Universidad de Valladolid

Doctoral Program of
Information and Telecommunication Technologies

Doctoral Thesis

Computer aided diagnosis of pediatric sleep apnea through the analysis of airflow and oximetry signals: from ensemble learning to explainable deep learning algorithms

Thesis presented by **Jorge Jiménez García**
to apply for the *Ph.D. degree*
from the *University of Valladolid*

Directed by:
Dr. María García Gadañón and Dr. Gonzalo C. Gutiérrez Tobal

2024
Valladolid, Spain

A mis padres
y a mis hermanos

Universidad de Valladolid

School of Telecommunications Engineering
Department of Signal and Communications Theory and Telematic Engineering

Research Stay for the International Mention

City:	Leuven (Belgium)
Faculty:	KU Leuven
Department:	Department of Electrical Engineering (ESAT)
Research group:	Biomedical Data Processing Research Team (BIOMED) at STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics
Dates:	01/09/2023–01/12/2023
Duration:	92 days (3 months)
Supervisor:	Prof. Dr. Maarten De Vos

Agradecimientos

En el final de esta etapa quiero expresar mi más sincero agradecimiento a todas aquellas personas que me han acompañado durante todos estos años. En primer lugar, quiero agradecer a mis directores de tesis, Dra. María García Gadañón y Dr. Gonzalo César Gutiérrez Tobal, haberme guiado por el buen camino y haber conseguido que ahora mismo esté escribiendo esta tesis doctoral. Gracias María, por la confianza que has depositado en mí desde aquel día en el que hablamos sobre la posibilidad de hacer mi trabajo de fin de grado, y por todo lo que he podido aprender durante todos estos años. Gracias Gonzalo, por todos los consejos que me has dado y por tener tanta paciencia conmigo. Espero que haya merecido la pena. También quiero dar las gracias al Prof. Roberto Hornero por ofrecerme la posibilidad de unirme al Grupo de Ingeniería Biomédica (GIB) de la Universidad de Valladolid. Gracias también a Jesús Poza, Carlos Gómez, Daniel Álvarez, Javier Gómez y Félix del Campo por compartir vuestras enseñanzas y experiencias. I would also like to express my gratitude to Drs. David Gozal and Leila Kheirandish-Gozal for their aid and collaboration during these years.

Quiero dar las gracias también a los compañeros y compañeras del GIB, con los que he compartido montones de experiencias, y también han hecho que este camino haya sido mucho más llevadero. Muchas gracias por vuestro compañerismo, por el buen ambiente de trabajo, y en definitiva por hacer fácil lo difícil entre todos. Gracias a los hoy doctores Fernando Vaquerizo, Verónica Barroso, Adrián Martín, Roberto Romero, Víctor Martínez, Eduardo Santamaría, Saúl Ruiz y Pablo Núñez. Y también muchas gracias a los que lo van a ser: Clara, Enrique, Aarón, Víctor R., Víctor G., Marcos, Diego, Sergio, Selene, María H., Rubén, Ana, Beatriz y Marina.

También me gustaría agradecer a mis padres su apoyo y sacrificio. Muchas gracias por estar ahí, por pensar y actuar siempre en nuestro beneficio. Y también quiero extender este agradecimiento a Álvaro y a toda mi familia. Gracias por

estar a mi lado.

I would also like to thank Professor Maarten de Vos for welcoming me into the Biomedical Data Processing research group (BIOMED) at the KU Leuven university. I had a fantastic time in Leuven (Belgium) with all the BIOMED researchers, and I have learned a lot from them. Thanks for showing me your group and the way you research, and for including me in your team. Thanks to Joran and Guido, with whom I had the pleasure to collaborate during the stay and I hope to continue working together in the future.

Finalmente, muchas gracias a Javier, a Raúl y a Jesús por todos los buenos ratos que hemos pasado y los que están por venir. A pesar de la distancia, siempre estáis conmigo cuando os necesito.

Gracias por confiar en mí

Abstract

Obstructive sleep apnea (OSA) is a sleep disorder in which intermittent obstruction or narrowing of the upper airway causes recurrent pauses and cessations of normal respiration (named apneas and hypopneas, respectively) during sleep. The most common symptoms of OSA in chidren are snore, labored respiration or breathing pauses, and daytime hypersomnolence. However, these symptoms can be subtle and not easily detected. Pediatric OSA affects 1% – 5% of children, with several negative consequences that range from cardiometabolical comorbidities such as hypertension and dyslipidemia to neurobehabioral disorders such as neurocognitive and attention deficits, and also hyperactivity. Early diagnosis of children at risk of OSA is crucial to access surgical or pharmacological treatment and diminish the chances of developing serious comorbidities.

The gold standard in the diagnosis of childhood OSA is the nocturnal, in-lab polysomnography (PSG), a sleep study that involves the recording of cardiorespiratory, neuronal, muscular, position, and movement signals while the patient is sleeping in a sleep laboratory. PSG signals are inspected to locate and quantify apnea or hypopnea events. The American Academy of Sleep Medicine (AASM) define apneas as a reduction $\geq 90\%$ in the airflow (AF) signal during at least two breathing periods. Likewise, hypopneas are defined as an AF reduction $\geq 30\%$ for at least two breathing cycles associated with either a drop in the oxygen saturation (SpO_2) signal $\geq 3\%$ (desaturation) or an electroencephalographic arousal. Pediatric OSA is diagnosed by computing the rate of apnea/hypopnea events per hour (e/h) of sleep (apnea-hypopnea index, AHI). OSA severity is defined according to AHI: no OSA (AHI < 1 e/h), mild OSA ($1 \leq$ AHI < 5 e/h), moderate OSA ($5 \leq$ AHI < 10 e/h), and severe OSA (AHI ≥ 10 e/h). Notwithstanding the preference of PSG to diagnose OSA in children, it has low availability due to scarce sleep units, high complexity and associated costs. These reasons delay the diagnosis of children affected by OSA, making it an underdiagnosed disease. Alternatives to

PSG comprise the analysis of less signals, which can also be recorded outside the sleep laboratories. This way, simplified tests involving the analysis of AF and SpO_2 are a suitable alternative since these signals summarize the information required to detect apneas and hypopneas.

The analysis of AF and SpO_2 signals can be further simplified by means of automatic signal processing algorithms. Moreover, these methods can comprise a pattern recognition stage that automatically detects signs of pathology and provide a simplified diagnosis of pediatric OSA. In this doctoral thesis, different advanced machine learning (ML) algorithms such as ensemble learning and more recent deep learning (DL) architectures combined with Explainable Artificial Intelligence (XAI) methods are proposed to facilitate the diagnosis of pediatric OSA through the automatic analysis of AF and SpO_2 signals. The research conducted throughout this doctoral thesis was conducted by assuming the hypothesis that automatic analysis of overnight AF and SpO_2 signals using advanced ML techniques such as ensemble learning, DL and XAI can help to simplify the diagnosis of childhood OSA. Thus, the main objective of this doctoral thesis was to study, develop, and validate advanced ML methods such as ensemble learning or DL together with new XAI techniques in the context of automatic analysis of AF and SpO_2 signals, so that these methods can be used to help diagnose pediatric OSA.

In order to conduct this research, two databases of overnight AF and SpO_2 signals from a total of 2,612 sleep studies have been used. The first database was provided by the University of Chicago (UofC) School of Medicine, and comprised 974 pediatric subjects with suspicion of OSA. The Childhood Adenotonsillectomy Trial (CHAT) public and multicentric database contained a total of 1,638 sleep studies performed to children with OSA symptoms.

The methodology deployed to achieve the main goal of this doctoral thesis was split into two main branches. The first branch was a feature-engineering methodology that encompassed feature extraction, selection and classification stages to estimate the presence of OSA and its severity from the most relevant and non-redundant information extracted from AF and SpO_2. Temporal, spectral, and nonlinear parameters were computed from AF and SpO_2, and the 3% oxygen desaturation index (ODI 3%) was also included in the analyses. Next, subsets of relevant and non-redundant features were obtained by applying the Fast Correlation-Based Filter (FCBF) algorithm on different combinations of AF- and SpO_2-derived information. The classification stage comprised the implementation of AdaBoost ensemble learning classifiers to estimate the OSA severity level from the subsets of relevant and complementary information of AF and SpO_2. The second branch

included the deployment of different DL architectures aimed at quantifying OSA severity by means of AHI estimation from the total number of detected apneic events. Moreover, the explainability of the developed DL models was addressed with XAI algorithms to obtain explanations about the DL models' predictions. The first DL architecture was a convolutional neural network (CNN) trained to process and analyze raw AF and SpO_2, while the second relied on a combination of the aforementioned CNN with a recurrent neural network (RNN), namely CNN+RNN, by means of a transfer learning approach. This architecture was further analyzed by means of the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm, a XAI method that derives explanatory heatmaps associated to the model's predictions.

The results obtained from the feature engineering approach revealed that the combination of AF- and SpO_2-derived features with ODI 3% by means of a multiclass AdaBoost classifier obtained the highest agreement with the actual OSA severity in terms of 4-class accuracy (Acc_4) and Cohen's kappa (k) (Acc_4=57.95%, k=0.3984). Nevertheless, the combination of AF with ODI 3% reached the maximum performance to diagnose pediatric OSA in terms of accuracy (Acc), sensitivity (Se), and specificity (Sp) in the AHI cutoffs of 1 e/h (Acc=81.28%, Se=92.06%, Sp=36.00%), 5 e/h (Acc=82.05%, Se=76.03%, Sp=85.66%), and 10 e/h (Acc=90.26%, Se=62.65%, Sp=97.72%). On the other hand, the DL architectures outperformed the previous feature-engineering approaches in terms of diagnostic performance. The Intraclass Correlation Coefficient (ICC) revealed a very high agreement of AHI estimation in CHAT (ICC=0.9546 for CNN; ICC=0.9465 for CNN+RNN) and UofC (ICC=0.8821 for CNN; ICC=0.9004 for CNN+RNN) test sets, which was also confirmed in terms of Acc_4 and k. The agreement was higher in CHAT (Acc_4=72.55%, k=0.6011 for CNN; Acc_4=74.51%, k=0.6231 for CNN+RNN), while the results in the UofC surpassed those reached with the AdaBoost models (Acc_4=61.79%, k=0.4469 for CNN; Acc_4=62.31%, k=0.4495 for CNN+RNN). Comparing both DL models, CNN+RNN outperformed CNN. A superior ability to diagnose OSA was obtained in all AHI cutoffs, with very high diagnostic performance in 1 e/h (Acc=87.25%, Se=87.03%, Sp=88.06%), 5 e/h (Acc=93.46%, Se=80.22%, Sp=99.07%), and 10 e/h (Acc=93.46%, Se=71.43%, Sp=96.97%) in the CHAT test set, whereas these metrics were also high in the UofC test set in 1 e/h (Acc=84.10%, Se=96.83%, Sp=30.67%), 5 e/h (Acc=84.62%, Se=82.88%, Sp=85.66%), and 10 e/h (Acc=90.51%, Se=78.31%, Sp=93.81%). These results indicate that the CNN+RNN model is the most accurate model among all the approaches covered in this doctoral thesis. The Grad-

CAM-derived explanatory heatmaps obtained from this model revealed that it focused in sudden AF amplitude changes and desaturations to detect and quantify OSA-related events. Nevertheless, these heatmaps also revealed that the model usually misses some hypopneas associated to arousals and is sensitive to abrupt variations in both signals caused by artifacts.

The characterization of AF and SpO_2 revealed that both signals exhibited relevant and complementary features according to the results obtained using the FCBF feature selection algorithm. SpO_2-derived ODI 3% was the most relevant and dominant feature, whereas the central tendency measure computed from AF was also relevant and non-redundant with ODI 3%. This complementary information from both signals reached the highest diagnostic performance in comparison with approaches that only included the information from one of these signals. Nevertheless, ODI 3% contributed the most to enhance OSA detection by means of AdaBoost. Overall, the multi-class AdaBoost models reached remarkable diagnostic performance compared to other approaches that also combined AF with ODI 3%. These results suggest that AF and SpO_2 can be complementary and useful together to detect pediatric OSA. However, the superior diagnostic ability of DL-based approaches was clearly demonstrated. The CNN model trained with both signals outperformed a very similar previous approach based solely on SpO_2 in 1 and 5 e/h, indicating that the contribution of AF was remarkable to enhance the utility of this CNN model. The architecture obtained by extending the aforementioned dual-channel CNN model to a CNN+RNN naturally surpassed all the previous approaches focused on detecting pediatric OSA. This highlights the usefulness of a DL architecture that combines different techniques to automatically learn the particularities of AF and SpO_2 signals. In addition, the use of Grad-CAM enabled the discovery of relevant OSA-related patterns of the input signals that are automatically detected by the DL algoritms and can aid users to reinforce their trust in the DL-based models. In view of these results, the methods proposed in this doctoral thesis could be used to develop a screening test of pediatric OSA that would alleviate the waiting lists of pediatric sleep laboratories.

In summary, the results achieved by the methods proposed in this research allow us to conclude that the automatic analysis of AF and SpO_2 based on ensemble and DL methods combined with XAI have demonstrated a remarkable diagnostic usefulness, and can be used to deploy alternative, simple, reliable and trustworthy screening methods to serve as an aid in the diagnosis of OSA in children.

Acronyms

AF	airflow
AASM	American Academy of Sleep Medicine
AC	alternating current
Acc	accuracy
Acc_4	four-class accuracy
AdaBoost	Adaptive Boosting
AHI	apnea-hypopnea index
AI	artificial intelligence
Bi-GRU	bidirectional gated recurrent unit
BMI	body mass index
CHAT	Childhood Adenotonsillectomy Trial
CNN	convolutional neural network
CSA	central sleep apnea
CTM	central tendency measure
DC	direct current
DFA	detrended fluctuation analysis
DFT	discrete Fourier transform
DL	deep learning
DNN	deep neural network
DT	decision trees
e/h	events per hour
ECG	electrocardiography
EEG	electroencephalography
EMG	electromiography
EOG	electroculography
f_s	sampling frequency

FCBF	Fast Correlation-Based Filter
FN	false negative
FP	false positive
FreqM	median frequency
Grad-CAM	Gradient-weighted Class Activation Mapping
Hb	hemoglobin
HRV	heart rate variability
HSAT	home sleep apnea test
ICC	intraclass correlation coefficient
JCR	Journal Citation Reports
JIF	Journal Impact Factor
LASSO	least absolute shrinkage and selection operator
LDA	linear discriminant analysis
LR	logistic regression
LR−	negative likelihood ratio
LR+	positive likelihood ratio
LZC	Lempel-Ziv complexity
M1F-M4F	statistical moments in the frequency domain
M1T-M4T	statistical moments in the time domain
MaxF	maximum in the frequency domain
MedF	median in the frequency domain
MedT	median in the time domain
MinF	minimum in the frequency domain
ML	machine learning
NN	neural network
NPV	negative predictive value
NSRR	National Sleep Research Resource
O_2Hb	oxyhemoglobin
ODI	oxygen desaturation index
OSA	obstrucive sleep apnea
PPG	photoplethysmography
PPV	positive predictive value
PR	pulse rate
PRV	pulse rate variability
PSD	power spectral density
PSG	polysomnography
PVDF	polyvinylidene fluoride

ReLU	rectified linear unit
RNN	recurrent neural network
RP	respiratory polygraphy
RRV	respiratory rate variability
SA	sleep apnea
SampEn	sample entropy
SaO_2	blood oxygen saturation
SDB	sleep disordered breathing
Se	sensitivity
SHAP	Shapley additive explanations
Sp	specificity
SpecEn	spectral entropy
SpO_2	oximetry / peripheral blood oxygen saturation
SU	symmetrical uncertainty
SVM	support vector machine
TD	time distributed
TN	true negative
TP	true positive
TST	total sleep time
WOS	Web of Science™
XAI	explainable artificial intelligence

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This doctoral thesis focused on the automatic analysis of respiratory airflow (AF) and peripheral blood oxygen saturation (oximetry, SpO_2) signals to help in the diagnosis of pediatric obstructive sleep apnea (OSA). With the objective of automating the analysis of these two overnight signals, several algorithms were specifically designed to process and characterize the signals, extract useful information related to the disease, and finally determine the presence and severity of OSA. The approaches covered in this doctoral thesis relied on machine learning (ML) techniques such as ensemble learning, deep learning (DL), and explainable artificial intelligence (XAI). ML and DL models were developed to estimate the presence and severity of OSA, whereas XAI methods were used to analyze the patterns present in the signals that DL algorithms link to the presence of OSA. The results of these studies were published in three journal articles, all of them indexed in the Web of Science™(WOS) Journal Citation Reports (JCR). Thus, this doctoral thesis is presented as a compendium of publications.

The articles included in the compendium of publications propose novel techniques applied to AF and SpO_2 to enhance the diagnosis of OSA in children. The thematic consistency is introduced in Section 1.1. The research conducted in this doctoral thesis is framed into the fields of biomedical engineering, biomedical signal processing, and ML, which are exposed in Section 1.2 devoted to the general context of this thesis. The medical background of the doctoral thesis is centered in pediatric OSA (Section 1.3), a prevalent sleep disorder with important negative consequences concerning the development of children. The diagnosis of pediatric OSA is described in Section 1.4, along with their limitations, disadvantages, and possible alternatives. In this view, the state-of-the-art alternatives that encompass

the use of data-driven models such as ML, DL, and XAI to aid in the diagnosis of childhood OSA are rewiewed in Section 1.5.

1.1 Compendium of publications: thematic consistency

OSA is a type of sleep disordered breathing (SDB) syndrome caused by recurrent, intermittent obstructions of the airway, either total (apneas) or partial (hypopneas), during sleep (Dehlink and Tan, 2016). The prevalence of OSA in children ranges between 1% and 5% according to epidemiological studies (Kaditis et al., 2016b; Marcus et al., 2012), but the actual rate of children who suffer OSA may be higher and represents a child health concern in many countries. Childhood OSA is associated to a variety of cardiometabolical and neurobehavioral comorbidities, such as hypertension, dyslipidemia, and neurocognitive deficits (Marcus et al., 2012; Tauman and Gozal, 2011). Timely diagnosis of children with suspicion of OSA is crucial to avoid its negative consequences, but remains challenging due to the complications of the most accepted diagnostic procedures. The gold standard for OSA diagnosis is an in-lab overnight polysomnography (PSG), a complex and expensive test, with limited availability, and uncomfortable for children (Dehlink and Tan, 2016; Marcus et al., 2012). The disadvantages of PSG hinder the access to a timely diagnosis of OSA, resulting in long waiting lists. Therefore, the simplification of OSA diagnosis is required to enhance the access to a convenient examination and possible treatment of affected children (Brockmann et al., 2018).

The research conducted in this doctoral thesis was aimed at providing tools to automatically analyze overnight AF and SpO_2, two signals directly involved in PSG, and enabling a simplified diagnosis of pediatric OSA directly from them. The three papers included in the compendium of publications assessed the diagnostic ability of overnight AF and SpO_2 to predict OSA, each of them proposing different ML, DL and XAI algorithms. All scientific contributions proposed a novel methodology to automatically process overnight AF and SpO_2, extract useful information from these signals, and finally derive a prediction of the severity of OSA. The proposed ML methods covered from advanced feature-engineering algorithms, such as ensemble learning, to the most recent DL architectures combined with XAI. Ensemble learning methods encompassed exhaustive feature extraction, selection, and classification stages, whereas DL-based approaches unified the aforementioned stages, enabling an automated algorithm to predict OSA directly from the raw sig-

nals. Finally, the last paper of the compendium addressed the interpretability of DL models by applying a XAI algorithm that uncovers the black-box nature of current complex data-driven models. Figure 1.1 illustrates the thematic consistency of the articles that constitute this doctoral thesis, which are introduced in the next paragraphs of this section.

The three articles in the compendium of publications included the assessment of the methods in populations of children with symptoms suggestive of OSA that underwent PSG testing. To do so, AF and SpO_2 signals were processed and an-

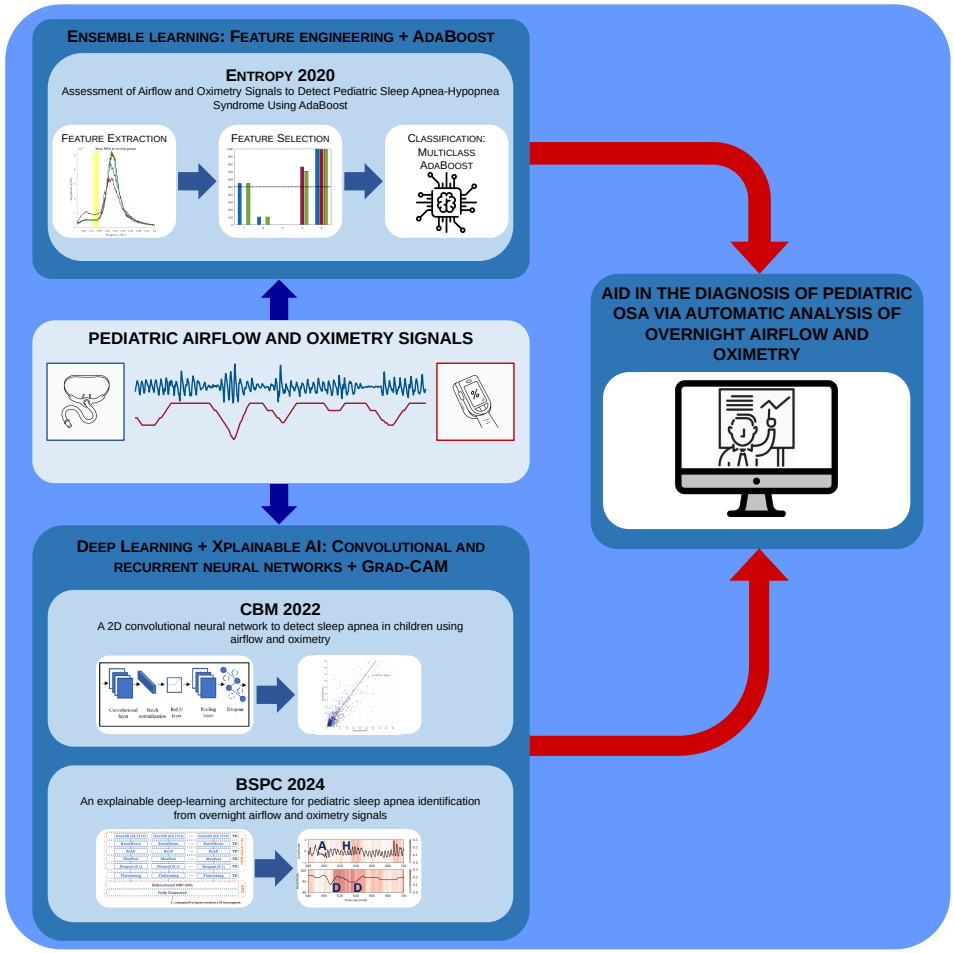


Figure 1.1: Thematic consistency between the publications of this doctoral thesis. AI: Artificial Intelligence; BSPC: Biomedical Signal Processing and Control; CBM: Computers in Biology and Medicine; Grad-CAM: Gradient-weighted Class Activation Mapping; OSA: Obstructive Sleep Apnea.

alyzed using different methodologies. The first article of the compendium was focused on assessing the diagnostic ability of AF and SpO_2 signals to detect pediatric OSA either alone and combined (Jiménez-García et al., 2020). Both signals were exhaustively characterized by extracting time-domain statistics, spectral and nonlinear features, as well as the 3% oxygen desaturation index (ODI 3%) computed from SpO_2. The OSA severity was obtained using the multiclass AdaBoost ensemble learning algorithm fed with combinations of relevant and non-redundant features. The feature sets that included information from both signals (especially ODI 3%) reached the highest accuracy, suggesting the complementarity of AF and SpO_2 to automatically diagnose pediatric OSA (Jiménez-García et al., 2020). The second and third articles of the compendium proposed DL models to automatically diagnose OSA from overnight AF and SpO_2. A two-dimensional convolutional neural network (CNN), as well as a combination of CNN with a recurrent neural network (RNN) were developed and tested in these articles (Jiménez-García et al., 2022, 2024). The target of CNN and CNN + RNN architectures was the computation of the apnea-hypopnea index (AHI), the main marker of pediatric OSA, from the amount of detected apneic events in these two signals. Additionally, the Gradient-weighted Class Activation Mapping (Grad-CAM) XAI algorithm was used to explain the behavior of the CNN + RNN algorithm and identify relevant OSA-related patterns found by the DL model from the raw signals (Jiménez-García et al., 2024). In summary, the three approaches included in this research covered different state-of-the-art algorithms to aid in the diagnosis of pediatric OSA and also described the transition from a classical ML perspective to a novel explainable DL framework.

Titles, authors, journals, and abstracts of the articles included in the compendium of this doctoral thesis are shown in the next three pages. The journals in which the articles were published also include their respective Journal Impact Factor (JIF). Due to this doctoral thesis is presented as a compendium of publications, the complete articles have been included in Sections 7.1, 7.2, and 7.3 for a suitable understanding of this dissertation.

Assessment of Airflow and Oximetry Signals to Detect Pediatric Sleep Apnea-Hypopnea Syndrome Using AdaBoost (Jiménez-García et al., 2020).

Jorge Jiménez-García, Gonzalo C. Gutiérrez-Tobal, María García, Leila Kheirandish-Gozal, Adrián Martín-Montero, Daniel Álvarez, Félix del Campo, David Gozal, and Roberto Hornero. *Entropy*, vol. 22 (6), pp. 670, 2020. JIF in 2020: 2.524, Q2 in "Physics, Multidisciplinary" (JCR-WOS).

Abstract: The reference standard to diagnose pediatric Obstructive Sleep Apnea (OSA) syndrome is an overnight polysomnographic evaluation. When polysomnography is either unavailable or has limited availability, OSA screening may comprise the automatic analysis of a minimum number of signals. The primary objective of this study was to evaluate the complementarity of airflow (AF) and oximetry (SpO_2) signals to automatically detect pediatric OSA. Additionally, a secondary goal was to assess the utility of a multiclass AdaBoost classifier to predict OSA severity in children. We extracted the same features from AF and SpO_2 signals from 974 pediatric subjects. We also obtained the 3% Oxygen Desaturation Index (ODI) as a common clinically used variable. Then, feature selection was conducted using the Fast Correlation-Based Filter method and AdaBoost classifiers were evaluated. Models combining ODI 3% and AF features outperformed the diagnostic performance of each signal alone, reaching 0.39 Cohens's kappa in the four-class classification task. OSA vs. No OSA accuracies reached 81.28%, 82.05% and 90.26% in the apnea–hypopnea index cutoffs 1, 5 and 10 events/h, respectively. The most relevant information from SpO_2 was redundant with ODI 3%, and AF was complementary to them. Thus, the joint analysis of AF and SpO_2 enhanced the diagnostic performance of each signal alone using AdaBoost, thereby enabling a potential screening alternative for OSA in children.

