al., 2009), SpO_2 + PRV (Garde et al., 2014), and HRV (Martín-Montero et al., 2021). Logistic regression (LR) (Hosmer & Lemeshow, 1989) is a standard in binary classification and has been also evaluated with SpO_2 (Marcos et al., 2009; Álvarez et al., 2010, 2013) and AF (Barroso-García et al., 2017), in both adults and children. LR uses the logistic formulae to transform the output resulting from a linear regression into a non-linear posterior probability (Hosmer & Lemeshow, 1989), that is, given the predictors, the probability of belonging to the OSA-positive class – as defined by the AHI cut-off used. Accordingly, LR avoids the limitation of the linear relationship assumption.

This limitation can be also minimized with more complex and modern methods such as artificial neural networks (ANNs) and support vector machines (SVMs) (Bishop, 2006). SVMs are machine learning algorithms that transform the data into a higher-dimensional space so that the distance between data points with different labels – in this case, OSA positive and negative – is maximized (Bishop, 2006). This is equivalent to choosing a decision boundary between classes for which the distance to the closest data point, the so-called margin, is maximized (Bishop, 2006). Accordingly, the decision boundary is defined by several of these data points termed support vectors. Some examples of SVM binary classification in OSA context can be found applied to SpO_2 (Álvarez et al., 2013) and ECG (Khandoker et al., 2009; Chen et al., 2015). On the other hand, several ANNs have been evaluated in OSA binary classification approaches (Marcos et al., 2008; Morillo & Gross, 2013), being multi-layer perceptron (MLP) one common approach that has become one of the most successful machine learning methods in any problem. ANNs are algorithms inspired in the biological neural networks, such as the human brain. Accordingly, MLP arrange computing units, also known as perceptrons or neurons, in several massively connected layers: input, hidden, and output (Bishop, 2006). The input layer is composed of one neuron for each feature or variable used as predictor. These input neurons are connected through weights with all the neurons in the next layer, which is part of the hidden layers. There can be as many hidden layers as the designers may consider appropriate. However, one single hidden layer is known to be able to provide universal approximations (Bishop, 2006). This means that, provided that your data gather information enough for your problem, one single hidden layer should suffice to model the function that transform your predictors into your desired target. In any case, both the number of hidden layers and the number of neurons per hidden layer are hyperparameters of the model to be tuned during the training process. Finally, each neuron of the last hidden layer – if there is more than one – is connected to all the neurons in the output layer, which in the case of the binary classification approach is a single neuron that offers the posterior probability of belonging to the OSA-positive class. During the training process of the MLP (and other ANNs), all the weights connecting all the neurons of the network are optimized using the well-known backpropagation algorithm (Bishop, 2006), which is one of the most remarkable milestones of machine learning. Another feature of ANNs is that each neuron has an associated activation function that combines the outputs – including weights – from previous layers into a single output, being logistic or softmax functions typically used in classification and linear functions in regression problems (Bishop, 2006).

8.3.1.2 Multiclass Classification

In recent years, as more sleep data has been available for scientific purposes, the focus of OSA diagnosis simplification has gone from binary classification to the determination of both OSA presence and severity, which naturally fits multiclass classification. There exist AHI thresholds for the definition of OSA severity categories in both adults and children, being the latter much more restrictive. Nowadays, the most clinically used ones are probably as follows (Flemons et al., 1999; Tan et al., 2014, 2017):

- *Adults*: no OSA if AHI < 5 e/h; mild OSA if 5 e/h ≤ AHI < 15 e/h; moderate OSA if 15 e/h ≤ AHI < 30 e/h; and severe OSA if 30 e/h ≤ AHI
- *Children*: no OSA if AHI < 1 e/h; mild OSA if 1 e/h ≤ AHI < 5 e/h; moderate OSA if 5 e/h ≤ AHI < 10 e/h; and severe OSA if 10 e/h ≤ AHI

As in the case of binary classification, several multiclass approaches have been already evaluated in OSA context. LDA and LR models were also used in the multiclass problem along with SpO_2 data (Gutiérrez-Tobal et al., 2019), the latter needing an additional "one-vs.-all" strategy to upgrade the binary approach. MLP and other ANNs have been also developed with both SpO_2 data (Gutiérrez-Tobal et al., 2019), SpO_2 + AF data (Barroso-García et al., 2021), and clinical, anthropometric, and demographic variables (Skotko et al., 2017). In this regard, from an implementation point of view, only minor changes in the architecture are needed to develop multiclass ANNs instead of binary ones, such as equaling the number of output neurons to the number of classes. The interested readers should notice, however, that data requirements usually increase as more classes are targeted and that multiclass overall performance tends to be lower than the binary one.