A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry (Jiménez-García et al., 2022).

Jorge Jiménez-García, María García, Gonzalo C. Gutiérrez-Tobal, Leila Kheirandish-Gozal, Fernando Vaquerizo-Villar, Daniel Álvarez, Félix del Campo, David Gozal, and Roberto Hornero. *Computers in Biology and Medicine*, vol. 147, pp. 105784, 2022. JIF in 2022: 7.7, Q1(D1) in "MATHEMATICAL & COMPUTATIONAL BIOLOGY" (JCR-WOS).

Abstract: The gold standard approach to diagnose obstructive sleep apnea (OSA) in children is overnight in-lab polysomnography (PSG), which is labor-intensive for clinicians and onerous to healthcare systems and families. Simplification of PSG should enhance availability and comfort, and reduce complexity and waitlists. Airflow (AF) and oximetry (SpO_2) signals summarize most of the information needed to detect apneas and hypopneas, but automatic analysis of these signals using deep-learning algorithms has not been extensively investigated in the pediatric context. The aim of this study was to evaluate a convolutional neural network (CNN) architecture based on these two signals to estimate the severity of pediatric OSA. PSG-derived AF and SpO_2 signals from the Childhood Adenotonsillectomy Trial (CHAT) database (1638 recordings), as well as from a clinical database (974 recordings), were analyzed. A 2D CNN fed with AF and SpO_2 signals was implemented to estimate the number of apneic events, and the total apnea-hypopnea index (AHI) was estimated. A training-validation-test strategy was used to train the CNN, adjust the hyperparameters, and assess the diagnostic ability of the algorithm, respectively. Classification into four OSA severity levels (no OSA, mild, moderate, or severe) reached 4-class accuracy and Cohen's Kappa of 72.55% and 0.6011 in the CHAT test set, and 61.79% and 0.4469 in the clinical dataset, respectively. Binary classification accuracy using AHI cutoffs 1, 5 and 10 events/h ranged between 84.64% and 94.44% in CHAT, and 84.10%–90.26% in the clinical database. The proposed CNN-based architecture achieved high diagnostic ability in two independent databases, outperforming previous approaches that employed SpO2 signals alone, or other classical feature-engineering approaches. Therefore, analysis of AF and SpO_2 signals using deep learning can be useful to deploy reliable computer-aided diagnostic tools for childhood OSA.

An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals (Jiménez-García et al., 2024).

Jorge Jiménez-García, María García, Gonzalo C. Gutiérrez-Tobal, Leila Kheirandish-Gozal, Fernando Vaquerizo-Villar, Daniel Álvarez, Félix del Campo, David Gozal, and Roberto Hornero. *Biomedical Signal Processing and Control*, vol. 87, part B, pp. 105490, 2024. JIF in 2022: 5.1, Q2 in "ENGINEERING, BIOMEDICAL" (JCR-WOS).

Abstract: Deep-learning algorithms have been proposed to analyze overnight airflow (AF) and oximetry (SpO_2) signals to simplify the diagnosis of pediatric obstructive sleep apnea (OSA), but current algorithms are hardly interpretable. Explainable artificial intelligence (XAI) algorithms can clarify the models-derived predictions on these signals, enhancing their diagnostic trustworthiness. Here, we assess an explainable architecture that combines convolutional and recurrent neural networks (CNN + RNN) to detect pediatric OSA and its severity. AF and SpO_2 were obtained from the Childhood Adenotonsillectomy Trial (CHAT) public database (n = 1,638) and a proprietary database (n = 974). These signals were arranged in 30-min segments and processed by the CNN + RNN architecture to derive the number of apneic events per segment. The apnea-hypopnea index (AHI) was computed from the CNN + RNN-derived estimates and grouped into four OSA severity levels. The Gradient-weighted Class Activation Mapping (Grad-CAM) XAI algorithm was used to identify and interpret novel OSA-related patterns of interest. The AHI regression reached very high agreement (intraclass correlation coefficient > 0.9), while OSA severity classification achieved 4-class accuracies 74.51% and 62.31%, and 4-class Cohen's Kappa 0.6231 and 0.4495, in CHAT and the private datasets, respectively. All diagnostic accuracies on increasing AHI cutoffs (1, 5 and 10 events/h) surpassed 84%. The Grad-CAM heatmaps revealed that the model focuses on sudden AF cessations and SpO_2 drops to detect apneas and hypopneas with desaturations, and often discards patterns of hypopneas linked to arousals. Therefore, an interpretable CNN + RNN model to analyze AF and SpO_2 can be helpful as a diagnostic alternative in symptomatic children at risk of OSA.

1.2 Context: biomedical engineering, biomedical signal processing, and artificial intelligence

The research presented in this doctoral thesis is framed in the biomedical engineering field. This is a research-oriented category of engineering that covers the application of diverse technologies in biology and medicine. Biomedical engineering is thus an interdisciplinary field in which engineers and clinicians create synergies to meet the needs of healthcare systems (Bronzino, 2006). This can be achieved by the application of chemical, mechanical, material, electrical, electronic, and optical engineering to the understanding, modification or control of biological systems, development of new devices to monitor physical activity, detection and treatment of diseases, management of large amounts of biological and medical data, etc. The field of biomedical engineering encompasses several research areas such as biomechanics, biomaterials, biotechnology, bioinformatics, and biosensors among others (Bronzino, 2006). This doctoral thesis addressed the aid in the diagnosis of pediatric OSA through a biomedical engineering perspective, applying biomedical signal processing and artificial intelligence techniques to physiological data recordings.

Biomedical signals represent the variations of a certain measurement (electromagnetic, optical, acoustic, chemical, etc.) obtained from the human body throughout time, and are frequently recorded to analyze the behavior of different groups of organs such as the cardiovascular, respiratory or nervous systems (Bronzino, 2006; Rangayyan, 2015). These recordings are useful for clinicians to identify pathological patterns and diagnose diseases. However, the interpretation of biomedical signals is a tedious task that requires expert knowledge and valuable time from the medical staff. Moreover, these signals are usually affected by noise and/or artifacts that hide the characteristics of both normal and abnormal patterns present in the biomedical recordings (Rangayyan, 2015). Consequently, signal processing solutions have been deployed to aid in the understanding of biomedical signals, to provide tools to automatically extract information from them, and to serve as a clinical decision support to diagnose several diseases (Rangayyan, 2015). The use of biomedical signal processing algorithms contribute to automate the analysis and interpretation of medical recordings, additionally reducing the human subjectivity, and potentially increasing the diagnostic reliability. The analyses carried out in this thesis relied on biomedical signal processing methods to extract useful OSA-related information from AF and SpO_2 recordings.

The major novelty of this research is related to the application of artificial intel-

ligence (AI) techniques such as ML, DL, and XAI, to perform a pattern recognition task on biomedical signals. AI is the suite of techniques and technologies that allow the development of intelligent computer programs, that is, algorithms that solve problems by applying a kind of reasoning gathered from domain knowledge. This can be achieved by applying a set of rules that allows to solve the problem or by learning a mathematical, data-driven model that infers the solution (Aggarwal, 2021; Ertel, 2017). This latter approach is the fundamental idea of ML, which refers to the set of algorithms that are able to create a model that learns from existing data and has the capability to generalize (i.e., to accurately provide the desired output on unseen data) (Aggarwal, 2021; Ertel, 2017). The learning approach can be either unsupervised, in which the algorithm finds intrinsic patterns in the data to group or segment them, or supervised, in which examples are labeled with the expected output and the algorithm try to match this solution from input data (Aggarwal, 2021; Ertel, 2017). Regarding the output, supervised learning approaches can be split into classification (categorical output) and regression (continuous variable) (Witten et al., 2011). This doctoral thesis focused exclusively on the implementation of supervised ML methods, given that the diagnosis of pediatric OSA was addressed by detecting the presence and severity of OSA through classification and regression algorithms.

Classical supervised ML techniques were based on simple mathematical models such as linear regression, discriminant analysis, logistic regression (LR), decision trees (DT), and support vector machines (SVM) (Witten et al., 2011). However, these simple models were limited by their ability to derive a sufficiently accurate model in some datasets (i.e., high bias). When these mathematical models evolved to more complex, nonlinear solutions, they became more sensitive to the particularities of limited training data, thus failing to retain their generalizability (i.e., high variability) (Witten et al., 2011). Ensemble learning is an umbrella term that gathers all ML approaches aimed at producing an accurate and generalizable model by combining the outputs of several diverse, simple models (Kuncheva, 2014; Sagi and Rokach, 2018). The idea behind ensemble learning is that a committee of experts will take better decisions by reaching a consensus than any of them alone (Sagi and Rokach, 2018). This way, ensemble learning approaches solve a complex problem by dividing it into simple parts and combining their solutions, in a divide-and-conquer philosophy (Kuncheva, 2014). In this doctoral thesis, an ensemble learning algorithm was implemented to derive the diagnosis of pediatric OSA trough the information extracted from AF and SpO_2 signals (Jiménez-García et al., 2020).

Neural networks (NNs) fall into another specific branch of ML algorithms inspired in the behavior of large, complex interconnections of brain neurons. They mimic the propagation and processing of electric stimulus throughout the nervous system (Goodfellow et al., 2016). Feed-forward NNs became widespread in many tasks during the early 2000s, but the research interest in developing deep NNs (DNNs) such as CNNs and RNNs was limited due to computational demands. However, the enthusiasm on DL algorithms expanded after the *AlexNet* CNN algorithm largely surpassed all image recognition methods that competed in the 2012 ImageNet challenge (Goodfellow et al., 2016). This way, the development of DNN-, CNN- and RNN-based applications exploded, becoming the state-of-the art in almost all pattern recognition tasks (Lecun et al., 2015). The characteristics of DL algorithms facilitate the development of pattern recognition methodologies able to learn from raw data (i.e., images, signals, text, etc.) rather than human-engineered features (Goodfellow et al., 2016). Thus, DL unifies the learning process by automatically identifying the relevant patterns in the data and utilize them to make predictions (Lecun et al., 2015). In this doctoral thesis, CNN and RNN algorithms were implemented to automatically analyze AF and SpO_2 signals and estimate the presence and severity of OSA (Jiménez-García et al., 2022, 2024).

Despite the high capabilities of current ML and DL algorithms to automatically process data and derive top-performing models, they have to face an important drawback related to their transparency. Complex ML models, especially DL approaches, usually fail to supply explanatory information about the decision making process (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020). This shortcoming has motivated part of the AI research community to address the interpretability of ML and DL architectures by proposing models, algorithms, and methodologies into the framework of XAI (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020). This new field is aimed at proposing interpretable ML solutions, that is, developing accurate models that also have the capability to explain the decision-making process (Barredo Arrieta et al., 2020). This way, final users can understand and manage AI, and finally trust in the ML-derived predictions (Barredo Arrieta et al., 2020). The deployment of XAI-grounded solutions is crucial in the medical field involving the use of ML or DL as a diagnostic aid tool, since both clinicians and patients need to comprehend and trust the decision-making process. In this view, several XAI methods have been applied in biomedical data such as clinical features or medical images, but these methods have not been extensively used in other data sources like biomedical signals (Loh et al., 2022). The application of XAI has been addressed in this doctoral thesis in order to interpret the patterns

of AF and SpO_2 signals associated to the presence of apneic events detected by our DL architecture (Jiménez-García et al., 2024).

This doctoral thesis focused on the analysis of AF and SpO_2 signals to derive ensemble and DL models that can serve as an aid in the diagnosis of pediatric OSA. Within a biomedical engineering research framework, we deployed biomedical signal processing techniques together with various ensemble learning and DL approaches such as AdaBoost, CNNs, and RNNs to estimate the presence and severity of OSA in children, and also enhanced the explainability of DL approaches using Grad-CAM. Thus, the topics addressed in this section constitute the research framework of this doctoral thesis.

1.3 Pediatric obstructive sleep apnea (OSA)

1.3.1 Definition and prevalence

Sleep apnea (SA) is characterized by respiratory cessations during sleep. Concretely, OSA is a sleep disorder in which the upper airway is recurrently blocked or narrowed during sleep, leading to intermittent episodes of obstructive apneas or hypopneas, respectively (Dehlink and Tan, 2016; Moffa et al., 2020). These obstructions interrupt the normal respiratory AF, are linked to increased respiratory effort, and cause episodes of oxyhemoglobin saturation (blood oxygen saturation, SaO_2) reductions (Bitners and Arens, 2020; DelRosso, 2016). Apnea events may also occur without obstruction during sleep, leading to Central Sleep Apnea (CSA), characterized by the absence of cerebral stimuli to continue breathing and therefore by the absence of respiratory effort as well (Bitners and Arens, 2020). Mixed apneas have characteristics of both obstructive and central apnea during the event (Bitners and Arens, 2020). The presence of OSA or CSA leads to sleep disruptions that alter the normal development of sleep stages and affected children end up suffering restless sleep and its consequences (Dehlink and Tan, 2016).

The American Academy of Sleep Medicine (AASM) defined specific rules to identify apneas and hypopneas in children from sleep studies (Berry et al., 2012). According to these rules, AF reductions $\geq$90% with respect to previous normal respiration during at least 2 breathing periods are considered apneas, while hypopneas are defined as AF reductions $\geq$30% for two or more breaths followed by a drop $\geq$3% in the SpO_2 levels or an arousal (Berry et al., 2012). The AASM establishes more restrictive criteria to score apneic events in children compared to adults, as long as the minimum duration is 10 seconds for adults and 2 breaths

(5-7s) for children (Berry et al., 2012).

Pediatric OSA is a prevalent sleep disorder. The most common symptoms of OSA include snore, breath pauses, daytime sleepiness, and morning headaches, but these and many other symptoms can be subtle and not easily noticed by parents and caregivers (Tauman and Gozal, 2011). Snoring children usually have suspicion of suffering OSA, and approximately represent 7.2% of the pediatric population (Li et al., 2010; Marcus et al., 2012). However, only a percentage of these children actually have OSA (Tauman and Gozal, 2011). An international study involving 4,191 pediatric subjects worldwide revealed that 72.9% of children referred to a PSG study due to suspicion of OSA (reported snore and/or breathing pauses) were finally diagnosed with any degree of OSA severity, suggesting that the majority of symptomatic children actually have OSA (Hornero et al., 2017). This result is coherent with those obtained in some epidemiological studies that estimated the prevalence of OSA is up to 5.7% while others reported a more conservative rate between 1% and 5% (DelRosso, 2016; Marcus et al., 2012). Notwithstanding this high prevalence, the actual rate of affected children may be higher due to a large number of cases that are not properly diagnosed as a consequence of the unavailability and large waiting lists of sleep laboratories (Brockmann et al., 2018).

1.3.2 Causes and consequences

The main cause of OSA is the normal respiration disturbance following upper airway collapse. In children, this is due to the relaxation of muscles and other tissues surrounding the naso- and oropharynx combined with the upper airway narrowing caused by adenotonsillar hypertrophy, obesity, craniofacial abnormalities and neuromuscular disorders (Moffa et al., 2020). Adenoids and tonsils hypertrophy usually favors the development of OSA in children, since enlarged adenoids and tonsils can occlude the nasopharynx and oropharynx during sleep, respectively (Moffa et al., 2020). Figure 1.2 shows different degrees of tonsillar hypertrophy that can lead to throat obstruction. The higher the degree of hypertrophy, the more collapsible the upper airway is (Brodsky, 1989). Adenotonsillar hypertrophy is a normal condition in the childhood and coincides with the pediatric OSA peak of prevalence (from 2 to 8 years old) (DelRosso, 2016). Another common but less frequent cause of larynx obstruction is the presence of adipose tissue in the neck of obese children, making obesity a risk factor of pediatric OSA (DelRosso, 2016; Moffa et al., 2020). Maxillofacial or craniofacial anatomy abnormalities such as macroglossia, retrognathia, micrognathia, and midfacehypoplasia can also favor

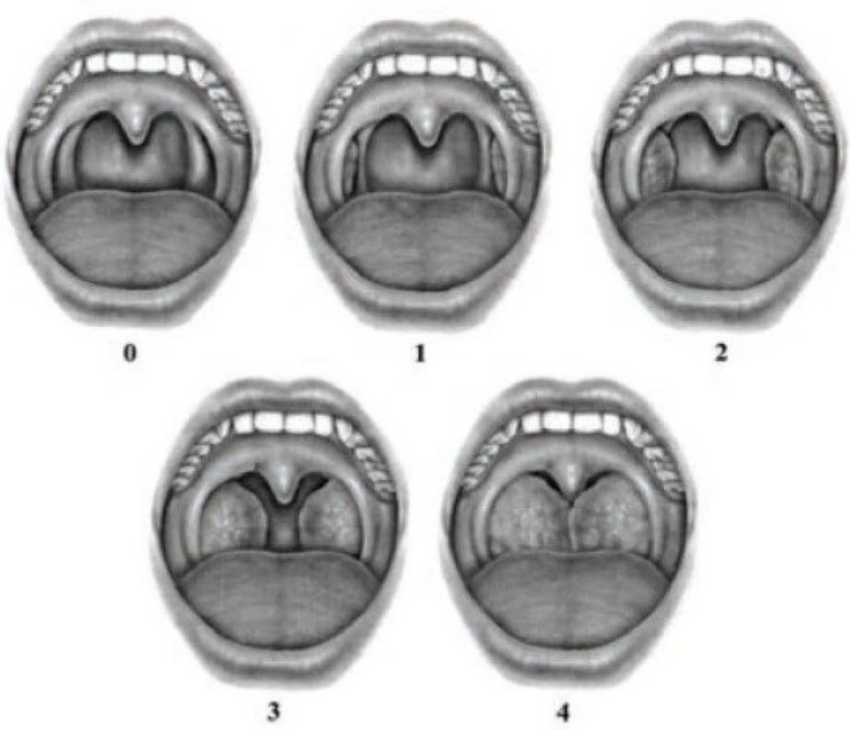


Figure 1.2: Degrees of tonsillar hypertrophy ranked from 0 (absence) to 4 (almost complete obstruction). Adapted from Callén Blecua (2017) with permission (CC BY-NC-ND 3.0).

airway obstruction and predispose to childhood OSA (Bitners and Arens, 2020; DelRosso, 2016). These abnormalities are present in children with conditions like Down syndrome, Prader Willi syndrome, or Pierre Robin sequence, who are frequently affected by OSA (Bitners and Arens, 2020; DelRosso, 2016; Moffa et al., 2020).

Childhood OSA affects the normal development of sleep, leading to restless sleep, enuresis, daytime hypersomnolence, along with growth and development alterations (Joosten et al., 2017). The negative outcomes of OSA are classified mainly in two groups: neuro-behavioral and cardio-metabolical (Blechner and Williamson, 2016). Behavior alterations like hyperactivity, impulsivity, and aggressivity, as well as neurocognitive deficits like attention, executive function, and language deficits have been associated to the presence of OSA in children (Bitners and Arens, 2020). Regarding cardiometabolical alterations, the most common manifestations are metabolic syndrome, pulmonary and arterial hypertension and

long-term endothelial dysfunction (Blechner and Williamson, 2016). It has also been observed that obesity can be both cause and consequence of pediatric OSA since the daytime sleepiness and tiredness can limit the physical activity of children (Blechner and Williamson, 2016). These negative outcomes of pediatric OSA impact the future well-being of children, and stress the importance of early diagnosis of children at risk in order to mitigate future comorbidities.

1.4 OSA diagnosis in children: polysomnography (PSG)

Pediatric specialists should assess children with symptoms related to OSA in primary care facilities and refer them to the sleep specialist to confirm the diagnosis (Meltzer and Paisley, 2023). Overnight sleep studies are recommended to diagnose SDB related diseases such as OSA. The gold standard to diagnose pediatric OSA is an in-lab PSG, in which neuronal, cardiorespiratory, muscular, position and movement signals are recorded during the night and subsequently analyzed (Berry et al., 2020; Jon, 2009). These recordings are obtained by placing electrodes and other sensors on the patient's body (Jon, 2009). Respiratory channels such as oronasal AF, nasal pressure, thoracic and abdominal movements, as well as SpO_2 reflect the changes in normal respiration due to OSA (Jon, 2009; Stowe and Afolabi-Brown, 2020). As mentioned in Section 1.3.1, the information of AF and SpO_2 signals included in the PSG is used by sleep specialists to score apnea and hypopnea events in sleep studies (Berry et al., 2020; Mazzotti et al., 2018). The rate of apnea or hypopnea events per hour (e/h) of sleep is the AHI, which constitutes the main indicator of OSA presence and severity (Bitners and Arens, 2020; Moffa et al., 2020). The AHI thresholds of 1, 5, and 10 e/h are established to define mild, moderate, and severe OSA. This way, the 4-level classification of pediatric OSA is the following (Bitners and Arens, 2020; Moffa et al., 2020):

- No OSA: AHI $<$ 1 e/h.
- Mild OSA: 1 e/h $\leq$ AHI $<$ 5 e/h.
- Moderate OSA: 5 e/h $\leq$ AHI $<$ 10 e/h.
- Severe OSA: AHI $\geq$ 10 e/h.

These thresholds are considerably lower than those used to define OSA severity in the adult population (5, 15, and 30 e/h), indicating that the criteria to score apneic

events (introduced in Section 1.3.1) to diagnose pediatric OSA is more stringent (Berry et al., 2012). This is due to the children-specific characteristics of upper airway obstructions during sleep, which are more subtle than those observed in adults (e.g., lower duration and less severe desaturations). As a consequence, adult criteria is unable to correctly identify obstructive events in children (Rosen et al., 1992). Moreover, the negative consequences of OSA in the pediatric population have tightened the diagnostic requirements. While 5 e/h is the AHI cutoff used to diagnose the lowest OSA severity in adults, this threshold is established to refer surgical treatment in children (Kaditis et al., 2016b).

1.4.1 Alternatives to polysomnography

Notwithstanding the preference of PSG to diagnose pediatric OSA, its shortcomings have motivated the search for diagnostic alternatives. The main drawback of PSG is the limited availability, caused by the scarcity of pediatric sleep laboratories at hospitals. Reasons why PSG is relatively unavailable are related with the costs, both material and personnel (Dehlink and Tan, 2016; Tan et al., 2015). The equipment to perform PSG includes electroencephalography (EEG), electrooculography (EOG), electrocardiography (ECG), electromiography (EMG), respiratory monitoring, pulse oximetry, etc., that make it thorough but expensive (Jon, 2009; Mazzotti et al., 2018; Riha et al., 2023). Regarding personnel costs, the interpretation of PSG requires specialized medical staff to manually score sleep stages, apneic events, and other findings in the physiological signals (Borrelli et al., 2023). OSA diagnosis through PSG is thus demanding and inefficient. Another limitation of performing PSG in children is the disturbance of spending the night hospitalized in a sleep laboratory with several sensors and wires attached to the body, that could be distressing for them and result in a non-representative sleep study (Stowe and Afolabi-Brown, 2020). The aforementioned causes enlarge the waiting lists and delay the access to a convenient diagnosis.

The efforts to mitigate the drawbacks of PSG have focused on developing diagnostic alternatives that reduce the complexity and costs while increase the availability of a proper OSA diagnosis (Brockmann et al., 2018). This can be achieved by using less signals and enabling the possibility to perform the tests at patients' home with portable equipment (Tan et al., 2015). Sleep studies can be classified in 4 types according to the number of signals involved in the recording and the possibility of performing them outside the hospital (Riha et al., 2023):

- Type 1 (In-lab PSG): Involves the recording of up to 32 biomedical signals

in an attended sleep laboratory.

- Type 2 (Portable PSG): The sleep studies can be performed at patient's home with a portable device that records a minimum of 7 physiological signals: EEG, EOG, EMG, ECG, SpO_2, AF, and respiratory effort.
- Type 3 (Respiratory polygraphy, RP): Also known as Home Sleep Apnea Test (HSAT), it comprises the recording of 4 to 7 signals, including ECG, SpO_2 and two or more respiratory channels such as AF (nasal pressure and/or thermistor) and respiratory movements (thoracic and/or abdominal effort).
- Type 4 (Single or dual channel approaches): The simplest sleep studies that do not meet Type 3 criteria are usually recordings of nocturnal pulse oximetry and/or single channel AF.

All these alternatives to perform sleep studies have advantages and drawbacks. Type 1 and 2 devices are more complex but include the recording of EEG, EOG, and EMG channels, which is useful to identify sleep stages and derive sleep parameters such as latency, wake after sleep onset and total sleep time, among others (Stowe and Afolabi-Brown, 2020). On the other side, type 3 and 4 devices greatly reduce complexity and costs compared to type 1 and 2 approaches, since they only focus on cardiorespiratory signals involved in the identification of apneas and hypopneas (Tan et al., 2015). HSAT is well established as an alternative to PSG in adullt OSA, but there was much more skepticism about its effectiveness in children (Oceja et al., 2021; Tan et al., 2015). Recent studies comparing RP against PSG to diagnose childhood OSA pointed out the feasibility of these simplified tests, especially to primarily diagnose moderate or severe OSA (Alonso-Álvarez et al., 2015; Chiner et al., 2020; Tan et al., 2014). In addition to this, single channel approaches have also gained relevance to evaluate OSA in symptomatic children (Kaditis et al., 2016a). In summary, all these simplifications of PSG allow cost reduction, increased availability, and patient comfort.

Following the trend of aforementioned simplified diagnostic tests, this doctoral thesis was aimed at reducing the number of required signals. Here, AF and SpO_2 recordings were exclusively used to derive an automatic diagnosis of OSA and its severity based on the AHI. As mentioned in Sections 1.3.1 and 1.4, these two signals are mainly involved in the detection of apneas and hypopneas, and therefore to derive the AHI. The AF signal is presented in Section 1.4.2, and the SpO_2 signal is introduced in Section 1.4.3.

1.4.2 The airflow (AF) signal

Respiratory AF measures the flow of inhaled/exhaled air throughout time. The most accurate way to quantify AF is the pneumotachography, but it is unfeasible to perform in sleep studies since it requires placing an obtrusive mask in the patient's face (Roebuck et al., 2014; Shokoueinejad et al., 2017). AF can be easily and non-invasively measured during sleep using nasal pressure and/or oronasal thermal sensors. The former detect the differential pressure during inhalation and exhalation while the latter respond to temperature changes (the sensor is cooled during inhalation and warmed during exhalation). This way, both AF signal waveforms represent the cyclic pattern of respiration (Roebuck et al., 2014). According to the AASM recommendations, thermistor is the choice to score apneas, while nasal pressure is preferred to score hypopneas. Nevertheless, thermistor is also the recommended alternative to score hypopneas if nasal pressure sensor is not available, and they are able to register both nasal and mouth breathing (Berry et al., 2012). In addition, the improved sensitivity of polyvinylidene fluoride (PVDF) sensors make oronasal thermistors a reasonable choice to record respiratory AF and score both apneas and hypopneas using only one channel (Berry et al., 2012; Shokoueinejad et al., 2017).

Figure 1.3 represent intervals of the AF signal linked to normal respiration as well as during apnea and hypopnea events. It can be seen that the amplitude of AF oscillations determine the presence of these apneic events. However, some AF reductions $\geq$ 30 % are not considered hypopneas unless they are associated to a desaturation $\geq$ 3% or an arousal.

1.4.3 The oximetry (SpO_2) signal

The persistent occurrence of apneic events diminish the amount of oxygen available in the lungs to be transferred to the hemoglobin (Hb) molecules of the blood. This results in intervals of hypoxemia, that is, low levels of oxygen in the blood (Marcus et al., 2012). SaO_2 is the percentage of Hb molecules carrying oxygen (oxyhemoglobin, O_2Hb) with respect to the total Hb in the arterial blood (Chan et al., 2013). This parameter can be continuously monitored with a pulse oximeter, which computes SpO_2 and pulse rate (PR) from optical photoplethysmography (PPG) signals obtained from peripheral tissues such as a finger, toe, or ear lobe (Allen, 2007). This is achieved by illuminating them from one side and receiving the trespassing light on the other side (Allen, 2007; Chan et al., 2013). Pulse oximeters exploit the optical characteristics of peripheral tissues and their vari-

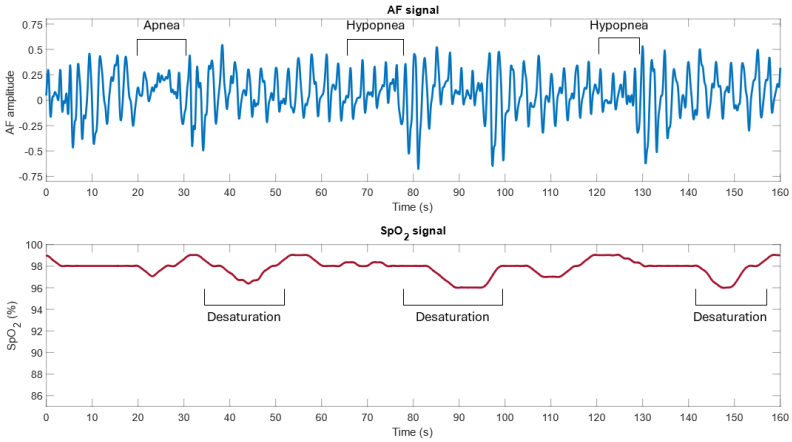


Figure 1.3: Example of airflow (AF) and oximetry (SpO_2) signal with presence of apnea and hypopnea events, together with their associated desaturations.

ations in presence of pulsatile arterial blood circulation. These devices obtain PPG signals at two different wavelengths in the spectrum of red (660 nm) to near-infrared (940 nm) light, which amplitude substantially differ for different SpO_2 levels (Allen, 2007; Chan et al., 2013). Then, the amplitude ratios between direct current (DC) and alternating current (AC) components of PPG are continuously measured to compute the SpO_2 signal (Chan et al., 2013).

Figure 1.3 shows an example of the SpO_2 signal with recurrent desaturation events caused by consecutive apneas and hypopneas. These events begin with a drop of the SpO_2 level few seconds after the apnea/hypopnea is observed in the AF signal, and end after normal respiration is reestablished, so they are delayed with respect to the apneic events.

The use of pulse oximeters has become widespread and portable SpO_2 recorders, viewed as a PSG alternative, have been proposed to simplify OSA diagnosis (Chan et al., 2013; del Campo et al., 2018). Some studies have indicated the diagnostic usefulness of SpO_2 as a screening tool in adults, but the scientific evidence is more scarce in the case of children (del Campo et al., 2018). In this context, SpO_2-derived parameters such as ODI 3% or McGill score have shown good predictive power in symptomatic children (Kaditis et al., 2016a). Although these parameters have shown good predictive power for moderate-to-severe OSA, they still miss the mildest cases (i.e., underestimate the actual OSA severity) (Brouillette et al., 2000; Kaditis et al., 2016a; Van Eyck and Verhulst, 2018). To overcome this issue, some researchers suggested that SpO_2 can be complemented

with other signals such as AF or PR in order to increase its diagnostic ability (Alonso-Álvarez et al., 2015; Van Eyck and Verhulst, 2018). This can be achieved by using both sources to automatically compute OSA severity via ML approaches (Gutiérrez-Tobal et al., 2022; Van Eyck and Verhulst, 2018).

1.5 State-of-the-art: machine learning approaches to aid in the diagnosis of pediatric OSA

Due to the aforementioned disadvantages of PSG-based sleep studies, the research focus has pointed to the automatic processing of these biomedical recordings (Mazzotti et al., 2018). This way, a reduced set of PSG signals such as EEG, AF, SpO_2, PPG, or ECG can be analyzed using computer-based algorithms aimed at detecting signs of OSA pathology and investigate novel biomarkers related to the disease (Mazzotti et al., 2018). These algorithms can also provide features beyond classical descriptors such as sleep efficiency, ODI 3%, or the percentage of sleep time spend with $SpO_2 < 90\%$ using spectral, nonlinear and many other techniques (Mazzotti et al., 2018; Mendonca et al., 2019). Moreover, these characterization methods have been combined with several ML models with the objective of determining the presence of apneic events in the signals as well as deriving an OSA diagnosis (Mendonca et al., 2019; Uddin et al., 2018). However, a large part of this research has been done in adult OSA cohorts, and the implementation of these techniques in childhood OSA is still pending (Gutiérrez-Tobal et al., 2022).

Regarding the application of automatic methods to detect pediatric OSA, the most commonly used signals are AF , SpO_2, ECG-derived heart rate variability (HRV), and PPG-derived pulse rate variability (PRV) (Bertoni and Isaiah, 2019; Gutiérrez-Tobal et al., 2022). It is worth noting that parameters such as OSA symptoms, clinical findings, ODI, and other sociodemographic variables have also been used to derive automatic models based on LR (Chang et al., 2013; Skotko et al., 2017; Wu et al., 2017). The study of Calderón et al. (2020) focused on a comparison of LR, SVM, and AdaBoost models fed with oximetric variables such as ODI 3% and 4%, the number of SpO_2 drops greater than certain levels, and the percentage of time spent under 90% and 92% of SpO_2, obtained from a database of 453 subjects. Other studies evaluated ML algorithms such as LR, bootstrap aggregation, random forest, and SVM trained with demographical and clinical data, OSA-related findings, ODI, and actigraphy-derived parameters (Bertoni et al.,

2020).

ECG is one of the most frequently analyzed signals in sleep studies, given that the effects of respiratory events are also noticeable in the cardiac pattern (Mazzotti et al., 2018). Apneas and hypopneas produce changes in the HRV as a response to the lack of oxygen, characterized by repeated episodes of bradycardia and tachycardia. This characteristic has been exploited in some studies aimed at assessing the ECG and HRV. Shouldice et al. (2004) applied temporal and spectral techniques to classify ECG signal segments into normal and apneic and derive a rate of apneic segments on the overnight signal. Other recent studies proposed the spectral analysis of the HRV signal in specific frequency bands, which were more suitable to characterize pediatric OSA (Martín-Montero et al., 2021b). This approach was further extended to the analysis of bispectral features and the diagnostic ability of a feed-forward NN was also assessed (Martín-Montero et al., 2021a).

Undoubtedly, the ease of use of pulse oximeters together with the diagnostic usefulness of pulse oximetry signals have placed these devices in an advantageous place to substitute PSG. The majority of studies dealing with the diagnosis of pediatric OSA analyzed SpO_2, PRV, or PPG signals, either alone or combined. The studies of Gil et al. (2010) and Lazaro et al. (2014) analyzed the PPG waveforms, focusing on the amplitude fluctuation decreases and the pulse transit time variability as well as the spectral analysis of PRV, respectively. To do that they utilized a database of 21 pediatric subjects and linear discriminant analysis (LDA) classifiers to derive the OSA diagnosis. The usefulness of PRV has also been assesed in the study conducted by Dehkordi et al. (2016). Temporal and spectral parameters, together with detrended fluctuation analysis (DFA) were computed and presented to a least absolute shrinkage and selection operator (LASSO) classifier to identify patients with and without SDB. The combination of PRV and SpO_2 have also been addressed by Garde et al. (2014), who combined statistical, spectral, and non-linear features from both signals in a LDA classification model. The PhoneOximeter™mobile device has been used in these studies to record a total of 146 PPG signals, with PRV and SpO_2 subsequently obtained from that signal. This approach was further assessed with a modified set of SpO_2 and PRV features as well as LR models to detect three severity levels of SDB (Garde et al., 2019).

Another group of studies involved the utilization of single-channel SpO_2 alone. A multicenter study involving 13 different hospitals assessed the diagnostic usefulness of a NN-based algorithm using a database of 4,191 children worldwide (Hornero et al., 2017). To our knowledge, this is the largest database of pedi-

atric sleep studies aimed at OSA detection. In the development of the algorithm, ODI 3%, time-domain statistics, spectral and nonlinear features were extracted from overnight SpO_2, and the final NN model was trained with a subset of non-redundant features (Hornero et al., 2017). This model was further validated in a subsequent study involving 432 additional patients (Xu et al., 2019). Looking for a more thorough characterization of SpO_2 behavior in the presence of OSA, other studies proposed techniques such as symbolic dynamics and DFA to enhance the diagnostic ability of this signal (Álvarez et al., 2018; Vaquerizo-Villar et al., 2018a). Alternatively, bispectral and wavelet analyses were also considered (Gutierrez-Tobal et al., 2018; Vaquerizo-Villar et al., 2018b,c). Regarding the ML models, these studies mainly implemented LR, SVM, or NN algorithms. Other classifiers such as LDA and quadratic discriminant analysis (QDA) have been also assessed in similar studies (Crespo et al., 2018).

The AF signal has been analyzed using automatic methods to fully characterize its behavior in presence of normal respiration and apneic events. Preliminary studies addressed the spectral analysis of single channel AF as well as the irregularity and variability of this signal using spectral entropies and central tendency measure (CTM), respectively (Barroso-García et al., 2017; Gutiérrez-Tobal et al., 2015). The latter analyses also involved the AF-derived respiratory rate variability (RRV) (Barroso-García et al., 2017). LR classifiers were implemented to derive the final binary OSA classification (Barroso-García et al., 2017; Gutiérrez-Tobal et al., 2015). Delving into this characterization, recurrence plots and more advanced bispectral and wavelet analyses were also proposed (Barroso-García et al., 2020, 2021a,b). The features obtained from these methods were used to derive the AHI with different single-layer NN approaches. Of note, one of these studies also assessed an AdaBoost classifier to predict OSA severity (Barroso-García et al., 2021b). It was observed from these studies that the diagnostic ability of AF-driven models increases when the ODI 3% is also included in the feature sets, highlighting the complementarity of AF with this oximetric index (Barroso-García et al., 2020, 2021a,b).

As mentioned in the previous paragraphs, most of the ML models used to predict pediatric OSA followed the traditional pipeline of pattern recognition algorithms: feature extraction, selection and classification. In spite of the widespread implantation of DL to evaluate adult OSA, the number of studies that applied DL models to detect OSA in children has been more limited (Gutiérrez-Tobal et al., 2022; Mostafa et al., 2019). One study utilized wavelet scalograms of the AF signal recorded from a nasal pressure sensor to detect events of obstructive apnea and

hypopnea, as well as central apneas, with DL (Crowson et al., 2023). Several pre-trained CNN architectures were adapted to predict the presence of apneic events using transfer learning and a database of 936 AF segments from 28 patients, but no subject-based classification was performed (Crowson et al., 2023). To our knowledge, only Vaquerizo-Villar et al. (2021) proposed a CNN algorithm to predict pediatric OSA severity from single-channel SpO_2 recordings. This method was developed using a public database of 1,638 sleep studies and externally validated and tested in two proprietary datasets comprising 980 and 587 subjects. Similarly, the study of García-Vicente et al. (2023) presented a CNN-based architecture to derive the AHI from overnight ECG. In this case, the algorithm was deployed and tested using a public database comprising 1,610 recordings. None of the aforementioned DL-based approaches to detect OSA from AF, SpO_2, or ECG involved the analysis of these signals jointly.

The application of XAI techniques to complement ML- or DL-based decision-making algorithms is increasingly common in the medical field. The most widespread XAI approaches in biomedical applications are Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) in the case of feature-based ML models, and Grad-CAM is the preferred choice when dealing with DL models such as CNNs (Loh et al., 2022). However, there is a lack of studies aimed at interpreting pediatric OSA detection models. As proof of this, previously mentioned studies did not apply any XAI technique to reveal OSA-related patterns learned by their architectures. Only one study applied the SHAP framework to increase the interpretability of a feature-based XGBoost that combined ODI with heart rate statistics and other clinical variables (Ye et al., 2022). Vaquerizo-Villar et al. (2023) interpreted sleep-related EEG patterns using CNN-derived Grad-CAM heatmaps aimed at classifying sleep stages in children. This XAI algorithm was also used in other sleep-related studies in adults, encompassing EEG-based sleep stages detection (Dutt et al., 2022; Kuo et al., 2021), and apneic events detection from cardiorespiratory signals (Serrano Alarcón et al., 2023).

After reviewing the scientific contributions to the topic of pediatric OSA detection, three key points can be extracted: (i) these approaches typically involved only one signal (mainly SpO_2, AF, PPG, PRV, ECG, and HRV) or combinations of two different sources such as PPG+PRV, SpO_2+PRV, or AF+ODI; (ii) classical ML models such as LR, LDA, SVM and shallow NNs have been widely assessed, but the number of studies involving ensemble learning approaches is more reduced; (iii) there is a scarcity of studies that have implemented and assessed DL algorithms

to detect and quantify pediatric OSA, especially those that are interpretable by human users (e.g., by applying XAI techniques). AF and SpO_2 fluctuations define the presence of apneas or hypopneas, and they have also shown their usefulness and complementarity in the detection of adult OSA (Álvarez et al., 2010, 2020; Gutiérrez-Tobal et al., 2013; Marcos et al., 2012). In this context, the AdaBoost ensemble learning method also has been successfully tested using AF and SpO_2 separately (Gutierrez-Tobal et al., 2016; Gutiérrez-Tobal et al., 2019). Regarding the application of DL on physiological signals, there is a continuous and fast growth of this type of architectures (Faust et al., 2018; Mostafa et al., 2019). Although several approaches have been presented in the last years using different OSA-related signals and DL architectures (Choi et al., 2018; Erdenebayar et al., 2019; Leino et al., 2021; Nikkonen et al., 2021), their deployment in the field of pediatric OSA is still limited. For all these reasons, the approaches covered in this thesis involved the analysis of AF and SpO_2 signals using ensemble learning methods such as AdaBoost and DL architectures such as CNNs and RNNs combined with XAI algorithms like Grad-CAM. The complementarity of these two overnight signals was assessed by means of feature engineering and AdaBoost in the first article of the compendium (Jiménez-García et al., 2020), whereas the two subsequent papers focused on evaluating the diagnostic ability of two DL architectures (CNN and CNN+RNN, respectively) trained with raw AF and SpO_2 data (Jiménez-García et al., 2022, 2024). Finally, the explainability of the CNN+RNN architecture was addressed in the last article (Jiménez-García et al., 2024).

The medical and technical context, together with the description of the compendium of publications that constitutes the present doctoral thesis have been introduced in this chapter. The hypothesis and objectives of this research are introduced in Chapter 2 and the methodology is described in Chapter 3.

Chapter 2

Hypothesis and objectives

Automatic signal processing together with increasingly popular AI algorithms have boosted the development of computer aided diagnosis solutions. As mentioned in Section 1.5, the simplification of PSG for pediatric OSA diagnosis has been addressed using different ML and DL algorithms that processed the information of biomedical signals such as AF and SpO_2 in order to estimate the presence and severity of OSA. The works that constitute this doctoral thesis are framed into this research line, focused on assessing ensemble and DL algorithms together with XAI to aid in the diagnosis of OSA in children. The hypothesis and objectives of this thesis are introduced in this chapter. Section 2.1 summarizes the research hypotheses of this work, and Section 2.2 present the main and specific objectives of the thesis.

2.1 Hypotheses

The definitions of apnea and hypopnea events direcly involve AF and SpO_2 signals (Berry et al., 2012). Approaches aimed at simplifying PSG such as HSAT or other Type 4 sleep studies have usually included these two signals, which provide enough information to score apneas, hypopneas, and their associated desaturations (Alonso-Álvarez et al., 2015; Kaditis et al., 2016a). Thus, ML models could combine both information sources to maximize their joint diagnostic ability and also minimize the number of signals to be analyzed. In this sense, it has been regarded that the information of AF can be complemented with oximetric indices such as ODI 3% (Barroso-García et al., 2020). It is assumed in this work that *the complementary information of AF and SpO_2 suffices to help detect pediatric OSA*

using ML-based approaches.

As mentioned in Section 1.5, ML-based approaches have been proposed in the literature to deal with the automatic analysis of polysomnographic recordings in the OSA detection context. However, these studies have been more scarce in the pediatric population (Gutiérrez-Tobal et al., 2022). In addition, the data-driven approaches covered classical and widespread ML models such as LDA, LR, SVMs, and shallow NNs (Barroso-García et al., 2020; Crespo et al., 2018; Gutiérrez-Tobal et al., 2015; Hornero et al., 2017; Vaquerizo-Villar et al., 2018c). These ML algorithms usually failed to reach the diagnostic accuracies shown in the case of adults. Ensemble learning algorithms have not been extensively tested in the context of childhood OSA, and existing approaches were restricted to assess the diagnostic value of clinical, oximetry- or actigraphy- derived features (Bertoni et al., 2020; Calderón et al., 2020). Moreover, ensemble learning models like AdaBoost accomplished high diagnostic ability in adults (Gutierrez-Tobal et al., 2016; Gutiérrez-Tobal et al., 2019). For these reasons, the research carried out in this thesis assumed that *more advanced ML algorithms, such as those belonging to the ensemble learning family, can aid to enhance the diagnostic accuracy of the most widespread ML-based approaches in pediatric OSA.*

Notwithstanding the potential shown by the algorithms that combine feature-engineering and ML techniques, these are obviously limited by the capability of domain experts to obtain, compute, select and analyze useful descriptors. Moreover, some studies found that an exhaustive signal characterization towards OSA detection in children did not contribute to find truly relevant and complementary features (Hornero et al., 2017). DL algorithms can overcome this limitation since they are capable to learn complex features with a high abstraction level directly from raw data (Lecun et al., 2015). Thus, this research is carried out under the assumption that *DL-based approaches can learn the necessary information to detect pediatric OSA directly from AF and SpO_2 raw data.*

An important shortcoming of DL models arises from their lack of interpretability. The application of XAI methods is becoming increasingly common in the medical field, mainly aimed at seeking the signs of pathology learned by the models and interpreting these signals patterns (Loh et al., 2022). This is advantageous for enhancing the trust of end users in these AI-derived diagnostic aids. The DL architectures proposed in this thesis were interpreted considering that *XAI methods can aid to identify relevant patterns linked to the presence of OSA in the AF and SpO_2 signals from pediatric patients.*

Based on the aforementioned considerations, this doctoral thesis is grounded

in the following hypothesis:

> "*The global hypothesis of this research is that the automatic analysis of overnight AF and SpO_2 signals using state-of-the-art data-driven algorithms together with XAI techniques can help to simplify the diagnosis of childhood OSA.*"

2.2 Objectives

The main objective of this doctoral thesis is to study, develop and validate advanced ML and DL methods together with new XAI techniques in the context of automatic analysis of AF and SpO_2 signals, so that these methods can be used to help diagnose pediatric OSA. To achieve this main objective, the following specific objectives are proposed:

I. To elaborate and analyze a database of nocturnal AF and SpO_2 recordings from PSG performed on pediatric subjects with suspected OSA, including their sociodemographic and clinical data related to the presence and severity of the disease.

II. To evaluate the complementarity of the information extracted from AF and SpO_2 signals using feature-engineering in conjunction with classification- and regression-aimed pattern recognition methods, in order to leverage their joint diagnostic performance. Determine if the combination of these two signals outperforms each of them separately.

III. To evaluate the diagnostic ability of a selection of advanced ML and DL methods trained with useful and complementary information of AF and SpO_2 as well as with raw AF and SpO_2 signals, respectively, all of them aimed at estimating AHI and/or classifying OSA severity from these nighttime records.

IV. To identify the most relevant AF and SpO_2 patterns that DL methods link to the presence of apneas and/or hypopneas and use to detect OSA by using XAI techniques.

Chapter 3

Materials and Methods

This chapter describes the databases used in the different studies of the compendium of publications and summarizes the methodology conducted in this doctoral thesis. Two different databases, one public and other private, are described in Section 3.1. Regarding the signal acquisition process, Sections 3.1.1 and 3.1.2 describe technical considerations about AF and SpO_2 recordings of both databases, respectively. Section 3.2 presents a scheme of the signal processing and analysis methodology that was implemented in the articles that constitute this research. Note that the databases utilized in this doctoral thesis, as well as the methodology described in the second part of this chapter are thoroughly explained in the articles that constitute the compendium of publications (see Chapter 7).

3.1 Databases

The first database used in the studies conducted in this thesis was provided by the Comer Children's Hospital, University of Chicago (UofC) School of Medicine (Chicago, IL, USA). This database comprised 974 pediatric subjects with ages up to 13 years old. The parents or legal caretakers of all children gave their informed written consent for participating in the study. The research protocols accomplished the Declaration of Helsinki and were approved by the Ethics Committee of the Comer Children's Hospital (#11-0268-AM017, #09-115-B-AM031, and #IRB14-1241). All children presented common symptoms of the disease such as snore, respiratory pauses during sleep, awakenings, or daytime hypersomnolence (Hornero et al., 2017). They were referred to the hospital's sleep laboratory between the years 2012 and 2014. Sleep studies were performed with Type 1 PSG

equipment (Polysmith®, Nihon Kohden America Inc., Irvine, CA, USA), including up to 32 neuronal, cardiorespiratory and other physiological signals. Subjects were evaluated following the AASM rules to identify apneas and hypopneas and derive the AHI described in Berry et al. (2012). Accordingly, they were classified in the four common OSA severity groups: no OSA, mild, moderate, and severe OSA (see Section 1.4) (Hornero et al., 2017). Of the included subjects, 803 (82.44%) had any level of OSA (see Table 3.1).

The sociodemographic and clinical data extracted from the UofC database are shown in Table 3.1. The 974 subjects were randomly split into independent training (584) and test (390) sets, ensuring that no statistically significant differences were present between them ($p \geq 0.01$, Mann-Whitney U test for numeric variables, Chi-square test for categorical variables) with regard to age, body mass index (BMI), sex, AHI and OSA severity. This database was used in the three studies of the present doctoral thesis (Jiménez-García et al., 2020, 2022, 2024).

The second source of PSG data used in this thesis is the public database of the Childhood Adenotonsillectomy Trial (CHAT), provided by the National Sleep Research Resource (NSRR). This is a multicenter study that involved several sleep centers across the United States of America (Children's Hospital of Pennsylvania, Philadelphia, PA; Cincinnati Children's Medical Center, Cincinnati, OH; Rainbow Babies and Children's Hospital, Cleveland, OH; Boston Children's Hospital, Boston, MA; Cardinal Glennon Children's Hospital, St. Louis, MI; Montefiore Medical Center, Bronx, NY) (Marcus et al., 2013; Redline et al., 2011). The Number of Clinical Trial of this dataset is NCT00560859. The CHAT study was aimed at assessing the effectiveness of the reference surgical treatment (adeno-

Table 3.1: University of Chicago (UofC) database: sociodemographic and clinical data obtained from the pediatric sleep studies.

	All	**Training set**	**Test set**
Subjects (n)	974	584 (59.96%)	390 (40.04%)
Age (years)	6.0 [3.0; 8.0]	6.0 [3.0; 8.0]	5.5 [3.0; 9.0]
Males (n)	599 (61.50%)	346 (59.25%)	253 (64.87%)
Females (n)	375 (38.50%)	238 (40.75%)	137 (35.13%)
BMI (kg/m^2)	18.02 [16.04; 22.04]	17.72 [16.05; 22.65]	18.18 [16.01; 21.06]
BMI z-score	-0.22 [-0.60; 0.37]	-0.24 [-0.61; 0.43]	-0.17 [-0.58; 0.27]
AHI (e/h)	3.80 [1.53; 9.35]	4.08 [1.71; 10.00]	3.30 [1.40; 7.87]
No OSA (n)	171 (17.56%)	96 (16.44%)	75 (19.23%)
Mild OSA (n)	398 (40.86%)	229 (39.21%)	169 (43.33%)
Moderate OSA (n)	176 (18.07%)	113 (19.35%)	63 (16.15%)
Severe OSA (n)	229 (23.51%)	146 (25.00%)	83 (21.28%)

Data are presented as median [interquartile range] or number (%). BMI: body mass index, AHI: apnea-hypopnea index, OSA: obstructive sleep apnea.

tonsillectomy) against a strategy of watchful waiting and treatment of symptoms (Marcus et al., 2013; Redline et al., 2011). The database consists of 1,638 PSG studies performed to children between 5 and 10 years old who were referred to the sleep units due to OSA suspicion. The studies were distributed in the groups according to their inclusion in the randomization study and the consequent tracking (Marcus et al., 2013; Redline et al., 2011):

- Baseline (453 subjects): children who met the inclusion criteria and were included in the randomization study.
- Follow-up (406 subjects): children belonging to the baseline group who were re-evaluated after 7 months.
- Non-randomized (779): children who were not included in the randomization study after the first evaluation.