Ensemble learning methods have been also used to address the multiclass problem. As deduced from its name, this family of machine learning methods conduct the classification task as the result of the combination of the classification of several single models, typically termed "base classifiers." These can be any of the abovementioned methods, but simpler ones are preferred to increase the generalization ability of the final classification (Witten et al., 2011). Bagging ensemble learning algorithms have been tested, including the remarkable random forest (RF) method used with SpO_2 data (Deviaene et al., 2019). Bagging is the acronym for "bootstrap aggregating," which indicates the basic methodology behind this method. In essence, the original data is subsampled with replacement to form, typically, a high number of bootstrap replicates of these data (Kuncheva, 2014). A different classifier is trained for each of these replicates, and its decision is only one vote for the final classification task, which is conducted based on the decisions from all classifiers. RF follows this elementary scheme using decision trees as base classifiers. In addition, RF includes more sources of variability in the training of its classifiers by randomly varying the features and the decision trees hyperparameters involved within each bootstrap iteration (Kuncheva, 2014). Boosting ensemble learning methods have been also applied in OSA-related multiclass tasks, as is the case of the well-known AdaBoost (for "adaptive boosting"), used with SpO_2 (Gutiérrez-Tobal et al., 2019), AF (Gutierrez-Tobal et al., 2016), and SpO_2 + AF (Jiménez-García et al., 2020; Barroso-García et al., 2021). In contrast to bagging, boosting methods are iterative algorithms in which each new classifier is trained using the same data, but accounting for the errors made by previous classifiers. In this regard, misclassified data points in previous iterations are weighted to give them more importance, thus increasing the chances to be rightly classified in the current and next iterations (Witten et al., 2011). Another difference with bagging is that the vote of each classifier is dependent on its error so that the ones with higher performance contribute more to the final decision (Witten et al., 2011).

8.3.2 Regression

The automatic AHI estimation is another popular approach when simplifying OSA diagnosis. Instead of training machine learning methods to directly assign subjects (or epochs) into different OSA severity categories (or events), this strategy looks for assigning an AHI to each subject. As the clinical use of AHI thresholds has evolved over the years, and there are still some limitations regarding the OSA severity categories and the actual health state of the patients (Penzel et al., 2015; Korkalainen et al., 2019), the AHI estimation has the advantage of being relatively transparent to future changes in thresholding criteria.

There exists an extensive literature focused on regression methods and OSA diagnosis simplification. They focus on both simpler algorithms, such as multiple linear regression applied to clinical data (Wu et al., 2017) and more complex methods already mentioned such as MLP or SVM applied to clinical (El-Solh et al., 1999), SpO_2 (Marcos et al., 2012; Hornero et al., 2017; Rolón et al., 2017; Xu et al., 2019; Rolon et al., 2020), AF (Álvarez et al., 2020; Barroso-García et al., 2021), and SpO_2 + AF data (Álvarez et al., 2020; Barroso-García et al., 2021). Moreover, boosting ensemble learning methods have been also evaluated, such as least-square boosting (LSBoost) using SpO_2 information (Gonzalo C. Gutiérrez-Tobal, Álvarez, et al., 2021a). In this regression task, rather than focusing on previously misclassified data points, the boosting algorithm LSBoost looks for computing the remaining residual error (between the actual and the estimated AHI) that was not able to be estimated in previous iterations (Bühlmann & Hothorn, 2007).

8.3.3 Machine Learning Performance Assessment and Validation

8.3.3.1 Underfitting and Overfitting

Machine learning faces two main issues regardless of the problem and the approach considered. The first one, underfitting, relates to the inability of the method to learn the function it is intended for. Two aspects are often behind underfitting, unsuitable learning algorithm or unsuitable input information, the latter caused by either data scarcity or low quality. However, provided that a proper study design has been conducted, underfitting is not the main drawback that machine learning may confront. Very often, overfitting is a much more challenging aspect. Overfitting refers to an excessive fitting of the machine learning algorithm to the training sample, thus resulting in poor generalization ability when evaluated in new (test) data (Bishop, 2006). Most of machine learning methods can be affected by overfitting. Accordingly, several strategies can be followed to minimize this effect. Increasing the size of the training set is a common option but it is not always possible. More usual are the methods based on "regularization." Under this name, there are a wide range of strategies based on adding a penalty term to the learning of the method during the training process, so that it can model a more general function instead of adjusting to the particularities of the training data (Bishop, 2006). The interested reader should be aware that this is not an issue that can be obviated, especially when using relatively complex methods such as ANNs or SVM. Even the ensemble learning methods, which have a natural well-known robustness against overfitting (Witten et al., 2011), can benefit from using regularization techniques (Bühlmann & Hothorn, 2007).