All PSGs were analyzed and apneic events were scored to derive the AHI according to the same standardized rules (Redline et al., 2011). More details of the protocol can be found in the supplementary material of Marcus et al. (2013).

The sociodemographic and clinical data extracted from the CHAT database are shown in Table 3.2. The common thresholds introduced in Section 1.4 were used to categorize children into the four OSA severity levels (no OSA, mild, moderate, and severe), with 1,283 out of 1,638 (78,33%) having at least mild OSA (Table 3.2). Similarly to the UofC database, the subjects were randomly split into training (1006) validation (326), and test (306) sets, with no statistically significant differences between groups ($p \geq 0.01$, Mann-Whitney U test). The CHAT database was used in the studies of the present doctoral thesis involving the development of DL architectures (Jiménez-García et al., 2022, 2024).

3.1.1 Airflow signals

AF signals were recorded as part of the PSG with an oronasal thermistor in both databases. The PVDF sensor Dymedix™was used to obtain the AF recordings included in the CHAT database, whereas the thermal sensor provided with the Polysmith® PSG equipment was used in the UofC database. The sampling frequencies of overnight AF signals were different for each dataset, being two in the UofC database (200 and 500 Hz), and ranging between 20 and 512 Hz in the CHAT database. The duration of overnight recordings was checked in both databases to ensure that each of them surpassed 3 hours of total recording time.

Table 3.2: Childhood Adenotonsillectomy Trial (CHAT) database: sociodemographic and clinical data obtained from the pediatric sleep studies.

	All	**Training set**	**Validation set**	**Test set**
Subjects (n)	1,638	1,006 (61.42%)	326 (19.90%)	306 (18.68%)
Age (years)	7.0 [6.0; 8.0]	7.0 [6.0; 8.0]	7.0 [6.0; 8.0]	6.9 [6.0; 8.0]
Males (n)	761 (46.5%)	471 (46.8%)	156 (47.9%)	134 (43.8%)
Females (n)	856 (52.3%)	520 (51.7%)	168 (51.5%)	168 (54.9%)
BMI (kg/m^2)	17.3 [15.5; 21.7]	17.4 [15.6; 21.6]	17.1 [15.4; 21.8]	17.5 [15.7; 21.7]
BMI z-score	-0.24 [-0.65; 0.48]	-0.21 [-0.66; 0.49]	-0.28 [-0.66; 0.46]	-0.26 [-0.60; 0.47]
AHI (e/h)	2.53 [1.16; 5.93]	2.62 [1.14; 5.90]	2.41 [1.21; 5.77]	2.32 [1.14; 6.15]
No OSA (n)	355 (21.67%)	219 (21.77%)	69 (21.17%)	67 (21.90%)
Mild OSA (n)	812 (49.57%)	496 (49.30%)	168 (51.53%)	148 (48.37%)
Moderate OSA (n)	253 (15.45%)	160 (15.90%)	44 (13.50%)	49 (16.01%)
Severe OSA (n)	218 (13.31%)	131 (13.02%)	45 (13.80%)	42 (13.73%)

Data are presented as median [interquartile range] or number (%). BMI: body mass index, AHI: apnea-hypopnea index, OSA: obstructive sleep apnea.

3.1.2 Oximetry signals

SpO_2 signals were obtained from the PSG studies and were recorded with pulse oximeters attached to a finger probe. The built-in pulse oximetry sensor available in the Polysmith® PSG equipment was used in the UofC database, and the Nonin 3012 pulse oximeter with a Nonin 8000J finger sensor was used to record the signals of the CHAT database. The overnight recordings were sampled at rates between 1 and 512 Hz in the CHAT database, and 25, 200 or 500 Hz were set in the UofC database. Similarly to AF, all recordings lasting less than 3 hours were discarded during the composition of both databases.

3.2 Methodology

The methodology conducted during this doctoral thesis is summarized in this section. This research was aimed at developing and testing ensemble and DL algorithms together with XAI capable of detecting pediatric OSA and its severity from AF and SpO_2 signals. The global methodology proposed to achieve the goals reported in Section 2.2 is composed of five stages:

i) A preprocessing stage to resample the AF and SpO_2 signals, reject possible artifacts and normalize their amplitude.

ii) Characterization of the overnight recordings to obtain feature sets that sum-

marize the information of AF and SpO_2.

iii) Feature selection to identify the most useful parameters extracted from AF and SpO_2, discarding redundant or irrelevant features.

iv) Classification of the patients' overnight recordings into OSA severity levels by means of an ensemble learning model.

v) Application of end-to-end DL-based architectures fed with minimally preprocessed (resampling and filtering) AF and SpO_2 signals to derive the AHI of the pediatric subjects. This approach encompasses three sequential steps:

- Segmentation of signals into epochs or sequences before being presented to the DL models.
- Automatic OSA detection and AHI estimation from the overnight signals by means of the DL models.
- Obtaining explanations of the operation of the DL models.

The schematic representation of this global methodology is shown in Figure 3.1. The signals were preprocessed before applying two different approaches to analyze them. As can be seen in the schema, two different paths were followed to solve the problem of detecting pediatric OSA and its severity. The first branch (left) encompasses the typical steps of a classical pattern recognition approach, which consists of three phases: (i) feature extraction, (ii) feature selection, and (iii) feature classification. This approach was deployed in the first article of the compendium (Jiménez-García et al., 2020). The second branch (right) involves the use of DL architectures to derive the AHI directly from the minimally preprocessed signals. Two different DL models were implemented, namely CNN and CNN+RNN. The latter architecture can be viewed as an extension of the CNN model which was further analyzed using the Grad-CAM XAI algorithm. Approaches covering DL models and XAI are described in the second and third articles of the compendium (Jiménez-García et al., 2022, 2024). All the methodologies proposed in this doctoral thesis were evaluated by means of their diagnostic ability, that is, the accuracy and reliability of the automatic models to detect the presence and severity of pediatric OSA. The rest of this section is organized according to the methodological steps introduced previously. Overnight AF and SpO_2 signals preprocessing is described in Section 3.2.1. Next, Sections 3.2.2 and 3.2.3 describe the conventional pipeline developed in the ensemble learning-based approach. The DL architectures and the XAI methods are introduced in Sections 3.2.4 and 3.2.5,

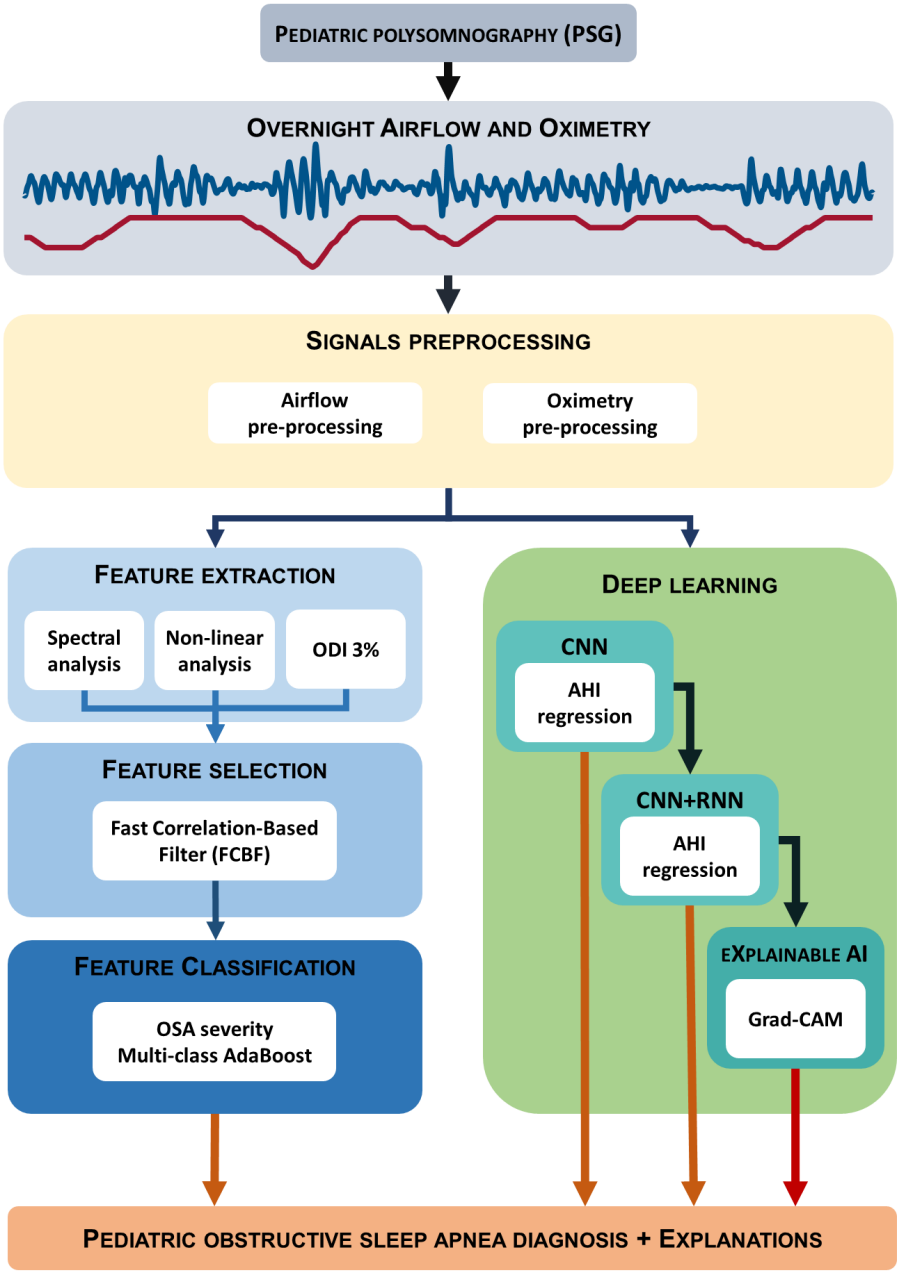


Figure 3.1: Summary of the methodology selected for the doctoral thesis.

respectively. Lastly, Section 3.2.6 gathers the methods utilized to quantitatively assess the results produced by the aforementioned methods.

3.2.1 Preprocessing

PSG-derived AF and SpO_2 signals were minimally preprocessed in order to homogeneize their sampling frequency (f_s), reduce noise, normalize their amplitudes and remove artifacts. It is noteworthy that the studies involving DL approaches were conducted using these preprocessed signals, but without performing artifact removal (Jiménez-García et al., 2022, 2024).

3.2.1.1 Airflow preprocessing

The AASM guidelines to score sleep studies include recommendations for the sampling frequencies of PSG-derived signals. The recommended AF f_s according to these rules is 100 Hz (Iber et al., 2007). Accordingly, 100 Hz was the f_s used in the first study (Jiménez-García et al., 2020). On the other hand, f_s was lowered to 10 Hz was used in the studies involving DL-based approaches in order to reduce the computational load (Jiménez-García et al., 2022, 2024). After resampling, AF signals were filtered using a low-pass filter between 0 and 1.5 Hz to reduce noise while preserving the respiratory oscillations (Jiménez-García et al., 2020, 2022). Lastly, the amplitude of AF was adaptively normalized according to the algorithm proposed by Várady et al. (2002) in order to obtain preprocessed AF recordings with an homogeneous range of values.

Artifacts were automatically removed from the AF signal by computing the statistical moments of preprocessed AF epochs in a 30 s basis (Jiménez-García et al., 2020). Concretely, epochs with a standard deviation ≤ 0.026 or ≥ 0.550 or having a kurtosis ≤ 1.320 were removed from overnight AF recordings (Barroso-García et al., 2020). Note that this procedure to remove artifacts from the overnight AF recordings was conducted in the study involving a feature-engineering approach exclusively (Jiménez-García et al., 2020).

3.2.1.2 Oximetry preprocessing

According to the AASM rules, the f_s of SpO_2 should be at least 10 Hz, being 25 Hz the recommended frequency (Iber et al., 2007). A $f_s = 25$ Hz was used in the first study to resample the SpO_2 signals (Jiménez-García et al., 2020), whereas 10 Hz was used in subsequent studies (Jiménez-García et al., 2022, 2024). This way, both AF and SpO_2 signals had equal and reduced f_s, which is more suitable to reduce the computational load of DL algorithms without signal information loss. SpO_2 registers were subsequently normalized by subtracting the mean value and dividing by the standard deviation of the entire recordings in order to standardize

them before being presented to the DL algorithms (Jiménez-García et al., 2022, 2024).

Artifacts in the SpO_2 signal were also removed in the first study by removing samples with very low oxygen saturation values (below 50%) and sudden variations in the SpO_2 ($\geq$ 4%/s) (Magalang et al., 2003). Again, this artifact removal stage was only conducted in the study that encompassed a feature-engineering workflow (Jiménez-García et al., 2020), and was omitted in the DL-based approaches (Jiménez-García et al., 2022, 2024).

3.2.2 Feature engineering approach: extraction and selection

3.2.2.1 Feature extraction

The feature extraction stage comprised the computation of time-domain statistics, as well as spectral and nonlinear parameters from the preprocessed AF and SpO_2 signals. First-to-fourth statistic moments in the time domain (M1T-M4T), together with the median (MedT) were the statistical features, while CTM, Lempel-Ziv Complexity (LZC), and Sample Entropy (SampEn) constituted the nonlinear parameters (Jiménez-García et al., 2020). CTM was used as a measure of the variability of the signals, which is based on first order differences (Cohen et al., 1996). Given the signal $x(n)$, a 2D plot is constructed with the differences between consecutive samples ($x(n+2)-x(n+1)$ vs. $x(n+1)-x(n)$), and CTM is defined as the rate of points inside a circular region of radius r around the coordinates origin (i.e., their distance to (0, 0) is lower than r) (Cohen et al., 1996):

$$CTM = \frac{\sum_{n=1}^{N-2} \delta(n)}{N-2}, \tag{3.1}$$

$$\delta(n) = \begin{cases} 1, & \text{if } \sqrt{(x(n+2)-x(n+1))^2 + (x(n+1)-x(n))^2} \leq r \\ 0, & \text{otherwise} \end{cases}. \tag{3.2}$$

The computation of CTM depends on the value of r, which was set for AF and SpO_2 independently. To do so, the r parameter value that maximized the magnitude of the Spearman's correlation coefficient ($|\rho|$) between CTM and the subjects AHI was selected (Barroso-García et al., 2017; Jiménez-García et al., 2020).

LZC was computed to measure the complexity of AF and SpO_2 recordings. This nonlinear parameter was computed by counting the number of subsequences contained within the signals (Lempel and Ziv, 1976). Given a time series of length

n, it was first binarized by using the median value as a threshold. Then, $c(n)$ accounts the number of different subsequences in the signal and LZC complexity is the division of $c(n)$ with $b(n)$, which is defined as the theoretical maximum value of $c(n)$ (Lempel and Ziv, 1976):

$$LZC = \frac{c(n)}{b(n)}, b(n) = \frac{n}{\log_2(n)} \tag{3.3}$$

SampEn is a nonlinear parameter aimed at measuring time series irregularity, as it assigns larger values to signals with higher entropy (Richman and Moorman, 2000). SampEn was computed from the conditional probability of two similar sequences of length m remained at a distance lower than r after the length of that sequences increase to $m+1$ (Richman and Moorman, 2000):

$$SampEn = -\log \frac{A^m(r)}{B^m(r)}, \tag{3.4}$$

in which $B^m(r)$ account the number of similar sequences of length m and $A^m(r)$ is the number of sequences that remained similar after increasing the lenght to $m+1$ (Richman and Moorman, 2000). Note that m and r values need to be fixed prior to compute SampEn. Thus, these parameters were optimized by searching m and r values that maximized the absolute value of Spearman's $|\rho|$ between SampEn and AHI (Jiménez-García et al., 2020).

The spectral analysis of the signals was conducted by computing the Power Spectral Density (PSD) by means of the Welch's method (Welch, 1967). Overnight AF and SpO_2 were segmented into epochs of 2^{16} and 2^{14} samples, respectively, with 50% overlap and using a Hamming window, and the periodogram of each segment was computed by means of the squared magnitude of the Discrete Fourier Transform (DFT) (Barroso-García et al., 2017; Jiménez-García et al., 2020). The estimates of AF and SpO_2 PSDs were calculated by averaging the DFT-derived periodograms of all segments in the respective signals (Welch, 1967).

Seven features were extracted from the pediatric OSA-related band of interest of each signal (0.134−0.176 Hz for AF, 0.020−0.044 Hz for SpO_2) (Jiménez-García et al., 2020). These 7 features in the frequency domain were: first-to-fourth statistic moments (M1F-M4F), median (MedF), maximum (MaxF), and minimum (MinF). Furthermore, the median frequency (FreqM) and three spectral entropies ($SpecEn^{1,2,3}$) were also computed from the frequency spectrum of both signals. FreqM is defined as the frequency that splits $PSD(f)$ into two regions with equal power.

$$\sum_{f=0}^{f=FreqM} PSD(f) = \frac{1}{2} \sum_{f=0}^{f=f_s/2} PSD(f). \tag{3.5}$$

SpecEn is an indirect measure of signals irregularity based on Shannon's entropy. It quantifies PSD uniformity since signals with a flatter spectrum (i.e., without dominant frequencies) are associated to high irregularity. SpecEn was computed as the Shannon entropy of the frequency distribution derived from the PSD.

$$SpecEn^i = -\frac{1}{log(M)} \sum_{f=0}^{f_s/2} PSD_n^i(f) \cdot \log[PSD_n^i(f)], i = 1, 2, 3, \tag{3.6}$$

where i is the SpecEn order, M is the PSD length, and PSD_n is the normalized PSD (i.e., the PSD divided by the total PSD power) (Jiménez-García et al., 2020):

$$PSD_n^i(f) = \frac{|PSD(f)|^i}{\sum_{f=0}^{f_s/2} |PSD(f)|^i}. \tag{3.7}$$

The set of extracted features was completed with the computation of ODI 3% from SpO_2 (Jiménez-García et al., 2020). This oximetric index is defined as the rate of SpO_2 drops $\geq$ 3% with respect to the previous baseline per hour of recording (Magalang et al., 2003).

3.2.2.2 Feature selection

Up to 39 features from both signals were extracted and distributed according to the information source to perform feature selection and subsequent classification. In order to test the complementarity of AF and SpO_2 information, as well as the computed features with ODI 3%, six subsets combining AF, SpO_2 and ODI 3% features were subsequently evaluated (namely AF, SpO_2, AF+SpO_2, AF+ODI 3%, SpO_2+ODI 3%, AF+SpO_2+ODI 3%) (Jiménez-García et al., 2020). The complementarity between AF and SpO_2, as well as the identification of relevant and non-redundant features from these signals to detect pediatric OSA were evaluated by means of the Fast Correlation-Based Filter (FCBF) feature selection algorithm (Yu and Liu, 2004). This is a filter method (i.e., it does not depend on any ML algorithm), so it can be combined with any ML model, including ensembles (Saeys et al., 2007). FCBF utilizes the symmetrical uncertainty (SU), a Shannon entropy-based measure of correlation between two variables, to compute relevance and redundancy. It is defined as (Yu and Liu, 2004):

$$SU(X_i|X_j) = 2 \cdot \frac{H(X_i) - H(X_i|X_j)}{H(X_i) + H(X_j)}, \tag{3.8}$$

where $H(X_i)$ and $H(X_j)$ are the Shannon entropy of the features X_i and X_j:

$$H(X_i) = -\sum\nolimits_{x_i \in X_i} p(x_i) \cdot log(p(x_i)), \tag{3.9}$$

$$H(X_j) = -\sum\nolimits_{x_j \in X_j} p(x_j) \cdot log(p(x_j)); \tag{3.10}$$

$H(X_i|X_j)$ is the Shannon entropy of feature X_i after the observation of X_j:

$$H(X_i|X_j) = -\sum\nolimits_{x_j \in X_j} p(x_j) \cdot \sum\nolimits_{x_i \in X_i} p(x_i|x_j) \cdot log(p(x_i|x_j)). \tag{3.11}$$

This way, relevance and redundancy are defined as the SU between a feature and the AHI ($SU(X_i|AHI)$), and the SU between two different features($SU(X_i|X_j)$), respectively. The FCBF algorithm consists of two steps: first, it ranks the features according to their relevance from highest to lowest $SU(X_i|AHI)$; then, it discards features X_i that have a greater SU with a more relevant feature ($SU(X_i|X_j)$) than their own relevance ($SU(X_i|AHI)$):

$$SU(X_j|AHI) \geq SU(X_i|AHI), \; and \; SU(X_i|X_j) \geq SU(X_i|AHI). \tag{3.12}$$

In order to increase the robustness of the feature selection stage, a bootstrapping procedure was implemented in conjunction with FCBF (Guyon and Elisseeff, 2003). A total of 1000 bootstrap-derived datasets were obtained by randomly sampling the training data with replacement (i.e., some examples may be picked more than once in a bootstrap replicate, while others are not selected). This way, all bootstrap replicates have the same size as the training set, but each of them contain diverse subsets of the original data (Witten et al., 2011). FCBF was applied to 1000 bootstraps, and features selected at least in 500 iterations (50% of total) formed the subset of selected features (Barroso-García et al., 2021a; Hornero et al., 2017; Jiménez-García et al., 2020; Vaquerizo-Villar et al., 2018a).

3.2.3 Feature classification: AdaBoost ensemble learning

The feature classification stage was aimed at classifying the features selected by the FCBF algorithm in the four levels of childhood OSA. This is the final stage of the feature-based pattern recognition approach presented in the first article of the compendium (Jiménez-García et al., 2020). This study was focused on comparing

the information of AF and SpO_2, and assessing their diagnostic ability, both separately and jointly. The latter goal was addressed deploying AdaBoost classifiers. These models are based on boosting, a branch of ensemble learning characterized by training simple base classifiers sequentially with differently weighted versions of the dataset examples (Freund and Schapire, 1997; Kuncheva, 2014). In this work, the AdaBoost.M2 variant was implemented since it allows multi-class classification of OSA severity.

AdaBoost training is an iterative process in which, at each iteration, the boosting algorithm gives more importance to the examples that previous base classifiers failed, and a new base classifier is trained according to the updated, importance-weighted data (Freund and Schapire, 1997; Kuncheva, 2014). LDA classifiers were used in this work as weak, base learners, like in previous studies (Gutiérrez-Tobal et al., 2019). The number of iterations (L) is a hyperparameter to be set in order to maximize model's performance. Given a training dataset of N labeled examples $(\mathbf{x_i}, y_i)$, $(i = 1, ..., N)$, and the distribution at iteration t $(t = 1, ..., L)$ $D_t(i)$ (Freund and Schapire, 1997):

$$D_t(i) = \frac{W_i^t}{\sum_{i=1}^{N} W_i^t}, \tag{3.13}$$

with W_i^t:

$$W_i^t = \sum_{y \neq y_i} w_{i,y}^t. \tag{3.14}$$

The weights $w_{i,y}^t$ are equal in all examples in the first iteration. Then, a LDA classifier is trained using the calculated distribution $D_t(i)$ and subsequently evaluated. The pseudo-loss ϵ_t associated to the LDA-derived prediction $h_t(\mathbf{x}, y)$ is calculated as (Freund and Schapire, 1997):

$$\epsilon_t = \frac{1}{2} \sum_{i=1}^{N} D_t(i)[1 - h_t(x_i, y_i) + \sum_{y \neq y_i} \frac{w_{i,y}^t}{W_i^t} h_t(\mathbf{x}, y)]. \tag{3.15}$$

From this pseudo-loss, the modified weight update coefficient at iteration t is defined as β_t. This coefficient includes a regularization term by means of a learning rate hyperparameter ν $(0 < \nu \leq 1)$, with $\nu = 1$ meaning no regularization:

$$\beta_t = \left(\frac{\varepsilon_t}{1 - \varepsilon_t} \right)^{\nu}. \tag{3.16}$$

The weights for iteration $t+1$ are calculated as (Freund and Schapire, 1997):

$$w_{i,y}^{t+1} = w_{i,y}^{t} \cdot \beta_t^{\frac{1}{2}(1+h_t(\mathbf{x_i},y_i)-h_t(\mathbf{x_i},y))} \tag{3.17}$$

AdaBoost algorithm generates a prediction by calculating the weighted vote (Freund and Schapire, 1997):

$$H(\mathbf{x}) = \arg\max_{y} \sum_{t=1}^{L} \log(\frac{1}{\beta_t}) \cdot h_t(\mathbf{x}, y) \tag{3.18}$$

This way, AdaBoost model assigns each subject's descriptive pattern to a level of OSA severity by means of the weighted majority vote of a large population of sequentially trained weak LDA classifiers (Freund and Schapire, 1997).

3.2.4 Deep learning: convolutional and recurrent neural networks

Two end-to-end approaches based on DL to analyze AF and SpO_2 were proposed in the second and third articles of the compendium (Jiménez-García et al., 2022, 2024). As in many other applications, DL approaches have outperformed classic ML paradigms, especially in the contexts of biomedical signal analysis and OSA detection (Faust et al., 2018; Lecun et al., 2015; Mostafa et al., 2019). A previous 1D CNN architecture showed its suitability to detect pediatic OSA from single-channel SpO_2 (Vaquerizo-Villar et al., 2021). In this work, dual-channel approaches were implemented to allow the joint analysis of AF and SpO_2 and derive the AHI. First, a 2D CNN architecture was deployed to process 5-min (300 s) epochs of these two signals jointly (sampled at 10 Hz, 3,000x2 samples) (Jiménez-García et al., 2022). It consisted of a stack of convolutional blocks of five layers: 2D convolution, Batch normalization, Rectified Linear Unit (ReLU), Max pooling, and Dropout (Jiménez-García et al., 2022; Vaquerizo-Villar et al., 2021). Each 2D convolutional layer ($i = 1, ..., NL$) generates NF feature maps $x_i^j(m,n)$ ($j = 1, ..., NF$) by applying convolution operations to the input signal epochs (first layer, $a_1(m,n)$), or the output feature maps generated in previous layers ($a_i(m,n)$, $i = 2, ..., NL$) (Goodfellow et al., 2016):

$$x_i^j(m,n) = \sum_{k=1}^{K} \sum_{l=1}^{2} w_i^j(k,l) \cdot a_i(m-k+1, n-l+1) + b_i^j, \tag{3.19}$$

where K is the length of the 2D filters of kernel size $(K, 2)$, with learned weights w_i^j and bias b_i^j. Then, the feature maps are normalized, and the ReLU activation function is applied (Goodfellow et al., 2016):

$$ReLU(x_i^j) = \max(0, x_i^j). \tag{3.20}$$

After this activation, feature maps size is reduced in the Max pooling layer by retaining the local maximum responses (Goodfellow et al., 2016). The final step of each block is a dropout operation that randomly sets to zero a percentage of the feature map's samples during training (Goodfellow et al., 2016). In this work, the dropout rate was fixed to 0.1 heuristically. After the convolutional blocks, a flattening layer reshapes the multidimensional feature maps to one-dimensional vectors, and a fully connected layer derive a prediction of the number of apnea/hypopnea events per epoch (Jiménez-García et al., 2022).