8.3.3.2 Validation Strategy

There exists an intimate relationship between overfitting and the way in which the machine learning models should be validated. As the risk for overfitting exists, evaluating the models using training data would most probably lead to overoptimistic performance results (Bishop, 2006; Witten et al., 2011). For the same reason, the hyperparameters needed for some of the abovementioned methods – including the regularization term – should be chosen based on the results from an independent dataset. Finally, a reliable performance should be derived from a third previously unseen (test) dataset. This would be a classic and robust validation strategy, which would include a training group for model parameter estimation, a validation group for hyperparameter tuning, and a test group for assessing the performance of the final model.

Ideally, there should be a validation group for each freedom degree of the machine learning method used, including hyperparameters and possible previous feature selection stages. However, data scarcity is very common in healthcare problems, and the use of only three groups (training/validation/test) is usually accepted. Several cautions need to be considered when distributing the data among these three groups. First, the more the data in your training set, the better the chances for developing a more accurate model. Second, the data distribution of the validation and test groups should be as similar as possible. This means that if your test group has 50% OSA-positive and 50% OSA-negative subjects, your validation group should have similar proportions. Third, machine learning methods tend to favor the correct classification of the majority classes in classification problems, as well as the range of values more represented in regression problems. Consequently, provided that the classification of all your classes (or the estimation within all range of values) is equally important, it is also advisable that your training data is well-balanced. Finally, commonly used proportions in data distribution include 60–80% for training and 10–20% for each of the validation and test groups. However, in the rare cases in which data availability is not a problem, these proportions can vary if the other advice is considered.

A final consideration is needed regarding data scarcity. As mentioned above, healthcare-related machine learning problems tend to lack data, and OSA diagnosis simplification is not an exception (Gonzalo C Gutiérrez-Tobal et al., 2021c). Accordingly, split data in three independent groups is not often possible. A usual solution is to emulate one of the groups (typically the validation or the test group) using statistical methodologies such as bootstrapping, jackknife, leave-one-out cross-validation, or k-fold cross-validation (Bishop, 2006; Witten et al., 2011).

8.3.3.3 Performance Statistics

Performance assessment in binary problems is based on different combinations of the number of true positive (TP), false negative (FN), true negative (TN), and false positive (FP) subjects or events. In this sense, sensitivity (Se, also known as recall), specificity (Sp), and accuracy (Acc) are important metrics to evaluate the percentage of positive, negative, and total number of subjects/events rightly classified, respectively:

$$\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} * 100 \tag{8.1}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} * 100 \tag{8.2}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} * 100. \tag{8.3}$$

Useful statistics are also positive and negative predictive values (PPV, also known as precision, and NPV), which account for the percentage of success when assigning a data point within one class (e.g., positive) or the another (e.g., negative):

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} * 100 \tag{8.4}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} * 100. \tag{8.5}$$

Moreover, positive and negative likelihood ratios (LR+ and LR−) account for the ratios of the true positive rate to the false positive rate and the false negative rate to the true negative rate, respectively. In the next definitions, Se and Sp are also taken as rates instead of percentages:

$$\mathrm{LR}+ = \frac{\mathrm{Se}}{1 - \mathrm{Sp}} \tag{8.6}$$

$$\mathrm{LR}- = \frac{1 - \mathrm{Se}}{\mathrm{Sp}}. \tag{8.7}$$

These metrics, however, are affected by class imbalance to some extent. Therefore, they are often complemented with the receiver-operating characteristics (ROC) analysis (Zweig & Campbell, 1993). ROC is based on a plot representing Se vs. 1-Sp (in unit proportion) computed for a range of possible decision thresholds from the same output, which in the case of binary machine learning could be the posterior probability of belonging to the class of interest. One possible application of this analysis is the estimation of a suitable threshold to act as a trade-off between Se and Sp (Zweig & Campbell, 1993), i.e., the threshold that minimizes biases due to class imbalance. Other possible uses include to measure the overall performance of a model and, in turn, the comparison of the performance of different models. In this sense, the perfect performance would be achieved by a machine learning model that reaches the point of the plot Se = 1 and 1-Se = 0.

To properly quantify the overall performance, however, it is common to estimate the area under the ROC curve (AROC), which may range between 0 and 1, showing AROC = 0.5 the less discriminative power (Zweig & Campbell, 1993).

All the abovementioned metrics can be also used for evaluating the performance of a multiclass classification approach in each of the thresholds used to determine OSA severity categories. In addition, specific statistics can be used to assess the overall performance in the multiclass problem. Cohen's kappa, which can be also used in binary classification, is one of the most helpful as it measures the agreement between the actual and the estimated class by correcting it by the agreement occurred by chance (Witten et al., 2011). Values closer to 1 (or 100%) mean higher agreement, whereas values closer to 0 indicate lower agreement. Acc adapted to the number of classes is another useful metric to evaluate multiclass performance.