The second dual-channel approach combined the aforementioned CNN with a RNN (CNN+RNN) to process 30-min sequences of AF and SpO_2 data split into six 5-min epochs (Jiménez-García et al., 2024). Figure 3.2 shows the developed CNN+RNN architecture, in which the aforementioned CNN layers were inserted into consecutive time distributed (TD) layers that process the 5-min epochs. This way, the stack of TD layers produced a sequence of feature maps, similarly to the

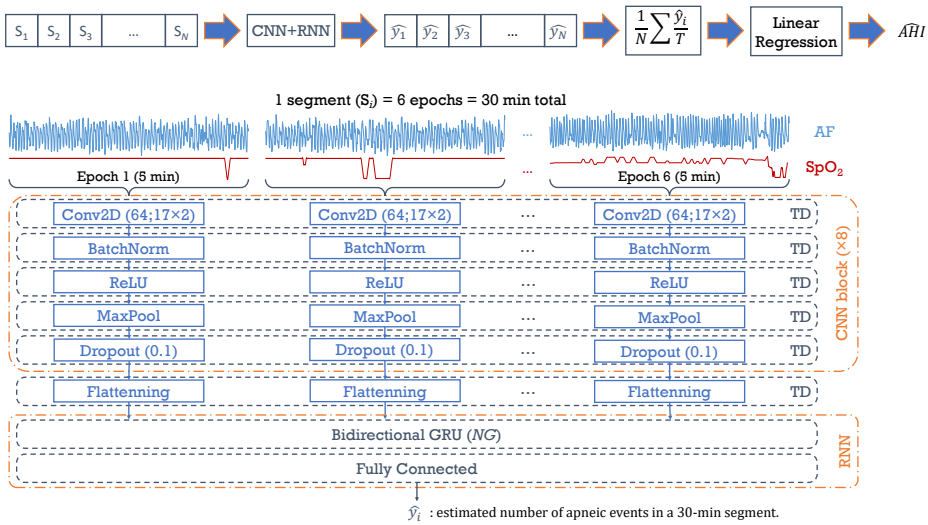


Figure 3.2: Convolutional and recurrent neural network (CNN+RNN) architecture. Adapted from Jiménez-García et al. (2024) with permission (CC-BY-NC-ND 4.0).

previous CNN approach, which is processed in the RNN (Jiménez-García et al., 2024; Korkalainen et al., 2020a). A bidirectional gated recurrent unit (Bi-GRU) layer of size NG ($NG = 1, 2, 4, ..., 64$) was implemented to analyze the temporal dependencies of the generated sequences and derive an estimate of the number of events per segment (Jiménez-García et al., 2024). The number of units in the Bi-GRU layer was a hyperparameter to be optimized, while recurrent dropout and dropout rates were fixed heuristically to 0.1 after preliminary assessments (Jiménez-García et al., 2024).

3.2.5 Explainable artificial intelligence: Gradient-weighted Class Activation Mapping (Grad-CAM)

The Grad-CAM XAI method algorithm was implemented to address the explainability of the developed CNN+RNN architecture, as well as to analyze the behavior of this DL architecture (Jiménez-García et al., 2024). Grad-CAM was used to obtain visual *post hoc* explanations of the model output that indicate zones in the input signals that the model identifies as relevant to make their predictions (Selvaraju et al., 2017). These visual explanations highlight parts of the input signals and are interpreted as heatmaps. In order to obtain explanations from an input example and its corresponding output, Grad-CAM computes the gradient of the model output ($\hat{y}$) with respect to the feature maps $x^j(m, n)$ generated in the last convolutional layer (Selvaraju et al., 2017):

$$g(m, n) = \frac{1}{NF} \sum_{j=1}^{NF} \frac{\partial \hat{y}}{\partial x_{NL}^j} \tag{3.21}$$

Then, it derives an explanatory heatmap through a ReLU activation applied to the dot product of the gradient and the feature map generated in the last layer (Selvaraju et al., 2017):

$$L_{GradCAM}(m, n) = ReLU(g \cdot x_{NL}). \tag{3.22}$$

In this work, the gradients and feature maps were computed in all convolutional layers to derive heatmaps from each of them. Since the generated heatmaps have different length in each convolutional layer, they were all resized to the input segment size and averaged. This way, the obtained heatmaps highlight both coarse and fine grained patterns that the model links to the presence of apneic events and/or their corresponding desaturations, since they have a large influence in the

model's prediction. This procedure to consider coarse and fine grained explanatory heatmaps was heuristically compared with the Guided Grad-CAM approach, which produced poorer results. Therefore, the user is able to understand what parts of the signal are important to detect apneas/hypopneas according to the model (Jiménez-García et al., 2024). When the automatic method accurately detects apneas in the input pattern, the heatmaps can be useful to identify them and verify that the model is correctly detecting OSA. On the contrary, when the prediction is inaccurate, Grad-CAM can highlight irrelevant patterns or miss the apneic events. In the first case, the trustworthiness of the model increases because the user can verify that the model is detecting apneas/hypopneas as expected and discover the useful patterns detected by the DL model. On the other hand, the user can discard patterns not linked to apneic events but highlighted as important by the model in order to supervise the algorithm's behavior.

3.2.6 Statistical analyses

3.2.6.1 Statistical tests and correlation

As described in Section 3.1, statistical tests were employed to ensure the absence of statistically significant differences in sociodemographical and clinical data between training, validation, and test datasets. Mann-Whitney U test and Chi-Square test were used to evaluate these differences according to the variable type to be assessed (numerical or categorical, respectively). Regarding the analysis of feature extraction techniques (Section 3.2.2.1), the extracted descriptors were assessed by means of the Kruskal-Wallis test in order to identify statistically significant differences among the 4 OSA severity groups. This non-parametric test was used because not all extracted features passed the Lilliefors normality test (Jiménez-García et al., 2020).

Further analysis of the features extracted in the first study was conducted by means of the Spearman's correlation (see Section 3.2.2.1) (Jiménez-García et al., 2020). This analysis was applied in the training set to find the optimum parameters of CTM (r) and SampEn (m, r), as well as to assess the positive or negative relationship of each extracted feature with AHI (Jiménez-García et al., 2020).

3.2.6.2 Validation strategies

Different validation strategies were used in this research to check the validity of the methods and the results obtained. As mentioned in Section 3.1, the databases included in this study were split whether into training and test sets (UofC), or

into training, validation and test sets (CHAT). This way, the training stage of all ensemble and DL algorithms deployed in this doctoral thesis was conducted in the training set of these databases. Once the final, optimized models were obtained, they were evaluated in the test sets of both databases. The approaches introduced below have been used in this doctoral thesis to validate the different architectures.

- **Hold-out**. As mentioned above, UofC and CHAT databases were split into two or three independent sets (training/test or training/validation/test). This way, the training set was used to deploy the models, the validation set was used to optimize them (e.g., by finding the optimum model configurations and/or hyperparameters), and the test set was used to assess their performance in unseen data (Witten et al., 2011). This validation approach is the most straightforward, but requires large databases to ensure a minimum amount of training data, which usually comprises a large proportion of the entire dataset (Witten et al., 2011). Moreover, the division should be conducted ensuring the similarity of data among the sets. The subjects of both databases were randomly split into these 2 or 3 groups, and statistical tests were used to evaluate possible differences in sociodemographic and clinical variables (see Section 3.1). In this research, model deployment encompassed parameters optimization within feature extraction techniques, feature selection, and ensemble and DL models training and optimization. When the database size is not large enough to make up to three partitions, other cross validation methods may be used to validate data-driven algorithms.

- **Bootstrapping**. A bootstrap method can be useful when the dataset size is small (Witten et al., 2011). As mentioned in Section 3.2.2.2, bootstrapping was implemented to increase the robustness of the FCBF-based feature extraction stage (Guyon and Elisseeff, 2003). Bootstrap uses random sampling with replacement and equal selection probability of the original set instances (N examples) to create bootstrap replicates of size N, in which some instances can be repeated whereas others are not included (Witten et al., 2011). The procedure is repeated M times in order to obtain M diverse bootstrap replicates and use them to train the models independently. Then, these trained algorithms are evaluated using both the samples included in the training replicate and those which were not selected in that iteration as a test set. The 0.632 bootstrap procedure then estimates the model performance (S) by means of the training and test figures of merit obtained in each bootstrap iteration (S^i_{train} and S^i_{test}, respectively) (Witten

et al., 2011):

$$S = \frac{1}{M}\sum_{i=1}^{M} 0.368 \cdot S_{train_i} + 0.632 \cdot S_{test_i}. \tag{3.23}$$

This validation method has been used in the first study of this doctoral thesis to optimize AdaBoost model's hyperparameters (L, ν) with $M = 1000$ iterations (Jiménez-García et al., 2020).

- **Stratified K-fold cross validation**. This is one of the most common validation methods (Witten et al., 2011). This procedure randomly splits the dataset into K partitions (folds) of equal size while maintaining the proportion of instances that belong to each class or group (e.g., healthy and diseased, different severity levels, etc.) (Witten et al., 2011). Once the database is split, $K-1$ folds are used to train the model and the remaining one is reserved to test it. Model performance is then evaluated in this test set. This process is conducted K times, each of them with a different fold to test the model. The estimated performance is finally calculated by averaging the performances obtained across the folds (Witten et al., 2011). In this research, the number of folds was set to $K = 10$, which is the most common choice. This validation method was used in the second and third articles of the compendium of publications to fix the optimum values of the CNN and CNN+RNN architectures (Jiménez-García et al., 2022, 2024).

3.2.6.3 Measures of agreement

The agreement between actual and estimated diagnosis of pediatric OSA, either by means of AHI or severity levels, was assessed in this research. As mentioned in the previous sections, AdaBoost models of the first article were aimed at classifying OSA severity in 4 levels: no OSA, mild, moderate, and severe OSA (Jiménez-García et al., 2020). Regarding the DL architectures, these were focused on estimating the total AHI from the overnight signals and categorizing the AHI estimates into the 4 severity levels (Jiménez-García et al., 2022, 2024). Bland-Altman plots were used to graphically represent the difference between actual and predicted AHI against the mean of these two observations (Bland and Altman, 1986). The agreement between actual and estimated AHI the Intraclass Correlation Coefficient (ICC) (Chen and Barnhart, 2008). The OSA severity was derived by thresholding the AHI estimates with the cutoffs introduced in Section 1.4 (1, 5, and 10 e/h). Confusion matrices were computed from the actual and estimated OSA severity,

and the classification into these four levels of the disease was evaluated using the 4-class accuracy (Acc_4) and the Cohen's kappa coefficient (k) (Cohen, 1960; Witten et al., 2011).

3.2.6.4 Diagnostic performance

The approaches covered in this doctoral thesis were analyzed in terms of their diagnostic performance, which was assessed in the three AHI cutoffs. All metrics were derived from the computation of correctly and incorrectly classified subjects. Depending on the actual and predicted diagnosis, 4 possible cases can occur (Fawcett, 2006):

- True negative (TN): when a subject with actual AHI<cutoff is correctly diagnosed as not having a certain degree of OSA (AHI<cutoff).
- False negative (FN): when a diseased subject (AHI≥cutoff) is incorrectly diagnosed as not having a certain degree of OSA (AHI<cutoff).
- True positive (TP): when a diseased subject (AHI≥cutoff) is correctly diagnosed as having a certain degree of OSA (AHI≥cutoff).
- False positive (FP): when a subject with actual AHI<cutoff is incorrectly diagnosed as having a certain degree of OSA (AHI≥cutoff).

According to the number of TN, FN, TP and FP, the following rates were calculated to assess diagnostic performance (Deeks and Altman, 2004; Fawcett, 2006; Witten et al., 2011):

- Sensitivity (Se) is the rate of diseased subjects correctly classified:

$$Se = \frac{TP}{TP + FN}. \tag{3.24}$$

- Specificity (Sp) is the rate of control subjects correctly classified:

$$Sp = \frac{TN}{TN + FP}. \tag{3.25}$$

- Accuracy (Acc) is the rate of subjects (diseased or not) correctly classified:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.26}$$

- Positive predictive value (PPV) is the rate of subjects classified as having a certain degree of OSA correctly classified:

$$PPV = \frac{TP}{TP + FP}. \tag{3.27}$$

- Negative predictive value (NPV) is the rate of subjects classified as not having at least a certain degree of OSA correctly classified:

$$NPV = \frac{TN}{TN + FN}. \tag{3.28}$$

- Positive likelihood ratio ($LR+$) is the ratio of probability of a positive test result (Se vs. the opposite of Sp):

$$LR+ = \frac{Se}{1 - Sp}. \tag{3.29}$$

- Negative likelihood ratio ($LR-$) is the ratio of probability of a negative test result (the opposite of Se vs. Sp):

$$LR- = \frac{1 - Se}{Sp}. \tag{3.30}$$

Chapter 4

Results

This chapter summarizes the results obtained with the different methods developed in this doctoral thesis, which covered from a feature engineering approach using ensemble learning to DL architectures combined with XAI. The results presented in this chapter are organized according to the methodological approaches followed in this work to deploy the predictive models (feature engineering, ensemble learning and DL), the XAI-derived explanatory information and their diagnostic performance. Results of feature extraction and selection stages are presented in Section 4.1, and the optimization results of ensemble and DL models are summarized in Section 4.2. The outcomes of the Grad-CAM XAI algorithm are evaluated in Section 4.3. Since the goal of all methods proposed in this research is to serve as an automatic tool to aid in the diagnosis of pediatric OSA, their diagnostic ability is the most common and important indicator of their usefulness. Consequently, the agreement between PSG-derived and model-estimated OSA severity, as well as the diagnostic performances reached by the models are summarized in Section 4.4. Note that the results introduced in this chapter are directly related to those published in the articles that constitute the compendium of publications (see Chapter 7).

4.1 Feature engineering: extraction and selection

4.1.1 Feature extraction

Among the feature extraction techniques used in this research, CTM and SampEn required parameters optimization. This procedure was aimed at maximizing the

absolute value of Spearman's $|\rho|$ between each feature and AHI for both signals independently, using the training set data from UofC database (Jiménez-García et al., 2020). The optimum r values of CTM were $r = 0.0004$ and $r = 0.0250$ in AF and SpO_2 signals, respectively. Likewise, the maximum $|\rho|$ reached by SampEn was obtained using $m = 2$, $r = 0.05\sigma$ in the AF, and $m = 3$, $r = 0.05\sigma$ in the SpO_2 (Jiménez-García et al., 2020). It is noteworthy that CTM obtained the maximum $\rho = 0.3979$ among all AF-derived features, and other time-domain and spectral parameters such as M4T ($\rho = 0.3580$), M1F ($\rho = 0.3492$), MedF ($\rho = 0.3591$), MinF ($\rho = 0.3588$), and SpecEn ($\rho = 0.3464$) reached significant but slightly lower correlations with the AHI as well. In all these cases, the Kruskal-Wallis test showed statistically significant differences($p \ll 0.01$) among severity groups in the UofC training set (Jiménez-García et al., 2020). ODI 3% was the feature that obtained the highest $\rho = 0.6918$, followed by other nonlinear and spectral parameters such as CTM ($\rho = -0.6187$), M1F ($\rho = 0.6773$), M2F ($\rho = 0.6352$), MedF ($\rho = 0.6753$), MaxF ($\rho = 0.6646$), and MinF ($\rho = 0.6504$). Again, the results of the Kruskal-Wallis test to examine statistically significant differences among severity groups in the UofC training set showed $p \ll 0.01$ in all mentioned SpO_2-derived features.

4.1.2 Feature selection

The feature selection stage conducted in the first study of this doctoral thesis combined FCBF with bootstrap to identify the most useful features (Jiménez-García et al., 2020). These formed the final subset of selected features if they were individually selected by FCBF at least 500 times (50% of total bootstrap iterations). Figure 4.1 shows the features that surpasses this threshold and were finally selected. It is worth to mention that CTM and quadratic SpecEn from the AF signal, as well as CTM and M4F from SpO_2 data were considered relevant and non-redundant by the FCBF-based approach in absence of ODI 3%. However, after including ODI 3% in the analyses, only CTM from AF and M4F from SpO_2 were complementary to it. ODI 3% emerged as the most relevant feature since it improved the diagnostic ability of AdaBoost when this variable was included in the feature sets. Indeed, the combination of all information sources (AF+SpO_2+ODI 3%) reached the highest diagnostic ability in terms of Cohen's kappa and 4-class accuracy (Jiménez-García et al., 2020).

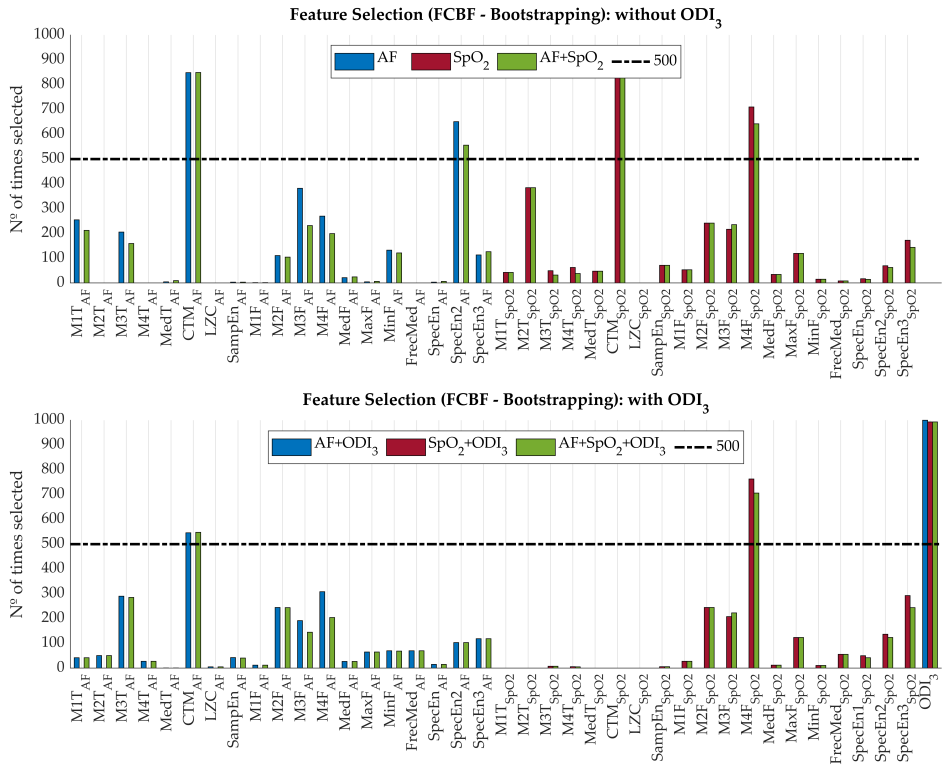


Figure 4.1: Results of the feature selection stage by means of FCBF with bootstrapping: without ODI 3% (top) and with ODI 3% (bottom) (Jiménez-García et al., 2020).

4.2 Ensemble and deep learning models hyperparameter optimization

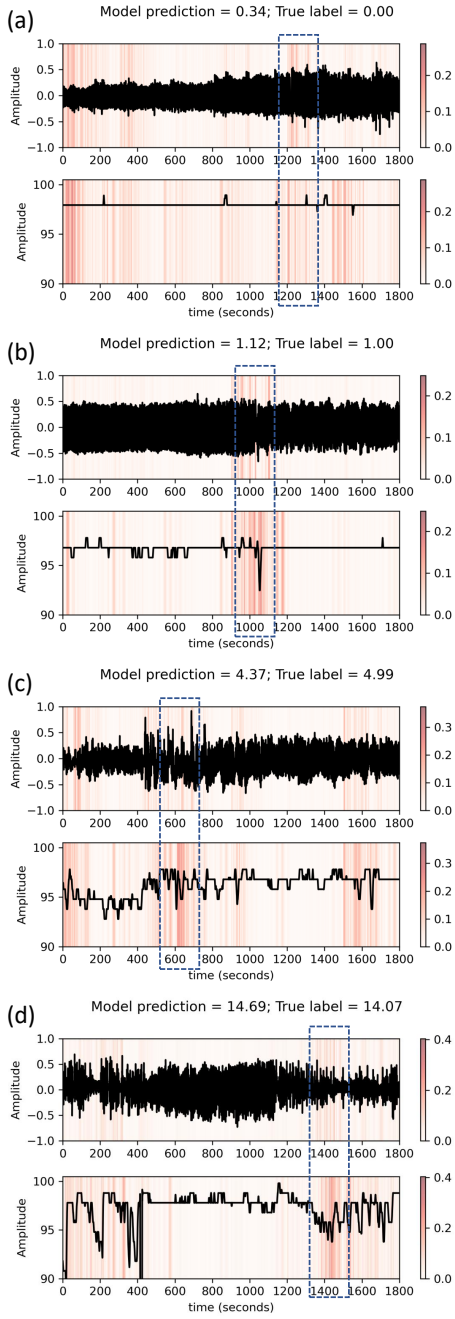
AdaBoost models training required hyperparameters L and ν selection. As mentioned in Section 3.2.3, L and ν optimization was aimed at maximizing Cohen's k in the UofC training set using 0.632 bootstrap validation. Moreover, the procedure was conducted for each feature subset independently (AF, SpO_2, AF+SpO_2, AF+ODI 3%, SpO_2+ODI 3%, AF+SpO_2+ODI 3%). Models trained without ODI 3% generally obtained lower k agreement ($k = 0.1453$ with $L = 3$, $\nu = 0.2$ for AF, $k = 0.2327$ with $L = 700$, $\nu = 0.5$ for SpO_2, $k = 0.2228$ with $L = 8000$, $\nu = 0.1$ for AF+SpO_2) than those including this oximetic variable ($k = 0.2926$ with $L = 500$, $\nu = 0.3$ for AF+ODI 3%, $k = 0.2918$ with $L = 700$, $\nu = 0.2$ for SpO_2+ODI 3%, $k = 0.2909$ with $L = 500$, $\nu = 0.3$ for AF+SpO_2+ODI 3%). It is noteworthy the great similarity between the maximum k values reached by the AdaBoost models

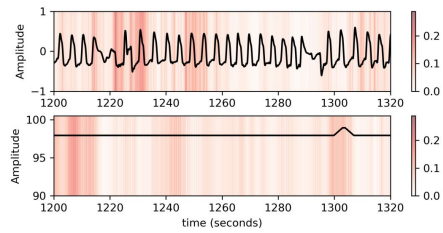
that included ODI 3% among the selected features. In addition, the AF subset reached the lowest maximum k with a small number of base classifiers, while the rest of optimum AdaBoost models were obtained using larger values of L.

Two different DL architectures were developed and tested in this research, a CNN and a combination of CNN+RNN (Jiménez-García et al., 2022, 2024). The optimum CNN architecture consisted of $NL = 8$ convolutional layers with $NF = 64$ 2D filters, each of them with an optimum kernel size ($K = 17, 2$) (Jiménez-García et al., 2022). It is important to note that this optimal set of structural hyperparameteres outperformed all other CNN configurations, reaching $k = 0.4807$ in a validation set formed by data from both UofC and CHAT data. It was also observed that larger values of NL, NF and K resulted in overfitting (i.e., increased performance decay from training to validation sets). This CNN model was reused to implement the CNN+RNN architecture by combining the CNN layers with a sequence analysis framework formed by TD and RNN layers, and following a transfer learning approach. This way, the CNN architecture hyperparameters remained unchanged while the optimization of the CNN+RNN was focused on the RNN. The optimum configuration of the Bi-GRU layer was accomplished by setting $NG = 4$ units, reaching the highest $k = 0.5077$ in the joint validation set using stratified 10-fold cross-validation (Jiménez-García et al., 2024).

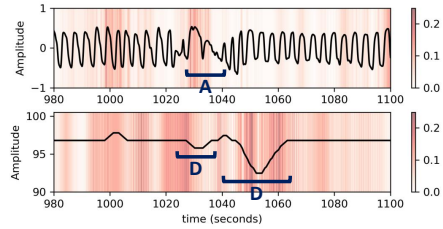
4.3 Model-derived explanations using Grad-CAM

The Grad-CAM algorithm was applied to the CNN+RNN architecture to obtain explanatory heatmaps about the model predictions. Figure 4.2 shows some examples of accurate predictions and their corresponding heatmaps. A zoom on relevant zones in the included segments is also provided to help in the visualization of highlighted patterns. As can be seen, the localization maps show a more reddish color in zones where the DL model detected signs of apneas, hypopneas or desaturations. The heatmap of Figure 4.2(a) point to possibly missed breaths, but the prediction was close to 0. The SpO_2 signal showed constant oxygenation and the heatmap did not highlight specific patterns. The apnea event shown in Figure 4.2(b) (indicated with A) was correctly detected and highlighted both in the AF and the SpO_2 patterns, in which the scored desaturations (D) coincide with stronger Grad-CAM heatmaps. Two consecutive events are correctly localized in Figure 4.2(c), in which two consecutive desaturations and the normal AF

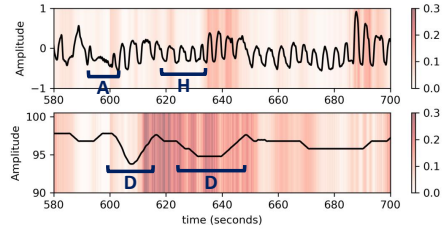




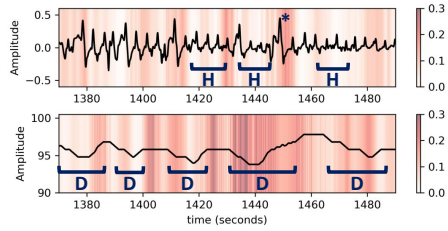
Interpretation: missed breaths within normal respiration, no desaturation.



Interpretation: airflow interruption followed by a desaturation ≥ 3%.



Interpretation: normal breathing after apnea and hypopnea; desaturations ≥ 3%.



Interpretation: irregular AF (hypopneas and arousal *), and mild desaturations.

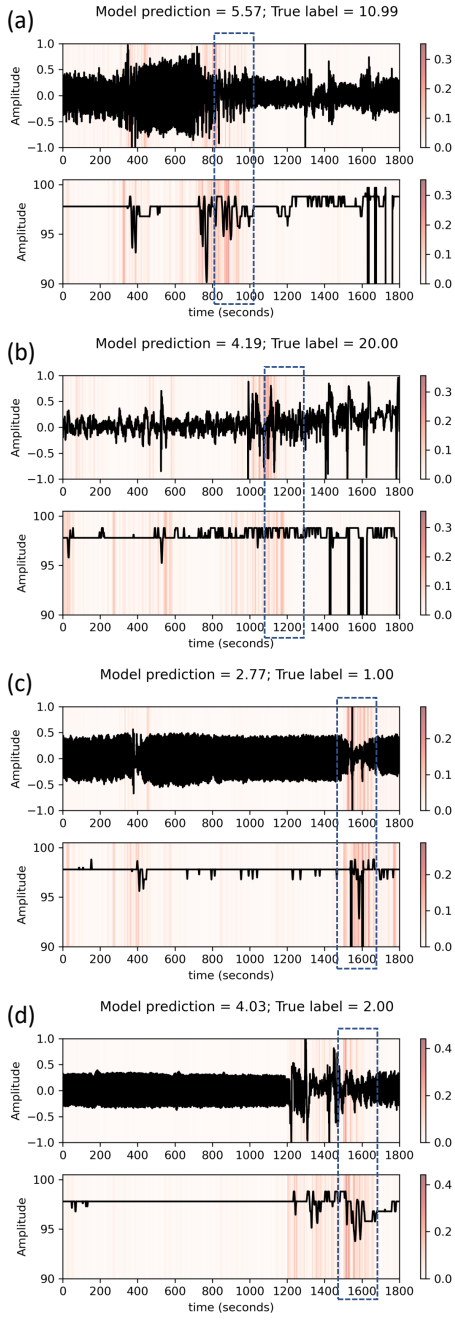
Figure 4.2: Grad-CAM heatmaps obtained from segments accurately predicted by the CNN+RNN architecture (Jiménez-García et al., 2024).

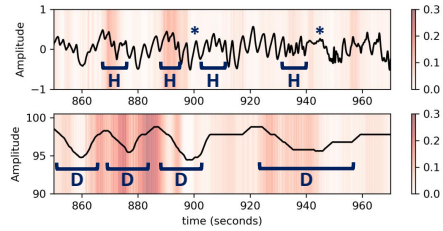
amplitude recovery after an hypopnea (scored with H) are highlighted. In this case, the strong heatmap highlights these desaturations in the SpO_2, while the AF heatmaps also point to this zone, but with lower intensity. Lastly, a segment with recurrent hypopneas and associated desaturations is represented in Figure 4.2(d). SpO_2 drops and recoveries, as well as zones around the hypopneas are indicative of a irregular AF pattern in which hypopneas and arousals (scored with an asterisk) are present. Concretely, the AF heatmap highlights with a stronger color some sudden changes in the respiratory pattern after several breaths with low AF amplitude.

Conversely, heatmaps corresponding to segments in which the CNN+RNN model failed to estimate the amount of apneic events were also observed. Figure 4.3(a) correctly point to several consecutive desaturations, but some hypopneas were not properly highlighted and probably not considered too. In this case, some hypopneas linked to arousals could have been missed by the algorithm, thus explaining why the amount of apneic events was clearly underestimated in this segment. A similar behavior was also observed in Figure 4.3(b), in which the AF heatmap did not highlight several hypopneas associated to arousals again. The fact that the model was unable to detect these consecutive AF cessations, together with the absence of SpO_2 drops in these hypopneas associated to arousals, led to underestimation of the number of apneic events in this segment. The highlighted, flat SpO_2 may suggest that the model did not detect the hypopneas because there were no associated desaturations, like it was observed in the previous example. Strong heatmaps were also observed in Figure 4.3(c), pointing to a zone without any scored apnea or hypopnea. However, the heatmap highlighted artifacts in both signals, and a noisy interval in the AF, and a clear desaturation. Lastly, a noisy interval of the AF signal was highlighted together with various desaturations in Figure 4.3(d). In this case, the DL model detected some apneic events that probably were not scored due to the low quality of the AF signal.

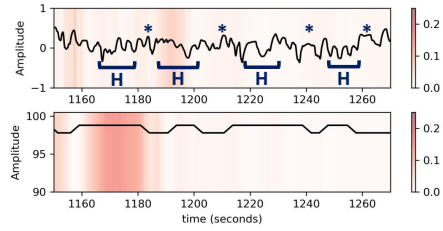
4.4 Diagnostic performance in the test set

The diagnostic performance of the approaches addressed in this thesis was assessed in the test sets of both databases. It is necessary to note that different subgroups of the UofC database were used to extract, analyze and select AF and SpO_2 features, as well as to develop and test the AdaBoost models. On the other hand, different subgroups of both CHAT and UofC databases were involved in the deployment and testing of the DL architectures. The results obtained in the UofC test set

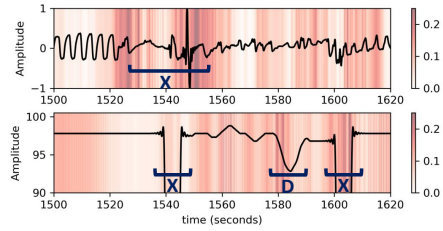




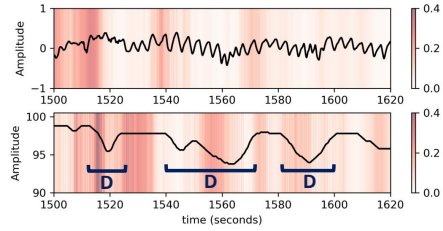
Interpretation: detected hypopneas with desat., missed hypopneas with arousals *.



Interpretation: undetected hypopneas without desaturation linked to arousals *.



Interpretation: artifacts (X) in AF and SpO_2, possibly hiding an apnea/hypopnea.



Interpretation: detected desaturations, possibly unscored apneas/hypopneas.

Figure 4.3: Grad-CAM heatmaps obtained from segments accurately predicted by the CNN+RNN architecture (Jiménez-García et al., 2024).

using the proposed ensemble and DL models can therefore be directly compared, and the results obtained in the CHAT test set are restricted to the DL models comparison.

4.4.1 Ensemble learning: AdaBoost

As mentioned in Section 4.2, 6 independent AdaBoost models were trained with subsets of relevant and non-redundant AF- and SpO_2-derived features (with and without ODI 3%). Among them, the AdaBoost model trained with AF and SpO_2 features, including ODI 3% (AF+SpO_2+ODI 3%), reached the highest $Acc_4 = 57.95\%$ and $k = 0.3984$ in the 4-class OSA severity classification. Nevertheless, it was also observed that the combination of AF features with ODI 3% (AF+ODI 3%) reached the same $Acc_4 = 57.95\%$ with slightly lower $k = 0.3930$ (Jiménez-García et al., 2020). The confusion matrices obtained in the UofC test set from both combinations are shown in Figure 4.4. It can be observed that both models overestimated OSA severity in a great proportion of children with no OSA according to the gold standard. Nevertheless, differences in the proportion of underestimated and overestimated subjects were not high in both models (Jiménez-García et al., 2020).

The minimal differences between the confusion matrices AF+ODI 3% and AF+SpO_2+ODI 3% were also noticeable when assessing the diagnostic ability

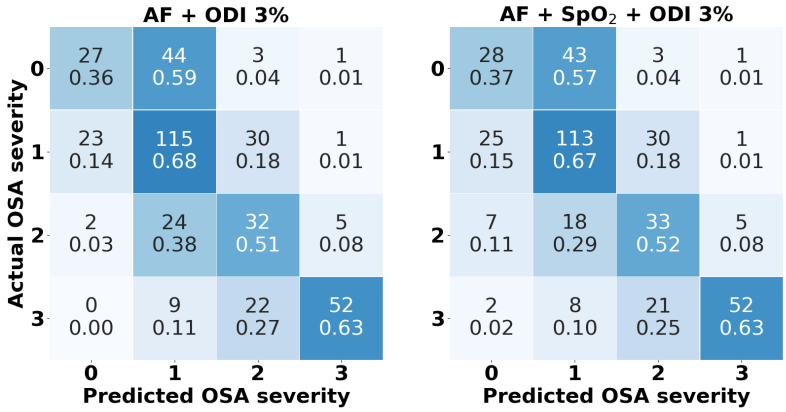


Figure 4.4: Confusion matrices obtained in the test set of the UofC database using the AdaBoost models (Jiménez-García et al., 2020). Results obtained in the AF+ODI 3% subset (left) and the AF+SpO_2+ODI 3% (right).

Table 4.1: Diagnostic performance obtained with the AdaBoost models in the AHI cutoffs 1, 5, and 10 e/h (Jiménez-García et al., 2020).

Subset	Cutoff	Se(%)	Sp(%)	Acc(%)	PPV(%)	NPV(%)	LR+	LR−
AF+ODI	1	92.06	36.00	81.28	85.80	51.92	1.4385	0.2205
	5	76.03	85.66	82.05	76.03	85.66	5.3002	0.2799
	10	62.65	97.72	90.26	88.14	90.63	27.4768	0.3822
AF+SpO_2 +ODI 3%	1	89.21	37.33	79.23	85.67	45.16	1.4235	0.2891
	5	76.03	85.66	82.05	76.03	85.66	5.3002	0.2799
	10	62.65	97.72	90.26	88.14	90.63	27.4768	0.3822

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

of the respective models in the common AHI cutoffs (1, 5, and 10 e/h). Table 4.1 shows the diagnostic ability of AdaBoost models in terms of the figures introduced in Section 3.2.6.4. In this case, it can be highlighted that the AF+ODI 3% subset reached slightly higher $Acc = 81.28\%$ than AF+SpO_2+ODI 3% in the lowest cutoff (1 e/h), as well as the highest $NPV = 51.92\%$. $Se = 92.06\%$ and $PPV = 85.80\%$ were also slightly higher in comparison with AF+SpO_2+ODI 3%, but it was also accompanied by a slightly lower $Sp = 36.00\%$ (Jiménez-García et al., 2020). The diagnostic performance of both AdaBoost models was exactly the same in 5 and 10 e/h, which also highlights the similarities between these two AdaBoost models.

4.4.2 Deep learning: CNN and CNN+RNN

The diagnostic ability of DL-based algorithms was tested in the two databases described in Section 3.1 (CHAT and UofC). The agreement between actual and predicted AHI was high, since the ICCs reached in the test sets by both models ranged between $ICC = 0.8821$ and $ICC = 0.9546$ and were higher in the CHAT test set (Jiménez-García et al., 2022, 2024). Bland-Altman plots are shown in Figures 4.5 (CHAT test set) and 4.6 (UofC test set), in which it can be observed that the standard deviation of the error is greater in the UofC test set. The limits of agreement were noticeably smaller in this dataset using CNN+RNN, which is also consequent with a slightly higher $ICC = 0.9004$ (vs. $ICC = 0.8821$ using CNN). However, this was not observed in the CHAT test set, since these limits were marginally lower using CNN. It is also noteworthy the slight AHI underestimation observed in CHAT, and the AHI overestimation in UofC data (Jiménez-García et al., 2022, 2024). Moreover, this estimation bias was lower in both test datasets using the CNN+RNN architecture, resulting in a more accurate

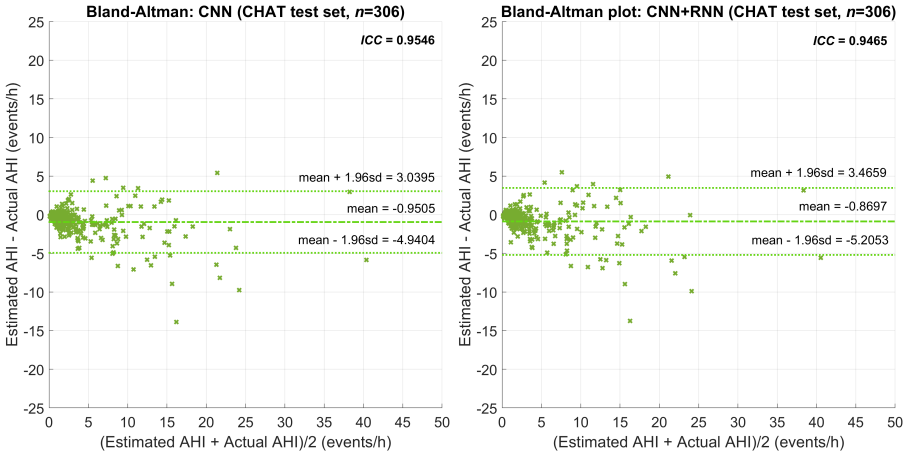


Figure 4.5: Bland-Altman plots of the AHI estimates obtained in the test set of the CHAT database using the DL models (Jiménez-García et al., 2022, 2024). Results obtained using the CNN (left) and the CNN+RNN models (right).

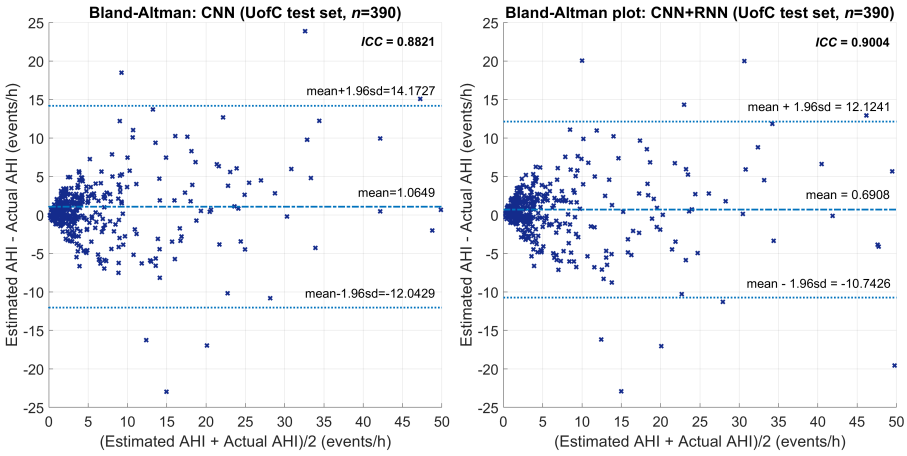


Figure 4.6: Bland-Altman plots of the AHI estimates obtained in the test set of the UofC database using the DL models (Jiménez-García et al., 2022, 2024). Results obtained using the CNN (left) and the CNN+RNN models (right).

algorithm (Jiménez-García et al., 2024).

Regarding the diagnostic agreement of CNN and CNN+RNN algorithms, the 4-class OSA severity classification was evaluated after deriving the four OSA severity levels by using the common AHI thresholds (1, 5, and 10 e/h). Figure 4.7 shows the confusion matrices obtained in the CHAT test set with the two DL architectures

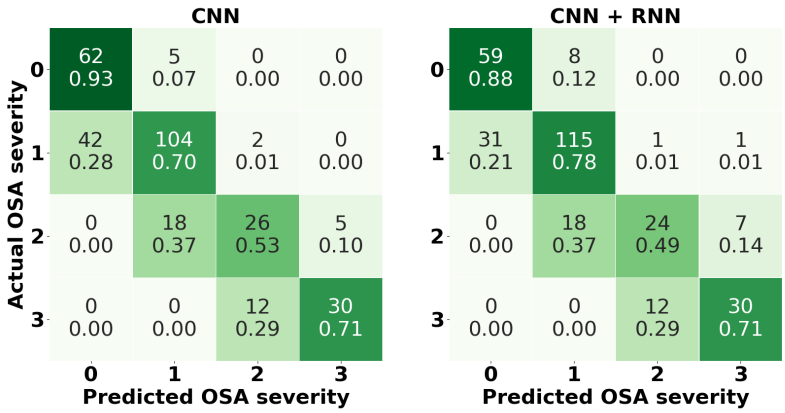


Figure 4.7: Confusion matrices obtained in the test set of the CHAT database using the DL models (Jiménez-García et al., 2022, 2024). Results obtained using the CNN (left) and the CNN+RNN models (right).

(Jiménez-García et al., 2022, 2024). The evolution from a simple CNN approach to a more sophisticated CNN+RNN slightly enhanced the diagnostic ability, obtaining higher $Acc_4 = 74.51\%$ and $k = 0.6231$ in the CHAT test set in comparison with CNN ($Acc_4 = 72.55\%$ and $k = 0.6011$). It can also be seen that the proportion of mild OSA subjects correctly classified improved sensibly. On the other hand, both DL models showed an OSA underestimation tendency given that the proportion of subjects that were classified at least one severity level behind their actual diagnosis is also noticeable. The same enhancement in terms of agreement was observed in the test set of the UofC database (Figure 4.8), where improved $Acc_4 = 62.31\%$ and $k = 0.4495$ were obtained using the CNN+RNN in comparison with the CNN architecture ($Acc_4 = 61.79\%$, and $k = 0.4469$) (Jiménez-García et al., 2022, 2024). Contrarily to what was observed in the CHAT test set, the DL algorithms frequently overestimated OSA severity, especially in subjects without OSA classified as having mild OSA.

The ability of CNN and CNN+RNN architectures to diagnose pediatric OSA in the three children-specific cutoffs is shown in Tables 4.2 (CHAT test set) and 4.3 (UofC test set). Regarding the results in the CHAT test set, the CNN+RNN outperformed the CNN in 1 e/h by reaching $Acc = 87.25\%$, $Se = 87.03\%$, and $NPV = 65.56\%$. However, the CNN model reached higher $Sp = 92.54\%$. Interestingly, both DL models obtained identical diagnostic performance in 5 e/h using the CHAT test dataset. On the other hand, the CNN showed slightly higher

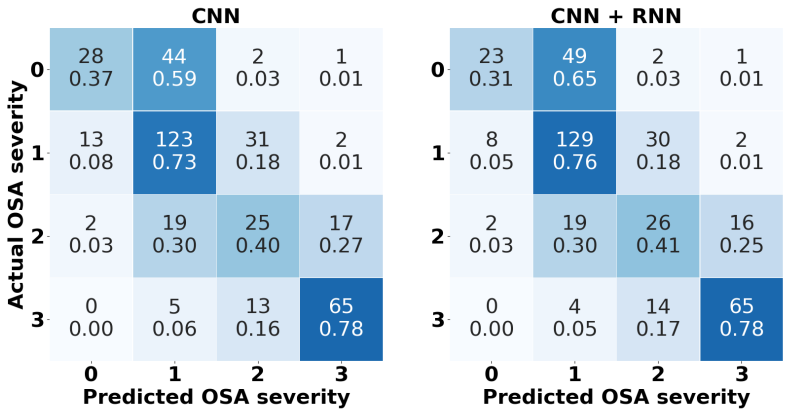


Figure 4.8: Confusion matrices obtained in the test set of the UofC database using the DL models (Jiménez-García et al., 2022, 2024). Results obtained using the CNN (left) and the CNN+RNN models (right).

$Acc = 94.44\%$, $Sp = 98.11\%$, and $PPV = 85.71\%$ in comparison with CNN+RNN in 10 e/h for CHAT data.

The comparison of CNN and CNN+RNN also showed very similar diagnostic metrics in the UofC test set (see Table 4.3), but the CNN+RNN architecture surpassed the CNN in most of them. Both models reached the same $Acc = 84.10\%$ in 1 e/h, but the CNN obtained higher $Sp = 37.33\%$ and $PPV = 86.46\%$, and the CNN+RNN surpassed it in terms of $Se = 96.83\%$ and $NPV = 69.70\%$. In 5 e/h, the CNN+RNN model marginally surpassed the CNN in all considered rates when they were evaluated in UofC test data. Finally, both models reached the same $Se = 78.31\%$ in 10 e/h, but the rest of diagnostic metrics favored the CNN+RNN architecture. However, these differences were not high. It is also worth to mention that the CNN+RNN algorithm reached the lowest $LR- = 0.1035$ in 1 e/h while also showed $LR+ = 12.6538$ in 10 e/h, indicating that this method shows high reliability to: (i) discard any level of OSA, and (ii) to establish a positive diagnosis of severe OSA. As stated at the beginning of this section, AdaBoost and both DL models can be directly compared in the same test set from the UofC database. In this sense, the CNN+RNN algorithm maximized Acc, Se, NPV, and $LR-$ in all cutoffs, while both AdaBoost models reached the highest Sp in 5 and 10 e/h. The most remarkable $Sp = 37.33\%$ in 1 e/h was simultaneously reached by the AdaBoost model trained with AF+SpO_2+ODI 3% and the CNN, but the latter model also obtained nearly optimal $Se = 95.24\%$ and the highest overall

Table 4.2: Diagnostic performance obtained in the CHAT database with the DL architectures for the AHI cutoffs 1, 5, and 10 e/h (Jiménez-García et al., 2022, 2024).

Model	Cutoff	Se(%)	Sp(%)	Acc(%)	PPV(%)	NPV(%)	LR+	LR–
CNN	1	82.43	92.54	84.64	97.52	59.62	11.0452	0.1899
	5	80.22	99.07	93.46	97.33	92.21	86.2363	0.1997
	10	71.43	98.11	94.44	85.71	95.57	37.7143	0.2912
CNN+RNN	1	87.03	88.06	87.25	96.30	65.56	7.2887	0.1473
	5	80.22	99.07	93.46	97.33	92.21	86.2363	0.1997
	10	71.43	96.97	93.46	78.95	95.52	23.5714	0.2946

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

Table 4.3: Diagnostic performance obtained in the UofC database with the DL architectures for the AHI cutoffs 1, 5, and 10 e/h (Jiménez-García et al., 2022, 2024).

Model	Cutoff	Se(%)	Sp(%)	Acc(%)	PPV(%)	NPV(%)	LR+	LR–
CNN	1	95.24	37.33	84.10	86.46	65.12	1.5198	0.1276
	5	82.19	85.25	84.10	76.92	88.89	5.5708	0.2089
	10	78.31	93.49	90.26	76.47	94.10	12.0211	0.2320
CNN+RNN	1	96.83	30.67	84.10	85.43	69.70	1.3965	0.1035
	5	82.88	85.66	84.62	77.56	89.32	5.7777	0.1999
	10	78.31	93.81	90.51	77.38	94.12	12.6538	0.2312

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

$PPV = 86.46\%$ in 1 e/h.

Chapter 5

Discussion

This doctoral thesis was aimed at deploying novel data-driven approaches to aid in the diagnosis of pediatric OSA using the information of AF and SpO_2 such as ensemble and DL models combined with XAI. Therefore, the diagnostic usefulness of the automatic analysis of AF and SpO_2 was assessed using three different approaches: a comparison of several AF- and SpO_2-derived combinations of features (with and without ODI 3%) using a multi-class AdaBoost classifier, a 2D CNN fed with epochs of both signals aimed at predicting the AHI, and an explainable CNN+RNN that not only estimates the presence and severity of OSA, but also provides explanatory localization maps by means of Grad-CAM. Each of these approaches revealed the complementarity of AF and SpO_2 to diagnose pediatric OSA by means of ensemble and DL. The combination of both signals performed better than any of them alone using AdaBoost, and DL methods outperformed other ML and DL methodologies regarded in similar previous studies. Furthermore, the CNN+RNN architecture provided a localization of the most relevant OSA-related patterns found in these signals by including a XAI algorithm. The most important findings of the research works carried out in this doctoral thesis, covering from feature-engineering to explainable DL, are outlined in this chapter. The diagnostic ability of these algorithms is analyzed and subsequently compared with similar studies in the next sections. The last part of this chapter takes into account the main limitations of this thesis. Note that the contents of this chapter outline the findings published in the studies that constitute the compendium of publications, so they have a close relationship with these articles (see Chapter 7).

5.1 Feature engineering: complementarity between AF and SpO_2 information

AF and SpO2 signals were characterized by extracting temporal, spectral, and nonlinear features together with ODI 3% (Jiménez-García et al., 2020). This implied the definition of a band of interest in the AF signal (0.134−0.176 Hz). This band was directly related with the presence and recurrence of apneic events since it covered frequencies around and below the half of the respiratory frequency (0.25−0.35 Hz). The OSA-characteristic AF pauses and cessations have a duration of at least 2 respiratory cycles, which is consistent with an increase of the spectral power in the band found in this study and also in other previous studies (Gutiérrez-Tobal et al., 2015; Jiménez-García et al., 2020). Some features extracted from this pediatric OSA-specific band showed significant correlation with the AHI (e.g, M1F, M2F, MedF, MinF, and MaxF), as well as other spectral features (SpecEn and $SpecEn^2$). Regarding the nonlinear analysis of AF, CTM showed the stongest correlation with AHI among AF-derived features. The utility of some of these features was also observed in a previous study aimed at evaluating the irregularity and variability of AF recordings in pediatric OSA (Barroso-García et al., 2017). With respect to the SpO2 signal, ODI 3% showed the highest correlation with AHI, and many others were strongly related with the latter. The usefulness of this oximetric variable has been highly reported in previous research involving ML approaches in childhood OSA (Barroso-García et al., 2021a,b; Hornero et al., 2017; Vaquerizo-Villar et al., 2018a). Nevertheless, some studies reported high redundancy among these features (Hornero et al., 2017; Vaquerizo-Villar et al., 2018a). Novel AF and SpO_2 characterization methods have been also proposed, but all of them show a strong dependency of ODI 3% to slightly increase their diagnostic performance (Barroso-García et al., 2020; Vaquerizo-Villar et al., 2018a), which reinforced the aforementioned considerations about the usefulness of ODI 3%. In summary, potentially useful OSA-related information may be extracted from both signals by means of spectral and nonlinear analyses with the objective of deploying a feature-based classifier to detect pediatric OSA.