As the definition of OSA severity classes, either binary or multiclass, is conducted based on AHI, the corresponding assessing metrics can be also used to evaluate regression approaches provided that the estimated AHI is properly transformed into the OSA-related classes. Moreover, there exist specific analytical tools to evaluate the similarity between the estimated and the actual AHI. One typical example is the intraclass correlation coefficient (ICC) (Chen & Barnhart, 2008), which measures the agreement between continuous variables. Accordingly, values closer to 1 indicate higher degree of agreement, whereas values closer to 0 mean lower degree of agreement. However, contrary to other statistics such as Pearson's correlation, ICC accounts for systematic errors to estimate agreement (Chen & Barnhart, 2008). Finally, a typical and very useful method to graphically assess the agreement in AHI estimations is the Bland-Altman plot (Bland & Altman, 1986). This method shows the difference between the estimated and the actual continuous variable against the mean of the two values (Bland & Altman, 1986). In addition, it provides possible bias for the estimation (the mean of the differences of all data points) and the limits of the agreement (mean ± 1.96*standard deviation of the differences of all data points). These limits are useful to evaluate whether the estimation can be used as a surrogate for the actual continuous variable (Giavarina, 2015).

8.4 Selected Results from the Literature

Table 8.1 displays some results selected from the literature regarding classic machine learning performance in OSA diagnosis simplification. Showing as many approaches as possible has been one important objective when selecting the results to be included in the table. Accordingly, there are studies focused on adults and children, using different overnight signals (alone and combined) and clinical data, and up to eight different machine learning methods. Validation strategies also vary among studies. In addition, these works are divided into the three main approaches explained above: binary classification, multiclass classification, and regression. The metrics included in the table have been chosen as a tradeoff between those reported in the studies and those highlighted as important in the previous sections. An interesting point is the range of methods that can be used to evaluate performance in each machine learning approach. Whereas binary classification is limited to very specific statistics, multiclass classification and, specially, regression approaches can be assessed with an increasing number of methods, thus providing a more complete picture of their performance. Unfortunately, not all the studies provided data to show or estimate all the statistics. Moreover, in some of the studies, the machine learning task focuses on the subjects, whereas in others it focuses on the apneic events. However, as we think that the most valuable approach implies to provide a final diagnosis, we only show those results that end up assigning subjects into one OSA class, regardless the specific purpose of the machine learning method.

As observed, very high diagnostic performance can be achieved using classic machine learning methods. Similarly, all the data involved

Table 8.1 Reported diagnostic performance in selected previous studies following binary classification, multiclass classification, and regression approaches

Study	Population (*N*)	Data source	ML method	Validation	AHI (e/h)	Se (%)	Sp (%)	Acc (%)	ICC	Kappa[a]
Binary classification										
Khandoker et al. (2009)	Adults (125)	ECG	SVM	loo-cv + test	10	92.3	93.8	92.9	–	–
Álvarez et al. (2013)	Adults (316)	SpO_2	SVM	Training + validation + test	10	95.2	80.0	84.5	–	–
Garde et al. (2014)	Children (146)	SpO_2 + PRV	LR	loo-cv + -fold cv	5	88.4	83.6	84.9	–	–
Martín-Montero et al. (2021)	Children (1738)	HRV	LDA	Training + test	5	63.8	84.7	82.8	–	–
Multiclass classification										
Gutierrez-Tobal et al. (2016)	Adults (317)	AF	AdaBoost	Training + Bootstrap + test	5	87.1	80.0	86.5	–	0.432
					15	85.9	72.9	81.0		
					30	74.2	90.6	82.5		
Skotko et al. (2017)	Children (102)	Clinical	LLM	Cross-validation	1	75.6	50.9	61.8	–	–
					5	72.2	54.8	57.8		
Deviaene et al. (2019)	Adults (5793)	SpO_2	RF	Training + 10-fold cv + test	5	83.5	88.0	84.3	–	0.547
					15	75.6	95.8	87.0		
					30	77.3	97.7	94.3		
Jiménez-García et al. (2020)	Children (974)	SpO_2 + AF	AdaBoost	Training + bootstrap + test	1	89.2	37.3	79.2	–	0.398
					5	76.0	85.7	82.1		
					10	62.7	97.7	90.3		
Regression										
El-Solh et al. (1999)	Adults (269)	Clinical	MLP	10-fold cv	10	94.9	64.7	–	0.850[b]	–
					15	95.3	60.0	–		
					20	95.5	73.4	–		
Hornero et al. (2017)	Children (4191)	SpO_2	MLP	Training + loo-cv + test	1	84.0	53.2	75.2	0.785	0.348
					5	68.2	87.2	81.7		
					10	68.7	94.1	90.2		
Álvarez et al. (2020)	Adults (239)	SpO_2 + AF	SVM	Training + loo-cv + test	5	95.6	83.3	94.8	0.930	0.610
					15	96.0	72.7	90.6		
					30	93.6	98.0	95.8		

(continued)

Table 8.1 (continued)

Study	Population (*N*)	Data source	ML method	Validation	AHI (e/h)	Se (%)	Sp (%)	Acc (%)	ICC	Kappa[a]
Gutiérrez-Tobal, Álvarez, et al. (2021a)	Adults (5793)	SpO_2	LSBoost	Training + validation + test	5	93.8	56.3	89.2	0.900	0.561
					15	87.0	84.1	85.3		
					30	82.2	96.3	94.6		

ML machine learning, *AHI* apnea-hypopnea index, *Se* sensitivity, *Sp* specificity, *Acc* 2 classes accuracy, *ICC* intraclass correlation coefficient, *ECG* electrocardiogram, *SVM* support vector machines, *loo-cv* leave-one-out cross-validation, SpO_2 oxygen saturation, *PRV* pulse rate variability, *LR* logistic regression, *HRV* heart rate variability, LDA linear discriminant analysis, *AF* airflow, *LLM*, logic learning machine, *RF* random forest, *MLP* multi-layer perceptron

[a] 4 class Cohen's kappa

[b] Pearson's correlation

in the studies can reach high statistics. The reader should notice that the results from those works with higher number of participants – and more independent groups in the validation strategy – should be initially considered as more robust. We kindly invite them to examine the original studies to evaluate whether this assumption is true. It is also observed that those studies focused on children achieved lower diagnostic performances. The reader can also check in the literature that this is not an effect due to the non-systematic selection of the studies, but a general tendency. Traditionally, the study of pediatric OSA has gathered much less attention than adult OSA. Consequently, efforts, resources, and data have been scarce, thus resulting in lower knowledge compared to adult OSA. In addition, the AHI rules for establishing pediatric OSA are tighter. All these limitations have favored that there is still a gap between the performances reached in adults and children.

Among the studies, Martin-Montero et al. (Martín-Montero et al., 2021) involve the non-randomized group of the CHAT database along with a private database from the University of Chicago, USA. Similarly, Deviaene et al. and Gutiérrez-Tobal et al. (Deviaene et al., 2019; Gonzalo C. Gutiérrez-Tobal, Álvarez, et al., 2021a) involved the SHHS database. For the sake of simplicity, only results from 5793 subjects (the SHHS1 subgroup) are shown. However, the studies also reported diagnostic results from the follow-up subgroup (SHHS2) with 2647 recordings and, in the case of Gutiérrez-Tobal et al., a high pre-test probability subgroup with 322 recordings from Hospital Universitario Rio Hortega from Valladolid, Spain.

8.5 Discussion and Conclusions

In this chapter, we focused on the most typical classic machine learning approaches involved in OSA diagnosis simplification, thus setting aside deep learning techniques. We have shown specific machine learning methods, the data regularly used with them, as well as large and easily available adult's and children's databases. We have also exposed useful strategies to measure and validate the performance of the machine learning methods, and we have shown a variety of studies in which this performance is high.

One first take-away idea to be highlighted is that there exists a wide range of successful machine learning methods applied to OSA diagnosis simplification. They covered both classification (either binary or multiclass) and regression approaches, the latter being more easily evaluated in depth. Other interesting key point is that many of the data from the PSG (SpO_2, AF, ECG/HRV, etc.) gather information enough to obtain accurate machine learning methods, as reflected by the high diagnostic performances shown in the studies involved in Table 8.1, and in many others referenced within this chapter. This implies that those methods to be evaluated in the future would need to not only justify a possible increase in the performance but also the eventual rise in data requirements and computational costs.

The studies we chose as examples also reflect a lack of homogeneity in the validation strategy. This is an issue that is also present in the scientific literature (Gonzalo C Gutiérrez-Tobal et al., 2021c) and hinders the comparison between the different methods. As mentioned in the past sections, the ideal training/validation/test strategy is greatly influenced by data scarcity. Accordingly, the problem is closely related to the different sample sizes of the studies, which involve a number of subjects ranging from moderate (102) to high (5793). This underlines the need to make available for the scientific community large databases such as CHAT and SHHS.

The previous idea is particularly important in the case of pediatric OSA. The gap between the machine learning performance in adults and children can be partially attributed to the more restrictive diagnosis rules for children. However, large samples such as CHAT can be very useful to increase the knowledge of pediatric OSA and develop more accurate machine learning models.

Finally, despite the high performance shown in several of the studies referenced in this chapter, it is very difficult to find machine learning-based systems implemented in real clinical environments. One possible reason for this issue is that clinicians and healthcare providers and managers perceive these methods as a black box, thus preventing them from completely relying on their predictions (Gonzalo C Gutiérrez-Tobal et al., 2021c). Accordingly, the machine learning designers who expect their work to be finally implemented will need to put extra efforts in explaining the decisions taken by their automatic models (Adadi & Berrada, 2018).