The results of the feature selection stage showed that the same descriptors selected separately from AF and SpO_2 were considered complementary within each other when these sources were joined (AF+SpO_2 set). This was also observed when the ODI 3% was added to the selection phase (AF+ODI 3%, SpO_2+ODI 3% AF+SpO_2+ODI 3% subsets) (Jiménez-García et al., 2020). This oximetric index was the most dominant variable of the study with regard to the FCBF selection

stage, suggesting that ODI 3% is definitely the most useful variable extracted from SpO_2 among feature-engineering approaches (Barroso-García et al., 2020; Gutiérrez-Tobal et al., 2015; Hornero et al., 2017; Jiménez-García et al., 2020; Vaquerizo-Villar et al., 2018a). In addition, CTM was the only AF-derived feature complementary with SpO_2-derived features and ODI 3% simultaneously. This suggests that AF variability, computed by means of CTM, might be complementary to the information of SpO_2, highly summarized in ODI 3% (Jiménez-García et al., 2020). Other studies also supported this finding using other AF descriptors, which reinforces the idea of completing the SpO_2 OSA-related information with AF parameters (Barroso-García et al., 2020, 2021a,b). Accordingly, FCBF-based feature selection indicated the most useful and complementary information jointly from AF and SpO_2, and thus susceptible of pattern recognition tasks.

AdaBoost models were trained with different combinations of relevant and non-redundant features according to the signals from which they were computed, which were evaluated in the test set. This ensemble learning algorithm was also evaluated in the context of adult OSA using AF and SpO_2 separately (Gutierrez-Tobal et al., 2016; Gutiérrez-Tobal et al., 2019). Other studies focused on pediatric OSA also implemented an AdaBoost classifier for pattern recognition (Barroso-García et al., 2021b; Calderón et al., 2020). The top-performing combinations were AF+ODI 3% and AF+SpO_2+ODI 3% (see Figure 4.4), which obtained the same $Acc_4 = 57.95\%$ ($k = 0.3930$ and $k = 0.3984$, respectively). Minimal differences in the confusion matrices revealed slightly higher $Acc = 81.28\%$ in 1 e/h due to increased $NPV = 51.92\%$ and $Se = 92.06\%$ when the AF+ODI 3% set was used, with equal diagnostic performances in 5 and 10 e/h (Table 4.1). This confirms that the highest diagnostic ability is reached by combining ODI 3% with AF-derived CTM. As mentioned in the previous paragraph, the approaches that combined AF with ODI 3% also showed this behavior (Barroso-García et al., 2020, 2021a,b). Similar results were observed in studies regarding adult OSA, in which the combination of both signals outperformed each of them alone (Álvarez et al., 2020). Thus, the contribution of AF information to detect pediatric OSA in combination with SpO_2 was proven useful and demonstrates the complementarity of both signals to aid in the diagnosis of this disease.

5.2 Deep learning approaches: optimum architectures

Two different DL architectures were proposed in this doctoral thesis to detect pediatric OSA from overnight AF and SpO_2. First, a 2D CNN was deployed to process these two signals (Jiménez-García et al., 2022), and then this model was extended to obtain a second CNN+RNN architecture (Jiménez-García et al., 2024). To date, only Vaquerizo-Villar et al. (2021) proposed a CNN to quantify OSA in children through the SpO_2 signal. Another recent approach from our group also assessed a DL architecture based on CNN to analyze ECG signals in the context of pediatric OSA (García-Vicente et al., 2023). The studies covered in the present research switched from single- to double-channel approaches by including the AF signal to the analysis, assuming that AF and SpO_2 channels are the most appropriate to detect apneic events and derive the AHI (Jiménez-García et al., 2022, 2024).

The first approach relied on a relatively simple stack of CNN blocks which core is a 2D convolutional layer, an architecture that has been successfully assessed in the context of OSA using the SpO_2 signal only (Mostafa et al., 2020; Vaquerizo-Villar et al., 2021). It was observed that the CNN performed better when 5-min epochs were used in comparison with 1-min inputs, suggesting that the model is capable of detecting groups of consecutive apneas and desaturations (clusters) that can last several minutes (Brouillette et al., 2000). The optimal performance in the validation set was obtained with a CNN of $NL = 8$ blocks, each of them including a 2D convolutional layer of $NF = 64$ filters with size ($K = 17, 2$). An increased number of layers and filters failed to generalize since the performance in the validation set decayed beyond the reported hyperparameters (Jiménez-García et al., 2022). Compared to the previous single-channel approach, the 2D CNN showed higher diagnostic ability by incorporating AF to the processing pipeline (Jiménez-García et al., 2022; Vaquerizo-Villar et al., 2021). This result suggests that incorporating AF together with SpO_2 can be advantageous, especially to discriminate between the mildest cases (AHI$<$ 5 e/h) (Jiménez-García et al., 2022).

The second DL approach completed the previously proposed CNN with a RNN able to model the temporal structure of the input data (Jiménez-García et al., 2024). Although CNN+RNN architectures have been proposed in the context of sleep analysis (Biswal et al., 2018; Korkalainen et al., 2020a,b), this novel approach was never tested in pediatric OSA detection. Another novelty of this approach was the deployment of a CNN+RNN from the previous CNN by means of trans-

fer learning, allowing the development of a more sophisticated architecture that leverages the pattern recognition ability of a simpler model. This simplifies the training and validation of the CNN+RNN model and also allows to estimate the AHI from larger segments. The RNN part was optimized by using a Bi-GRU layer of $NG = 4$ units that analyzed the sequences of CNN-derived feature maps. Although this value can be low, it may suffice to model the recurrence of apneic events distributed in large clusters throughout the 30-min input sequences (Jiménez-García et al., 2024). In addition, the results shown by this approach reinforces the choice of analyzing large time intervals that may contain clusters of repetitive OSA-related patterns (Brouillette et al., 2000; Vaquerizo-Villar et al., 2021). Finally, since this model was combined with a XAI algorithm, users can identify relevant patterns related to OSA detection highlighted by the network and trust in the automated decision-making process.

5.3 Explainable artificial intelligence: Grad-CAM heatmaps

As far as we know, the research done in this doctoral thesis includes the only study that proposes a XAI algorithm to interpret a DL model aimed at detecting pediatric OSA (Jiménez-García et al., 2024). Consequently, this is the first time that Grad-CAM is applied to explain a DL model aimed at detecting pediatric OSA using AF and/or SpO_2. Only one recent study focused on adult OSA addressed the explainability of a 1D CNN aimed at detecting apneic events from SpO_2, heart rate, thoracic and abdominal respiratory signals using Grad-CAM (Serrano Alarcón et al., 2023). Similar recent approaches using Grad-CAM were restricted to sleep staging, both in adults (Dutt et al., 2022), and also in children (Vaquerizo-Villar et al., 2023). Another study centered in pediatric OSA applied SHAP to explain a feature-based boosting model (Ye et al., 2022).

The Grad-CAM algorithm allowed us to visualize relevant patterns in AF and SpO_2 signals that DL models link to the presence of apneic events. As can be seen in Figures 4.2 and 4.3, sudden variations of the AF signal, as well as SpO_2 drops and recoveries are the main OSA-related patterns that the CNN+RNN model is able to associate with the presence of apneic events. For example, the heatmaps highlight zones in which the AF suddenly changes its normal oscillatory behavior, or present high-amplitude peaks that may be associated to an arousal. However, some highlighted AF patterns like isolated missed breaths are not accompanied by

a desaturation. Consequently, these explanatory heatmaps allow us to interpret irregular breathing patterns that the model can associate to apneas or hypopneas. Some examples are shown in Figure 4.2(a,b), in which regular and periodic breathing is interrupted, or when AF amplitude increases after some respiratory cycles (Figure 4.2(c,d)). However, it was also observed that SpO_2 heatmaps are stronger in case of desaturations, which is an important pointer of apneic events in children. Regions with recurrent and strong SpO_2 fluctuations were easily highlighted and can be interpreted as SpO_2 drops associated to apneas or hyponeas, even though the AF signal does not show a clear respiratory cessation (see Figure 4.3(a,d)). This can aid sleep technicians to check carefully the AF signal in order to account for recurrent hypopneas that are not easily detected by either a human scorer or the CNN+RNN architecture. Another interesting pattern highlighted by Grad-CAM correspond to regions with constant SpO_2 values. In these cases, the user can interpret that the absence of desaturations contributes to predict no apneas/hypopneas even though the AF pattern suggests the presence of one or more events Some examples can be seen in Figure 4.2(a) and Figure 4.3(b). If desaturations were present in these examples, the model probably would have detected these OSA-related events. Finally, it can be seen that some artifacts are also highlighted in the heatmaps, suggesting that the CNN+RNN model is sometimes prone to misinterpret them. The example shown in Figure 4.3(c) shows a strong heatmap in the AF signal during an artifact that could have been interpreted by the network as an apnea. Moreover, noise and low AF signal quality could be possibly hiding apneic events linked to a clear desaturation in Figure 4.3(d).

In general, it has been observed that Grad-CAM highlighted relevant patterns from both signals, which can contribute to understand the behavior of these two signals in presence of apneic events and how the model works. Moreover, these heatmaps can aid to discover hidden information about OSA particularities beyond AF cessations and SpO_2 fluctuations. In addition, the Grad-CAM heatmaps can aid the user to identify what patterns led the algorithm to incorrectly detect an apneic event, such as noise, signal loss, or movement artifacts.

5.4 Diagnostic performance of the proposed approaches

5.4.1 Comparison between ensemble and deep learning models

In this section, the algorithms proposed in this doctoral thesis to aid in the diagnosis of OSA in children are compared. As introduced in Sections 4.4 and 5.1, the most accurate ensemble learning models emerged from the combination of AF with SpO_2, especially including ODI 3% among the features. As can be seen in Figure 4.4 and Table 4.1, their diagnostic performance is very similar, being identical for the AHI cutoffs of 5 and 10 e/h. Only a slight difference favors the combination AF+ODI 3%, which reached the highest $Acc = 81.28\%$ and $Se = 92.06\%$, as well as the highest $NPV = 51.92\%$ in 1 e/h. This was also observed in the PPV-NPV and $LR-$-$LR+$ pairs (Jiménez-García et al., 2020). $Sp = 36.00\%$ is the weakest diagnostic metric, but was only slightly lower than the other configuration (Table 4.1). This result suggests that these models tend to estimate at least mild OSA in control subjects, which should be taken into account if a screening protocol is to be implemented. In this sense, a high rate of children should be re-evaluated to confirm if they are affected by OSA or not. With regard to the results reached in 5 e/h, the models showed moderate diagnostic ability as shown by $Acc = 82.05\%$, $Sp = 85.66\%$, and $NPV = 85.66\%$. This could be advantageous to discard surgical treatment in the case of getting a negative result in 5 e/h, because both Sp and NPV are high. The diagnostic ability in 10 e/h stands out by means of the high $Acc = 90.26\%$ and $PPV = 88.14\%$, as well as very high $LR+ = 27.4768$. These results are useful to directly derive children predicted as having severe OSA to surgical treatment, given that most of them would be confirmed in a hypothetical PSG-based assessment.

Notwithstanding the promising results of the AdaBoost models, their diagnostic performances were clearly surpassed by both DL architectures in all AHI cutoffs. The Bland-Altman plots shown in Figure 4.6 show high agreement between actual and estimated AHI, also reflected by the high ICCs. Nevertheless, the agreement was slightly lower in the UofC test database. In order to compare ensemble and DL algorithms, the results reached in the UofC database should be used for the comparison because this dataset was used in both approaches. With regard to 4-class classification of OSA severity, both DL models are one step ahead of AdaBoost in terms of accuracy ($Acc_4 = 62.31\%$ and $Acc_4 = 61.79\%$ vs.

Acc_4 = 57.95%, respectively) and Cohen's Kappa (k = 0.4495 and k = 0.4469 vs. k = 0.3930 and k = 0.3984, respectively). This is also reflected in the diagnostic performance in all cutoffs (Tables 4.1 and 4.3), which was higher using the CNN and CNN+RNN models. The same Acc = 84.10% in 1 e/h was reported in both DL approaches in comparison with Acc = 81.28% using AdaBoost. A better balance between Sp and NPV was observed in the CNN model, which may be advantageous to primarily discard subjects without OSA who were referred to PSG. This is also supported by $LR-$ = 0.1276 and $LR-$ = 0.1035, both in 1 e/h using the CNN and CNN+RNN models, respectively. These two rates are very close to 0.1, which would mean a strong evidence when the model discards OSA (Deeks and Altman, 2004). The CNN+RNN architecture was the best overall model in 5 e/h since it surpassed both CNN and AdaBoost in all the performance rates (Table 4.3). The moderate-to-high rates of correctly classified patients in 5 e/h encourage to use this model to deploy a screening protocol. The CNN+RNN model seems to be the ideal choice to primarily decide whether or not deriving children to further diagnosis by means of PSG or directly to surgical treatment. Lastly, the DL models also have high diagnostic value to directly derive children to surgical treatment due to their high Acc and Se while maintaining also high Sp, which is crucial to minimize possible false positives. The $LR+$ = 12.6538 was much lower compared to AdaBoost but it remained above 10, which could establish a great evidence when the model predicts severe OSA (Deeks and Altman, 2004).

The diagnostic ability of DL models was also assessed in the CHAT test set. In this case, ICC, k, and Acc_4 were higher than those obtained in the UofC dataset (Jiménez-García et al., 2022, 2024). For example, the Bland-Altman plots (Figures 4.5 and 4.6) show that the confidence interval of the AHI estimation error is much narrower in the CHAT database, which is consequent with the higher ICC obtained in this dataset. As a result, the OSA severity estimation results summarized in the confusion matrices of Figure 4.7 also show high agreement in the 4-class classification task. The CNN+RNN show slightly better agreement compared to CNN (Acc_4 = 74.51%, k = 0.6231 vs. Acc_4 = 72.55%, k = 0.6011, respectively), which confirms that the CNN+RNN resulted in a improved version of the CNN. With respect to the diagnostic performance in 1, 5, and 10 e/h, Table 4.2, the CNN+RNN show some advantages in comparison with CNN. In 1 e/h, the CNN+RNN reached an Acc = 87.25% with a balanced Se-Sp pair, along with high NPV = 65.56%. Again, these results suggest the usefulness of this proposal to discard the presence of OSA due to the high rate of patients that

were correctly classified as no OSA subjects ($LR- = 0.1473$ was low but slightly higher than the same rate observed in the UofC database). Both DL models obtained exactly the same diagnostic performance in the intermediate cutoff, with high $Acc = 93.46\%$ and near to excellent $Sp = 99.07\%$ and $PPV = 97.33\%$. This relevant result may allow to directly refer surgical treatment to those patients who tested positive using this automatic method, since the number of false positives (i.e., those who tested positive but do not need surgical treatment because their actual AHI<5 e/h) is very reduced. With respect to the diagnostic ability in 10 e/h, the CNN model was slightly more accurate ($Acc = 94.44\%$ vs. $Acc = 93.46\%$), but both Se and Sp were very similar within each other. These results indicate that the proposed DL models can also be used to directly recommend surgical treatment to those patients who receive a severe OSA diagnosis using the proposed models, because $LR+ > 10$ indicate strong evidence when the model detects severe OSA (Deeks and Altman, 2004).

The differences in the diagnostic ability across databases may be motivated by the discrepancies between scorers and the study design. PSG data corresponding to the CHAT database was scored using a more normalized research protocol in a common PSG reading center (Marcus et al., 2013), whereas the UofC database was obtained in a more clinical setting in which various scorers may have derived different interpretation of the signals (Collop, 2002). Moreover, only CHAT data could be used to train the DL models at a epoch/segment level (i.e., to detect the number of apneas/hypopneas) since only this database contained the annotations of apneic events. Nevertheless, both CHAT and UofC databases were used to validate the algorithms and optimize hyperparameters, so the chances of overfitting were reduced (Jiménez-García et al., 2022, 2024).

5.4.2 Clinical usefulness of CNN+RNN: screening protocol

According to these results, a screening protocol can be derived following the most accurate model (CNN+RNN) AHI estimates. This protocol might be implemented in low-resource settings that do not allow a proper PSG-based examination for children. This way, by recording only AF and SpO_2 with a portable Type 3 or 4 device and analyzing the overnight recording with the proposed CNN+RNN, a primary diagnosis can be obtained. The following medical decisions can be determined in view of the result provided by the algorithm:

1. The CNN+RNN algorithm predicts AHI<1 e/h: children may not need further diagnostic tests such as PSG, and surgery can also be discarded be-

cause only 1.63% of children who tested negative with this model actually had AHI≥5 e/h, which is the threshold to refer surgical treatment. Nevertheless, caregivers should watch out the chidren's symptoms and report them for further assessment.

2. The CNN+RNN algorithm predicts 1≤AHI<5 e/h: start non-surgical treatment of pediatric OSA symptoms (weight loss, anti-inflammatory drugs, etc.). In addition, patients should be followed up to re-evaluate their symptoms. Surgery should be initially discarded because only 11.99% of children with predicted 1≤AHI<5 e/h actually had moderate-to-severe OSA. If the symptoms persist, derive to PSG in order to confirm surgical treatment.

3. The CNN+RNN algorithm predicts 5≤AHI<10 e/h: derive children to PSG to confirm preliminary OSA diagnosis. Depending on PSG result, consider surgical treatment because 69.72% of children with this preliminary result had at least moderate-to-severe OSA.

4. The CNN+RNN algorithm predicts AHI≥10 e/h: recommend surgical treatment evaluation without performing PSG because only 3.28% of children with preliminary AHI≥10 e/h according to the model actually had AHI<5 e/h. If surgeons do not consider it appropriate to operate, propose pharmacological treatment.

Using this protocol, up to 78.45% of pediatric PSGs could potentially be avoided, drastically reducing the workload of sleep technicians and the waiting lists of the laboratories. This way, PSG could be inmediately available for those children that initially obtained 1≤AHI<5 with the automatic method and have persistent symptoms, as well as those who initially had 5≤AHI<10 regardless the symptoms.

5.5 Comparison with previous studies

As mentioned in Section 1.5, several ML methodologies have been proposed in the past to overcome the simplification of OSA diagnosis in children. These approaches combine signal processing algorithms with pattern recognition models to automatically detect pediatric OSA from a variety of biomedical recordings (Bertoni and Isaiah, 2019; Gutiérrez-Tobal et al., 2022). A meta-analysis involving the use of these pattern recognition methods to diagnose pediatric OSA has been recently carried out, which estimates the diagnostic performance of these ML-based approaches as: $Se = 84.9\%$, $Sp = 49.9\%$ (1 e/h), $Se = 71.4\%$, $Sp = 83.2\%$ (5 e/h),

and Se = 65.2%, Sp = 93.1% (10 e/h) (Gutiérrez-Tobal et al., 2022). Regarding the use of AF and SpO_2 data, it was observed that several characterizations have been addressed and different ML approaches have been tested. Table 5.1 summarizes the most recent studies that have applied pattern recognition to AF and/or SpO_2 to aid in the diagnosis of pediatric OSA. The diagnostic performance of these methods is also reported in Table 5.2.

As shown in Table 5.1, SpO_2 is the most used signal, either alone or combined with other sources such as PRV or AF. The most common features involved in these approaches were temporal, spectral and nonlinear parameters, but some studies also addressed bispectrum or wavelets. A large portion of these studies also included ODI 3% among other oximetric indices. Regarding pattern recognition methods, the most common were LR or shallow NNs. It is necessary to note that other studies also used ensemble learning methods such as AdaBoost or XGBoost, and very few implemented a DL architecture (CNN, RNN, etc.). The diagnostic performances reached by all these feature-based methods was very similar, with Acc =75.0%–83.2% in 1 e/h, Acc =78.5%–84.9% in 5 e/h, and Acc =89.0%–91.1% in 10 e/h. In comparison with these studies, the study elaborated within this doctoral thesis combining AF with ODI 3% showed a diagnostic ability similar to other approaches that combined ODI 3% with AF or other SpO_2 features (Barroso-García et al., 2021a,b; Vaquerizo-Villar et al., 2018a,c). This suggests that the investigation of novel approaches to characterize AF and/or SpO_2 is now mature, and is progressively moving forward to DL approaches that assume the task of learning their own representations of input data. Only four DL-based approaches were proposed in the context of pediatric OSA diagnosis, being two part of this doctoral thesis (Jiménez-García et al., 2022, 2024). It can be seen that these novel methods reached higher diagnostic performance than the previous ones: Acc =75.9%–84.1% in 1 e/h, Acc =83.9%–87.0% in 5 e/h, and Acc=90.3%–92.3% in 10 e/h. To note, the results reached in the study of (García-Vicente et al., 2023) were obtained using the CHAT database instead of UofC, which was used to elaborate this comparative table.

With regard to the DL approaches presented in this doctoral thesis, it can be seen that the models combining AF and SpO_2 surpassed the diagnostic ability of the only previous approach based on SpO_2 alone (see the last rows of Table 5.2). The CNN and CNN+RNN approaches accomplished the OSA detection with the highest Acc=84.1% in 1 e/h, and also reached remarkable Se, PPV and NPV in this cutoff in comparison with other approaches. The CNN+RNN was more accurate than the previous approaches, with remarkable Se, PPV and Sp. How-

Table 5.1: Summary of previous methodologies focused on the automatic OSA diagnosis in children using AF and/or SpO_2.

Study	Signal	Extraction	Selection	Pattern recognition	Validation	#Total/ #Test
Garde et al. (2014)	SpO_2, PRV	Temporal, Non-linear, Spectral	AUC optimization	LDA	Loo-cv / 4-fold-cv	146/146
Garde et al. (2019)	SpO_2, PRV	Temporal, Spectral, ODI 3%	Stepwise LR	LR	Loo-cv	207/207
Calderón et al. (2020)	SpO_2	Oximetric indices	–	LR,AdaBoost	10-fold-cv	453/453
Ye et al. (2022)	SpO2, HR	ODI, HR statistics		XGBoost		3139/628
Hornero et al. (2017)	SpO_2	Temporal, Spectral, Non-linear, ODI 3%	FCBF	NN	Holdout	4191/3602
Álvarez et al. (2018)	SpO_2	Anthropometrics, Temporal, Symbolic dynamics, ODI 3%	FSLR	LR	Bootstrap	142/142
Vaquerizo-Villar et al. (2018a)	SpO_2	ODI 3%, DFA	FCBF	NN	Holdout	981/392
Vaquerizo-Villar et al. (2018c)	SpO_2	Temporal, Spectral, Wavelet, ODI 3%	FCBF	SVM, LR	Holdout	981/392
Barroso-García et al. (2020)	AF, SpO2	Recurrence Plots, ODI 3%	FCBF	NN	Holdout	946/376
Barroso-García et al. (2021a)	AF, SpO_2	Bispectrum, ODI 3%	FCBF	NN	Holdout	946/376
Barroso-García et al. (2021b)	AF, SpO_2	Wavelet, ODI 3%	FCBF	MLP,AdaBoost	Holdout	946/376
Jiménez-García et al. (2020)	**AF, SpO_2**	**Temporal, Spectral, Non-linear, ODI 3%**	**FCBF**	**AdaBoost.M2**	**Holdout**	**974/390**
Vaquerizo-Villar et al. (2021)	SpO_2	–	–	CNN	Holdout	3196/935
García-Vicente et al. (2023)	ECG	–	–	CNN	Holdout	1610/299
Jiménez-García et al. (2022)	**AF, SpO_2**	–	–	**2D CNN**	**Holdout**	**2612/696**
Jiménez-García et al. (2024)	**AF, SpO_2**	–	–	**CNN+RNN**	**Holdout**	**2612/696**

ECG: electrocardiogram, SpO_2: blood oxygen saturation signal, AF: airflow signal, PRV: pulse rate variability signal, ODI 3%: 3% oxygen desaturation index, DFA: Detrended Fluctuation Analysis, AUC: area under the receiver operating characteristic curves, FSLR: forward stepwise logistic regression, FCBF: fast correlation based filter, LDA: linear discriminant analysis, LR: logistic regression, NN: neural network, SVM: support vector machine, AdaBoost: adaptive boosting, Loo-cv: leave-one-out cross validation.

Table 5.2: Comparison of the diagnostic performance obtained in other previous studies focused on the automatic OSA diagnosis in children.

Study	AHI	Se(%)	Sp(%)	Acc(%)	PPV(%)	NPV(%)	LR+	LR–
Garde et al. (2014)	5	88.4	83.6	84.9	76.9	92.6	5.4	0.1
	1	80.0	65.0	75.0	–	–	2.3	0.3
Garde et al. (2019)	5	85.0	79.0	82.0	–	–	4.1	0.2
	10	82.0	91.0	89.0	–	–	9.1	0.2
Calderón et al. (2020)	5	62.0	96.0	79.0	94.3	–	15.5	0.4
	1	90.3	100	90.5	100	16.7		
Ye et al. (2022)	5	82.0	93.8	85.7	96.7	69.9		
	10	84.8	92.1	89.8	83.1	93.0		
	1	84.0	53.2	75.2	81.6	53.7	1.8	0.3
Hornero et al. (2017)	5	68.2	87.2	81.7	68.6	87.0	5.3	0.4
	10	68.7	94.1	90.2	67.7	94.3	11.6	0.3
Álvarez et al. (2018)	5	73.5	89.5	83.3	82.0	84.3	10.4	0.3
	1	97.1	23.3	82.7	83.9	66.7	1.3	0.12
Vaquerizo-Villar et al. (2018a)[(1)]	5	78.8	83.7	81.9	74.2	86.9	4.8	0.25
	10	77.1	94.8	91.1	80.0	93.9	14.9	0.24
Vaquerizo-Villar et al. (2018c)[(1)]	5	71.9	91.1	84.0	83.8	84.5	14.6	0.3
	1	97.7	22.2	83.2	84.1	69.6	1.3	0.1
Barroso-García et al. (2020)[(1)]	5	78.7	78.3	78.5	68.5	86.0	3.6	0.2
	10	78.8	94.3	91.0	78.8	94.3	13.7	0.2
	1	98.0	15.3	82.2	83.0	65.0	1.2	0.1
Barroso-García et al. (2021a)[(1)]	5	81.6	83.0	82.5	74.2	88.3	4.9	0.2
	10	72.3	95.0	90.2	79.6	92.7	15.0	0.3
	1	80.3	68.1	78.0	91.5	44.9	2.6	0.3
Barroso-García et al. (2021b)[(1)]	5	68.0	90.3	81.9	80.8	82.5	7.2	0.4
	10	72.4	96.0	91.0	83.0	92.8	19.0	0.3
	1	92.1	36.0	81.3	85.8	51.9	1.4	0.2
Jiménez-García et al. (2020)[(1)]	5	76.0	85.7	82.1	76.0	85.7	5.3	0.3
	10	62.7	97.7	90.3	88.1	90.6	27.5	0.4
	1	90.8	36.4	80.1	85.4	49.1	1.4	0.25
Vaquerizo-Villar et al. (2021)[(1)]	5	76.0	88.6	83.9	79.8	86.2	6.7	0.3
	10	79.5	95.8	92.3	83.5	94.6	18.9	0.2
	1	84.2	46.2	75.9	84.9	44.8	1.6	0.3
García-Vicente et al. (2023)[(2)]	5	76.7	91.4	87.0	79.3	90.1	8.9	0.3
	10	53.7	98.1	92.0	81.5	93.0	27.7	0.5
	1	95.2	37.3	84.1	86.5	65.1	1.5	0.1
Jiménez-García et al. (2022)[(1)]	5	82.2	85.3	84.1	76.9	88.9	5.6	0.2
	10	78.3	93.5	90.3	76.5	94.1	12.0	0.2
	1	96.8	30.7	84.1	85.4	69.7	1.4	0.1
Jiménez-García et al. (2024)[(1)]	5	82.9	85.7	84.6	77.6	89.3	5.8	0.2
	10	78.3	93.8	90.5	77.4	94.1	12.7	0.2

AHI: apnea-hypopnea index, Se: sensitivity, Sp: specificity, Acc: accuracy, PPV: positive predictive value, NPV: negative predictive value, LR+: positive likelihood ratio, LR–: negative likelihood ratio. [(1)]: Results on UofC test set. [(2)]: Results on CHAT test set.

ever, the diagnostic ability of the SpO_2-based CNN surpassed both dual-channel approaches, suggesting that SpO_2 only would suffice to detect the most severe

cases (Jiménez-García et al., 2022; Vaquerizo-Villar et al., 2021). This is also supported by other studies that assessed the role of oximetry as an abbreviated test for pediatric OSA (Kaditis et al., 2016a).