To sum up, traditional machine learning methods have proven to be very useful in the automatic OSA diagnosis simplification. Accordingly, they are still valid options both to develop new proposals and to act as benchmark for future methods.

Acknowledgments This work was supported in part by "Ministerio de Ciencia e Innovación-Agencia Estatal de Investigación" and the "European Regional Development Fund (FEDER)" under projects PID2020-115468RB-I00 and RTC-2017-6516-1, in part by the "European Commission" and "FEDER" under project "Análisis y correlación entre la epigenética y la actividad cerebral para evaluar el riesgo de migraña crónica y episódica en mujeres" ("Cooperation Programme Interreg V-A Spain-Portugal POCTEP 2014–2020"), in part by Sociedad Española de Neumología y Cirugía Torácica (SEPAR) under project 649/2018, in part by Sociedad Española de Sueño (SES) under project "Beca de Investigación SES 2019," and in part by "Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Spain," through "Instituto de Salud Carlos III" co-funded with FEDER funds. The work of Fernando Vaquerizo-Villar was supported by a "Ayuda para contratos predoctorales para la Formación de Profesorado Universitario (FPU)" grant from the Ministerio de Educación, Cultura y Deporte (FPU16/02938). The work of Daniel Álvarez was supported by a "Ramón y Cajal" grant (RYC2019-028566-I) from the "Ministerio de Ciencia e Innovación-Agencia Estatal de Investigación" co-funded by the European Social Fund.

References

Acharya, U. R., et al. (2006). Heart rate variability: A review. *Medical and Biological Engineering and Computing, 44*(12), 1031–1051. https://doi.org/10.1007/s11517-006-0119-0

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* [Preprint]. https://doi.org/10.1109/ACCESS.2018.2870052

Álvarez, D., et al. (2010). Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. *IEEE Transactions on Biomedical Engineering, 57*(12), 2816–2824. https://doi.org/10.1109/TBME.2010.2056924

Álvarez, D., et al. (2013). Assessment of feature selection and classification approaches to enhance information from overnight oximetry in the context of sleep apnea diagnosis, *23*(5). https://doi.org/10.1142/S0129065713500202

Álvarez, D., et al. (2020). A machine learning-based test for adult sleep apnoea screening at home using oximetry and airflow. *Scientific Reports* [Preprint]. https://doi.org/10.1038/s41598-020-62223-4

Azarbarzin, A., et al. (2019). The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: The Osteoporotic Fractures in Men Study and the sleep heart health study. *European Heart Journal, 40*(14), 1149–1157. https://doi.org/10.1093/EURHEARTJ/EHY624

Bahammam, A. (2004). Comparison of nasal prong pressure and thermistor measurements for detecting respiratory events during sleep. *Respiration, 71*(4), 385–390. https://doi.org/10.1159/000079644

Barroso-García, V., et al. (2017). Irregularity and variability analysis of airflow recordings to facilitate the diagnosis of paediatric sleep apnoea-hypopnoea syndrome. *Entropy, 19*(9), 447. https://doi.org/10.3390/E19090447

Barroso-García, V. et al. (2021). Wavelet analysis of overnight airflow to detect obstructive sleep apnea in children. *Sensors*, *21*(4). https://doi.org/10.3390/s21041491

Berry, R. B., et al. (2012). Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events. *Journal of Clinical Sleep Medicine, 8*(05), 597–619.

Berry, R. B., et al. (2017). AASM scoring manual updates for 2017 (version 2.4). *Journal of Clinical Sleep Medicine, 13*(05), 665–666.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* [Preprint]. https://doi.org/10.1214/07-STS242

Chen, C. C., & Barnhart, H. X. (2008). Comparison of ICC and CCC for assessing agreement for data without and with replications. *Computational Statistics and Data Analysis, 53*(2), 554–564. https://doi.org/10.1016/j.csda.2008.09.026

Chen, L., Zhang, X., & Song, C. (2015). An automatic screening approach for obstructive sleep apnea diagnosis based on single-lead electrocardiogram. *IEEE*

Transactions on Automation Science and Engineering, *12*(1). https://doi.org/10.1109/TASE.2014.2345667

Deviaene, M., et al. (2019). Automatic screening of sleep apnea patients based on the SpO 2 signal. *IEEE Journal of Biomedical and Health Informatics* [Preprint]. https://doi.org/10.1109/JBHI.2018.2817368

El-Solh, A. A., et al. (1999). Validity of neural network in sleep apnea. *Sleep, 22*(1). https://doi.org/10.1093/sleep/22.1.105

Flemons, W. W., et al. (1999). Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research. *Sleep*. https://doi.org/10.1093/sleep/22.5.667

Garde, A., et al. (2014). Development of a screening tool for sleep disordered breathing in children using the phone oximeterTM," *PLoS One, 9*(11). https://doi.org/10.1371/journal.pone.0112959

Ghegan, M. D., et al. (2006). Laboratory versus portable sleep studies: A meta-analysis. *The Laryngoscope, 116*(6), 859–864. https://doi.org/10.1097/01.mlg.0000214866.32050.2e

Giavarina, D. (2015). Understanding bland altman analysis. *Biochemia Medica*, *25*(2). https://doi.org/10.11613/BM.2015.015

Gil, E., et al. (2010). "Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions," *Physiological Measurement*, *31*(9). https://doi.org/10.1088/0967-3334/31/9/015

Gutiérrez-Tobal, G. C., et al. (2013). Pattern recognition in airflow recordings to assist in the sleep apnoea-hypopnoea syndrome diagnosis. *Medical and Biological Engineering and Computing, 51*(12), 1367–1380. https://doi.org/10.1007/s11517-013-1109-7

Gutierrez-Tobal, G. C., et al. (2016). Utility of AdaBoost to detect sleep apnea-hypopnea syndrome from single-channel airflow. *IEEE Transactions on Biomedical Engineering, 63*(3), 636–646. https://doi.org/10.1109/TBME.2015.2467188

Gutiérrez-Tobal, G. C., et al. (2019). Evaluation of machine-learning approaches to estimate sleep apnea severity from at-home oximetry recordings. *IEEE Journal of Biomedical and Health Informatics, 23*(2), 882–892. https://doi.org/10.1109/JBHI.2018.2823384

Gutiérrez-Tobal, G. C., Álvarez, D., et al. (2021a). Ensemble-learning regression to estimate sleep apnea severity using at-home oximetry in adults. *Applied Soft Computing*, *111*. https://doi.org/10.1016/j.asoc.2021.107827

Gutiérrez-Tobal, G. C., Gomez-Pilar, J., et al. (2021b). Pediatric sleep apnea: The overnight electroencephalogram as a phenotypic biomarker. *Frontiers in Neuroscience*, 1448. https://doi.org/10.3389/FNINS.2021.644697

Gutiérrez-Tobal, G. C., et al. (2021c). Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis. *Pediatric Pulmonology* [Preprint].

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182. https://doi.org/10.1162/153244303322753616

Hornero, R., et al. (2017). Nocturnal oximetry-based evaluation of habitually snoring children. *American Journal of Respiratory and Critical Care Medicine, 196*(12), 1591–1598. https://doi.org/10.1164/rccm.201705-0930OC

Hosmer, D., & Lemeshow, S. (1989). *Applied logistic regression*. Available at: http://ecsocman.hse.ru/text/19164818/. Accessed: 2 Nov 2021.

Ian, G., Yoshua, B., & Aaron, C. (2016). *Deep learning* - Ian Goodfellow, Yoshua Bengio, Aaron Courville - Google Books. *MIT Press* [Preprint].

Iber, C., et al. (2007). The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specification. *Journal of Clinical Sleep Medicine* [Preprint].

Jiménez-García, J., et al. (2020). Assessment of airflow and oximetry signals to detect pediatric sleep apnea-hypopnea syndrome using AdaBoost. *Entropy* [Preprint]. https://doi.org/10.3390/E22060670

Karhu, T., et al. (2021). Longer and deeper desaturations are associated with the worsening of mild sleep apnea: The sleep heart health study," *Frontiers in Neuroscience* [Preprint]. https://doi.org/10.3389/fnins.2021.657126

Khandoker, A. H., Palaniswami, M., & Karmakar, C. K. (2009). Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings. *IEEE Transactions on Information Technology in Biomedicine*, *13*(1). https://doi.org/10.1109/TITB.2008.2004495

Korkalainen, H., et al. (2019). Mortality-risk-based apnea–hypopnea index thresholds for diagnostics of obstructive sleep apnea. *Journal of Sleep Research* [Preprint]. https://doi.org/10.1111/jsr.12855

Kuncheva, L. I. (2014). *Combining pattern classifiers: Methods and algorithms*. John Wiley & Sons.

Lázaro, J., et al. (2014). Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children. *IEEE Journal of Biomedical and Health Informatics* [Preprint]. https://doi.org/10.1109/JBHI.2013.2267096

Lin, Y. Y., et al. (2017). Sleep apnea detection based on thoracic and abdominal movement signals of wearable piezoelectric bands. *IEEE Journal of Biomedical and Health Informatics*, *21*(6). https://doi.org/10.1109/JBHI.2016.2636778

Marcos, J. V., et al. (2008) 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). https://doi.org/10.1016/j.cmpb.2008.05.006

Marcos, J. V., et al. (2009). Assessment of four statistical pattern recognition techniques to assist in obstructive sleep apnoea diagnosis from nocturnal oximetry. *Medical Engineering & Physics, 31*(8), 971–978. https://doi.org/10.1016/J.MEDENGPHY.2009.05.010

Marcos, J. V., et al. (2012). Automated prediction of the apnea-hypopnea index from nocturnal oximetry recordings. *IEEE Transactions on Biomedical Engineering, 59*(1), 141–149. https://doi.org/10.1109/TBME.2011.2167971

Marcus, C. L., et al. (2013). A randomized trial of adenotonsillectomy for childhood sleep apnea. *New England Journal of Medicine* [Preprint]. https://doi.org/10.1056/nejmoa1215881

Martín-Montero, A., et al. (2021). Heart rate variability spectrum characteristics in children with sleep apnea. *Pediatric Research, 89*(7), 1771. https://doi.org/10.1038/S41390-020-01138-2

Mendonça, F., et al. (2019). A review of obstructive sleep apnea detection approaches. *IEEE Journal of Biomedical and Health Informatics, 23*(2), 825–837. https://doi.org/10.1109/JBHI.2018.2823265

Morillo, D. S., & Gross, N. (2013). Probabilistic neural network approach for the detection of SAHS from overnight pulse oximetry. *Medical and Biological Engineering and Computing, 51*(3), 305–315. https://doi.org/10.1007/s11517-012-0995-4

Newman, A. B., et al. (2001). Relation of sleep-disordered breathing to cardiovascular disease risk factors: The sleep heart health study. *American Journal of Epidemiology*, *154*(1). https://doi.org/10.1093/aje/154.1.50

Penzel, T., et al. (2002). Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. *Medical & Biological Engineering & Computing, 40*(4), 402–407. https://doi.org/10.1007/BF02345072

Penzel, T., et al. (2003). Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *IEEE Transactions on Biomedical Engineering, 50*(10), 1143–1151. https://doi.org/10.1109/TBME.2003.817636

Penzel, T., Schöbel, C., & Fietze, I. (2015). Revise respiratory event criteria or revise severity thresholds for sleep apnea definition? *Journal of Clinical Sleep Medicine* [Preprint]. https://doi.org/10.5664/jcsm.5262

Quan, S. F., et al. (1997). The sleep heart health study: Design, rationale, and methods. *Sleep, 20*(12), 1077–1085.

Riedl, M., et al. (2014). Cardio-respiratory coordination increases during sleep apnea. *PLoS One, 9*(4). https://doi.org/10.1371/journal.pone.0093866

Rolón, R. E., et al. (2017). Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea–hypopnea detection. *Biomedical Signal Processing and Control* [Preprint]. https://doi.org/10.1016/j.bspc.2016.12.013

Rolon, R. E., et al. (2020). Automatic scoring of apnea and hypopnea events using blood oxygen saturation signals. *Biomedical Signal Processing and Control* [Preprint]. https://doi.org/10.1016/j.bspc.2020.102062

Skotko, B. G., et al. (2017). A predictive model for obstructive sleep apnea and Down syndrome. *American Journal of Medical Genetics, Part A* [Preprint]. https://doi.org/10.1002/ajmg.a.38137

Solà-Soler, J., et al. (2012). Multiclass classification of subjects with sleep apnoea-hypopnoea syndrome through snoring analysis. *Medical Engineering and Physics, 34*(9). https://doi.org/10.1016/j.medengphy.2011.12.008

Tan, H. L., et al. (2014). Overnight polysomnography versus respiratory polygraphy in the diagnosis of pediatric obstructive sleep apnea. *Sleep* [Preprint]. https://doi.org/10.5665/sleep.3392

Tan, H. L. et al. (2017). When and why to treat the child who snores? *Pediatric Pulmonology*, *52*(3), 399–412. https://doi.org/10.1002/ppul.23658

Uddin, M., et al. (2018). Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: A systematic review. *iopscience.iop.org* [Preprint]. https://doi.org/10.1088/1361-6579/aaafb8

Vaquerizo-Villar, F., et al. (2021). A convolutional neural network architecture to enhance oximetry ability to diagnose pediatric obstructive sleep apnea. *IEEE Journal of Biomedical and Health Informatics* [Preprint]. https://doi.org/10.1109/JBHI.2020.3048901

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, D., et al. (2017). A simple diagnostic scale based on the analysis and screening of clinical parameters in paediatric obstructive sleep apnoea hypopnea syndrome. *Journal of Laryngology and Otology* [Preprint]. https://doi.org/10.1017/S0022215117000238

Xu, Z., et al. (2019). Cloud algorithm-driven oximetry-based diagnosis of obstructive sleep apnoea in symptomatic habitually snoring children. *The European Respiratory Journal* [Preprint]. https://doi.org/10.1183/13993003.01788-2018

Zweig, M. H., & Campbell, G. (1993) Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*. https://doi.org/10.1093/clinchem/39.4.561