Overall, the results presented in this doctoral thesis confirm that DL architectures have a strong capability to automatically detect the patterns of AF and SpO_2 related to apneas, hypopneas, and their corresponding desaturations, and use them to estimate the presence of OSA and its severity. Furthermore, these DL models can be visually assessed with XAI methods to obtain explanations of their functioning and discover these possibly hidden patterns that the deep networks associate with signs of pathology.

5.6 Limitations of the study

Notwithstanding the usefulness of the research conducted in this doctoral thesis, it is necessary to note some limitations that could be addressed in the future.

Although the studies that constitute this doctoral thesis were conducted using two different databases covering a total of 2,612 pediatric patients, these can be complemented by adding more subjects. The first study of this thesis comprised a total of 974 subjects and the two subsequent works were conducted using an additional public database of 1,638 pediatric patients. Nevertheless, the differences in age, sex, or AHI between children recruited to form each database should be taken into account to analyze the results. Validation techniques such as bootstrap or 10-fold cross validation have been used to minimize the chances of overfitting in the the developed algorithms, improving their generalizability. Furthermore, a portion of our databases was always preserved to test the different deployments in a common set of unseen data. However, larger and more heterogeneous databases would represent the particularities of pediatric OSA and can further enhance the generalizability of the proposed automatic methods.

This study could also benefit from a more exhaustive analysis of the diagnostic performance in specific subgroups of the population that formed our database. Although sociodemographic and anthropometric variables were taken into account to divide equally the databases into training, validation, and test sets, as well as to avoid a lack of generalization, a stratification of the conducted tests was not assessed.

The disease covered in this doctoral thesis is also a frequent condition in adults. This study was exclusively focused on pediatric OSA, and the validity of the results are thus restricted to the pediatric population. Although pediatric OSA

specificities have motivated the development of this study, the broad similarity with adult OSA can make the algorithms presented in thesis a feasible starting point to propose similar solutions for adults.

All sleep studies utilized in this research were derived from Type 1 PSGs performed in sleep laboratories. These are attended settings with the supervision of experienced and skilled medical staff. Therefore, the validation of our methods in AF and SpO_2 signals recorded from an unattended and/or domiciliary setting would be desirable.

Regarding the recording of AF signals, the oronasal thermistor was selected to perform the present study instead of the pressure sensor. Notwithstanding the preference of thermistors to detect apneas, the AASM recommends using also the nasal pressure sensor to detect hypopneas. However, the thermistor is also considered an alternative to the pressure sensor when this is not available (Berry et al., 2012). Similarly, pulse oximeters are not able to detect hypopnea events associated to arousals. Therefore, a more exhaustive analysis of the detection this type of events should be taken into account. Overall, this study is naturally limited by the choice of AF and SpO_2 signals as data sources. Other overnight recordings such as PRV are also easy to obtain and could potentially complement those signals.

AF and SpO_2 signals were used to detect apneic events, but the AHI estimation also comprises the computation of the total sleep time (TST). In a PSG setting, TST is derived from the time spent in each of the sleep stages (Berry et al., 2012), but these states are determined by means of EEG/EOG analysis. The linear regression implemented in the DL-based architectures to perform the AHI estimation can partially overcome this issue, but an assessment of the information about wakefulness/sleep states to accurately compute TST from a convenient sensor needs to be addressed.

The feature engineering approaches covered in this doctoral thesis are naturally limited by the type of extracted features and the algorithms used to compute them. The natural way to overcome this limitation would be to investigate alternative signal processing and analysis methodologies. However, it can be seen in 5.2 that DL approaches have the potential to outperform most of the current feature extraction and selection methodologies.

With respect to the deployment of DL architectures, only two models have been assessed. Being DL a field in continuous expansion, the number of available architectures and algorithms is increasing. Therefore, the investigation of DL approaches in the context of pediatric OSA is far from being mature.

Finally, the explainability of the proposed DL models has been addressed by deploying an implementation of the Grad-CAM algorithm. Although this method has shown its usefulness in the detection of OSA-related patterns, complementary XAI methods need to be assessed, paying special attention to algorithms aimed at explaining time series classification (Theissler et al., 2022).

Chapter 6

Conclusions

As mentioned in the previous chapters, the ensemble learning, DL, and XAI approaches covered in the doctoral thesis allowed us to provide accurate and trustworthy automated diagnostic methods that can serve as an aid in the diagnosis of pediatric OSA by means of only AF and SpO_2 biomedical signals.

This chapter summarizes the contributions of this thesis, the main conclusions, and future research lines derived from these studies. The novel contributions of the research works conducted in this doctoral thesis are indicated in Section 6.1. The conclusions drawn from the research works of this compendium are listed in Section 6.2. Lastly, the possible future research lines that can continue and complete the investigations initiated in this doctoral thesis are introduced in Section 6.3.

6.1 Contributions

The research works conducted in this doctoral thesis represent a number of breakthroughs in the investigation of alternatives to diagnose pediatric OSA. These are the most important contributions of this doctoral thesis:

1) **A direct comparison of AF and SpO_2 signals to automatically detect pediatric OSA**. Previous studies in the context of childhood OSA diagnosis assessed the usefulness of the automatic analysis of single-channel AF or SpO_2 alone, but none of them proposed a comparison of the diagnostic ability of these signals, either alone or combined. In order to cover this gap, a database of overnight AF and SpO_2 signals from pediatric patients was elaborated and a exhaustive characterization of both signals was

performed. As a result, several features from AF and SpO_2 were obtained and evaluated. Moreover, a feature selection algorithm was used to identify the most relevant and complementary ones. The main discovery was that the CTM computed from AF was complementary with ODI 3%, the most discriminative feature of SpO_2. Thus, these two features have potential to detect OSA jointly with enhanced diagnostic performance.

2) **Assessment of ensemble learning methods using information from AF and SpO**2. To date, only classic and widespread ML algorithms such as LDA, LR, SVM, etc. were proposed to automatically detect childhood OSA. However, ensemble learning algorithms such as AdaBoost were not considered even though these methods showed state-of-the-art performance in a variety of contexts. This drawback was overcome by deploying multi-class AdaBoost models trained with the previously mentioned features of AF and/or SpO_2 to classify OSA severity into 4 levels: no OSA, mild, moderate, and severe OSA. As a result, an accurate and generalizable ensemble learning model was proposed as a diagnostic aid of pediatric OSA using the information of AF and SpO_2.

3) **A novel 2D CNN to process AF and SpO_2 jointly aimed at estimating the AHI**. To the best of our knowledge, only one previous study was aimed at detecting pediatric OSA using DL. This DL model was restricted to analyze SpO_2, thus not being able to simultaneously process 2 signals. The development of a new 2D CNN architecture aimed at analyzing overnight AF and SpO_2 enabled the analysis of these 2 signals jointly by means of a CNN. As a result, an accurate model was deployed and further evaluated in two pediatric OSA databases. The diagnostic ability of this model surpassed that showed by the SpO_2-based CNN, especially in 1 e/h and 5 e/h.

4) **A state-of-the-art CNN+RNN architecture**. In order to exploit the benefits of both CNNs and RNNs, a natural extension of the 2D CNN proposed in the previous study was developed. The combination of a CNN able to process AF and SpO_2 simultaneously with a RNN that analyzes the temporal dependencies of these signals in large segments was deployed. As a result, the diagnostic performance of this novel architecture surpassed those reached by previous approaches.

5) **A transfer learning approach to optimize the training and performance of the CNN+RNN**. Since the CNN+RNN algorithm optimization

is a cumbersome process, a transfer learning approach was implemented in order to reduce training time and focus on the optimization of the RNN part of the architecture. Therefore, the previously optimized CNN layers were transferred to the CNN+RNN model. As a result, the CNN+RNN model not only was faster to train, but also gained generalization ability.

6) **Explanatory heatmaps of the predictions derived from the CNN+RNN model using Grad-CAM**. An important breakthrough in the research on automatic diagnostic aid methods for OSA quantification was to add interpretability to the CNN+RNN model. This was done by generating Grad-CAM heatmaps that highlight the patterns this network associates to the presence of OSA-related abnormalities. As a result, users can trust in the decisions made by automatic diagnostic methods.

6.2 Main conclusions

The following conclusions can be drawn from the research conducted throughout this doctoral thesis:

1) The joint analysis of AF and SpO_2 signals is useful to automatically detect pediatric OSA. In this sense, relevant and complementary features from both signals were useful to derive automatic classification models. AF-derived CTM and SpO_2-derived ODI 3% were the most useful features from these signals, which showed their relevance to detect pediatric OSA and non-redundancy within each other.

2) Multi-class AdaBoost classifiers are able to accurately detect pediatric OSA severity from AF and SpO_2. The classifiers achieved remarkable diagnostic performance from a set of complementary features derived from AF and SpO_2 signals.

3) In the feature engineering approach, CTM and ODI 3% are the most useful features from AF and SpO_2, respectively. The combination of these features by means of AdaBoost enhanced and maximized the diagnostic ability of this classification algorithm in comparison with other combinations involving the use of only one signal alone.

4) A dual-channel approach to detect pediatric OSA involving AF and SpO_2 signals is advantageous against single-channel approaches. Both ensemble

and DL models achieved higher diagnostic performance in comparison with other approaches that only comprised one signal.

5) A 2D CNN architecture was useful to process and analyze AF and SpO_2 signals and to estimate the AHI from these overnight recordings. The proposed CNN-based approach achieved a remarkable diagnostic performance to detect pediatric OSA, especially when differentiating between control and mild OSA subjects, as well as discriminating mild and moderate OSA patients. The proposed AF and SpO_2-based CNN surpassed the diagnostic performance of previous DL models aimed at analyzing SpO_2 signals alone.

6) The combination of CNN with RNN was useful to detect pediatric OSA from AF and SpO_2. In this sense, the use of a transfer learning approach to develop a more sophisticated CNN+RNN model demonstrated its usefulness. This novel architecture outperformed previous CNN-based models as well.

7) The CNN+RNN architecture aimed at analyzing AF and SpO_2 was the most accurate DL model to estimate OSA severity by means of the AHI. To date, the performance of this algorithm is the state-of-the-art among automatic pediatric OSA detection using a reduced set of biomedical signals since it reached the highest diagnostic ability in comparison with similar approaches aimed at the same population.

8) The use of XAI methods such as Grad-CAM enabled the exploration of OSA-related patterns in AF and SpO_2 signals. Explanatory heatmaps highlighted specific parts of the input data which were relevant for the CNN+RNN model to make their predictions, and allowed us to interpret the functioning of this complex architecture.

9) The Grad-CAM heatmaps obtained from the CNN+RNN had the ability to justify why the CNN+RNN model predicted the presence of apneas or hypopneas in the input signals. Therefore, Grad-CAM contributed to enhance its diagnostic trustworthiness.

10) Explanatory heatmaps revealed relevant OSA-related patterns learned by the CNN+RNN algortithm. The patterns highlighted in the AF and SpO_2 signals by means of Grad-CAM were mainly related to desaturations and sudden changes in the amplitude of respiratory waves. These explanatory heatmaps might be useful for sleep technicians to analyze and interpret these signals with the objective of simplifying the diagnosis of pediatric OSA.

Based on the aforementioned statements, the global conclusion of this doctoral thesis is that the automatic signal processing and analysis of AF and SpO_2 based on ensemble learning and DL methods combined with XAI proposed in this research has a great diagnostic usefulness, and can be used to deploy alternative, simple, reliable and trustworthy screening methods to serve as an aid in the diagnosis of OSA in children.

6.3 Future research lines

At the same time that the present research was being developed, some interesting ideas were moved aside in order to be addressed in the future. The following research ideas could compensate the limitations exposed in Section 5.6:

1) Expanding our databases to cover a greater and more diverse population of children with suspicion of suffering OSA would be an interesting future goal. In this sense, the application of the proposed models in a wider population including children with a high risk of developing OSA in the future can be addressed.

2) It would also be desirable to identify subgroups of children at risk of OSA and assess the proposed methods specifically in these populations (e.g., obese children, with down syndrome, other malformations, etc.).

3) It would be feasible to propose, optimize and prospectively validate the algorithms presented in this thesis in a population of adults with suspicion of OSA with the objective of reaching a greater number of affected people.

4) In order to ensure the utility of deployed methods outside the hospital facilities, it would be desirable to include AF and SpO_2 signals from portable Type 3 or 4 devices recorded at the patient's home in our analyses. This can be useful to identify possible differences between in-lab and at-home recordings.

5) It would be desirable to assess other cardiorespiratory signals apart from AF and SpO_2 in order to enhance the detection of pediatric OSA. Including other useful cardiac signals such as ECG or HRV -or pulse oximetry-derived PRV- could be a future research line. In this sense, cardiac and/or pulse information can be incorporated to assess their response to apneas and/or hypopneas or obtain other useful sleep parameters (García-Vicente et al., 2023; Garde et al., 2019; Martín-Montero et al., 2021b, 2023).

6) More advanced ML and DL methodologies such as hybrid models, transformers, etc. could be considered to expand and optimize the use of automatic methods as an aid in the diagnosis of pediatric OSA.

7) Complementary XAI methods could be applied to increase the quality and quantity of the explanations about models' outcomes. In this regard, attention mechanisms or other model-agnostic XAI algorithms optimized to explain time series can be taken into consideration.

Chapter 7

Papers included in the compendium

7.1 Contribution 1: Jiménez-García et al. (2020)

Assessment of Airflow and Oximetry Signals to Detect Pediatric Sleep Apnea-Hypopnea Syndrome Using AdaBoost

Jorge Jiménez-García, Gonzalo C. Gutiérrez-Tobal, María García, Leila Kheirandish-Gozal, Adrián Martín-Montero, Daniel Álvarez, Félix del Campo, David Gozal, and Roberto Hornero. *Entropy*, vol. 22 (6), pp. 670, 2020. Impact factor in 2020: 2.524, Q2 in "Physics, Multidisciplinary" (JCR-WOS).

DOI: https://doi.org/10.3390/e22060670.

Abstract: The reference standard to diagnose pediatric Obstructive Sleep Apnea (OSA) syndrome is an overnight polysomnographic evaluation. When polysomnography is either unavailable or has limited availability, OSA screening may comprise the automatic analysis of a minimum number of signals. The primary objective of this study was to evaluate the complementarity of airflow (AF) and oximetry (SpO_2) signals to automatically detect pediatric OSA. Additionally, a secondary goal was to assess the utility of a multiclass AdaBoost classifier to predict OSA severity in children. We extracted the same features from AF and SpO_2 signals from 974 pediatric subjects. We also obtained the 3% Oxygen Desaturation Index (ODI) as a common clinically used variable. Then, feature selection was conducted using the Fast Correlation-Based Filter method and AdaBoost classifiers were evaluated. Models combining ODI 3% and AF features outperformed the diagnostic performance of each signal alone, reaching 0.39 Cohens's kappa in the four-class classification task. OSA vs. No OSA accuracies reached 81.28%, 82.05% and 90.26% in the apnea–hypopnea index cutoffs 1, 5 and 10 events/h, respectively. The most relevant information from SpO_2 was redundant with ODI 3%, and AF was complementary to them. Thus, the joint analysis of AF and SpO_2 enhanced the diagnostic performance of each signal alone using AdaBoost, thereby enabling a potential screening alternative for OSA in children.

7.2 Contribution 2: Jiménez-García et al. (2022)

A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry

Jorge Jiménez-García, María García, Gonzalo C. Gutiérrez-Tobal, Leila Kheirandish-Gozal, Fernando Vaquerizo-Villar, Daniel Álvarez, Félix del Campo, David Gozal, and Roberto Hornero. *Computers in Biology and Medicine*, vol. 147, pp. 105784, 2022. Impact factor in 2022: 7.7, Q1 in "Mathematical & Computational Biology" (JCR-WOS).

DOI: https://doi.org/10.1016/j.compbiomed.2022.105784.

Abstract: The gold standard approach to diagnose obstructive sleep apnea (OSA) in children is overnight in-lab polysomnography (PSG), which is labor-intensive for clinicians and onerous to healthcare systems and families. Simplification of PSG should enhance availability and comfort, and reduce complexity and wait-lists. Airflow (AF) and oximetry (SpO_2) signals summarize most of the information needed to detect apneas and hypopneas, but automatic analysis of these signals using deep-learning algorithms has not been extensively investigated in the pediatric context. The aim of this study was to evaluate a convolutional neural network (CNN) architecture based on these two signals to estimate the severity of pediatric OSA. PSG-derived AF and SpO_2 signals from the Childhood Adenotonsillectomy Trial (CHAT) database (1638 recordings), as well as from a clinical database (974 recordings), were analyzed. A 2D CNN fed with AF and SpO_2 signals was implemented to estimate the number of apneic events, and the total apnea-hypopnea index (AHI) was estimated. A training-validation-test strategy was used to train the CNN, adjust the hyperparameters, and assess the diagnostic ability of the algorithm, respectively. Classification into four OSA severity levels (no OSA, mild, moderate, or severe) reached 4-class accuracy and Cohen's Kappa of 72.55% and 0.6011 in the CHAT test set, and 61.79% and 0.4469 in the clinical dataset, respectively. Binary classification accuracy using AHI cutoffs 1, 5 and 10 events/h ranged between 84.64% and 94.44% in CHAT, and 84.10%–90.26% in the clinical database. The proposed CNN-based architecture achieved high diagnostic ability in two independent databases, outperforming previous approaches that employed SpO2 signals alone, or other classical feature-engineering approaches. Therefore, analysis of AF and SpO_2 signals using deep learning can be useful to deploy reliable computer-aided diagnostic tools for childhood OSA.

7.3 Contribution 3: Jiménez-García et al. (2024)

An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals

Jorge Jiménez-García, , María García, Gonzalo C. Gutiérrez-Tobal, Leila Kheirandish-Gozal, Fernando Vaquerizo-Villar, Daniel Álvarez, Félix del Campo, David Gozal, and Roberto Hornero. *Biomedical Signal Processing and Control*, vol. 87, part B, pp. 105490, 2024. Impact factor in 2022: 5.1, Q2 in "Engineering, Biomedical" (JCR-WOS).

DOI: https://doi.org/10.1016/j.bspc.2023.105490.

Abstract: Deep-learning algorithms have been proposed to analyze overnight airflow (AF) and oximetry (SpO_2) signals to simplify the diagnosis of pediatric obstructive sleep apnea (OSA), but current algorithms are hardly interpretable. Explainable artificial intelligence (XAI) algorithms can clarify the models-derived predictions on these signals, enhancing their diagnostic trustworthiness. Here, we assess an explainable architecture that combines convolutional and recurrent neural networks (CNN + RNN) to detect pediatric OSA and its severity. AF and SpO_2 were obtained from the Childhood Adenotonsillectomy Trial (CHAT) public database (n = 1,638) and a proprietary database (n = 974). These signals were arranged in 30-min segments and processed by the CNN + RNN architecture to derive the number of apneic events per segment. The apnea-hypopnea index (AHI) was computed from the CNN + RNN-derived estimates and grouped into four OSA severity levels. The Gradient-weighted Class Activation Mapping (Grad-CAM) XAI algorithm was used to identify and interpret novel OSA-related patterns of interest. The AHI regression reached very high agreement (intraclass correlation coefficient > 0.9), while OSA severity classification achieved 4-class accuracies 74.51% and 62.31%, and 4-class Cohen's Kappa 0.6231 and 0.4495, in CHAT and the private datasets, respectively. All diagnostic accuracies on increasing AHI cutoffs (1, 5 and 10 events/h) surpassed 84%. The Grad-CAM heatmaps revealed that the model focuses on sudden AF cessations and SpO_2 drops to detect apneas and hypopneas with desaturations, and often discards patterns of hypopneas linked to arousals. Therefore, an interpretable CNN + RNN model to analyze AF and SpO_2 can be helpful as a diagnostic alternative in symptomatic children at risk of OSA.

Appendix A

Scientific achievements

A.1 Publications

A.1.1 Papers indexed in the Journal Citation Reports (JCR)

1. **Jorge Jiménez-García**, Roberto Romero-Oraá, María García, María I. López-Gálvez, Roberto Hornero, "Combination of Global Features for the Automatic Quality Assessment of Retinal Images", *Entropy*, vol. 21 (3), pp. 311, March, 2019, DOI: 10.3390/e21030311. Journal Impact Factor in 2019: 2.494, Q2 in "Physics, Multidisciplinary" (JCR-WOS).

2. Roberto Romero-Oraá, **Jorge Jiménez-García**, María García, María I. López-Gálvez, Javier Oraá-Pérez, Roberto Hornero, "Entropy Rate Superpixel Classification for Automatic Red Lesion Detection in Fundus Images", *Entropy*, vol. 21 (4), pp. 417, April, 2019, DOI: 10.3390/e21040417. Journal Impact Factor in 2019: 2.494, Q2 in "Physics, Multidisciplinary" (JCR-WOS).

3. **Jorge Jiménez-García**, Gonzalo C. Gutiérrez-Tobal, María García, Leila Kheirandish-Gozal, Adrián Martín-Montero, Daniel Álvarez, Félix del Campo, David Gozal, Roberto Hornero, "Assessment of Airflow and Oximetry Signals to Detect Pediatric Sleep Apnea-Hypopnea Syndrome Using AdaBoost", *Entropy*, vol. 22 (6), pp. 670, June, 2020, DOI: 10.3390/e22060670. Journal Impact Factor in 2020: 2.524, Q2 in "Physics, Multidisciplinary" (JCR-WOS).

4. Adrián Martín-Montero, Gonzalo C. Gutiérrez-Tobal, Leila Kheirandish-Gozal, **Jorge Jiménez-García**, Daniel Álvarez, Félix del Campo, David Gozal, Roberto Hornero, "Heart rate variability spectrum characteristics in children with sleep apnea", *Pediatric Research*, vol. 89 (7), pp. 1771-1779, May, 2021, DOI: 10.1038/s41390-020-01138-2. Journal Impact Factor in 2021: 3.953, Q1 in "Pediatrics" (JCR-WOS).

5. **Jorge Jiménez-García**, María García, Gonzalo C. Gutiérrez-Tobal, Leila Kheirandish-Gozal, Fernando Vaquerizo-Villar, Daniel Álvarez, Félix del Campo, David Gozal, Roberto Hornero, "A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry", *Computers in Biology and Medicine*, vol. 147, pp. 105784, August, 2022, DOI: 10.1016/j.compbiomed.2022.105784. Journal Impact Factor in 2022: 7.7, D1 in "Mathematical & computational biology" (JCR-WOS).

6. Clara García-Vicente, Gonzalo C. Gutiérrez-Tobal, **Jorge Jiménez-García**, Adrián Martín-Montero, David Gozal, Roberto Hornero, "ECG-based convolutional neural network in pediatric obstructive sleep apnea diagnosis", *Computers in Biology and Medicine*, vol. 167, pp. 107628, December, 2023, DOI: 10.1016/j.compbiomed.2023.107628. Journal Impact Factor in 2022 (last year available): 7.7, D1 in "Mathematical & computational biology" (JCR-WOS).

7. **Jorge Jiménez-García**, María García, Gonzalo C. Gutiérrez-Tobal, Leila Kheirandish-Gozal, Fernando Vaquerizo-Villar, Daniel Álvarez, Félix del Campo, David Gozal, Roberto Hornero, "An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals", *Biomedical Signal Processing and Control*, vol. 87, pp. 105490, January, 2024, DOI: 10.1016/j.bspc.2023.105490. Journal Impact Factor in 2022 (last year available): 5.1, Q2 in "Engineering, Biomedical" (JCR-WOS).

A.1.2 Book chapters

1. Verónica Barroso-García, **Jorge Jiménez-García**, Gonzalo C. Gutiérrez-Tobal, Roberto Hornero, "Airflow analysis in the context of sleep apnea", in *Advances in the Diagnosis and Treatment of Sleep Apnea: Filling the Gap Between Physicians and Engineers*, pp. 241-254, Cham, Switzerland:

Springer International Publishing, Editors: Thomas Penzel and Roberto Hornero, 2022, DOI: 10.1007/978-3-031-06413-5_14.

A.1.3 International conferences

1. **Jorge Jiménez-García**, Joran Michiels, Guido Gagliardi, Gonzalo C. Gutiérrez-Tobal, María García, Clara García-Vicente, David Gozal, Maarten de Vos, Roberto Hornero, "An explainable artificial intelligence technique to interpret deep learning models aimed at detecting paediatric sleep apnoea from airflow and oximetry signals", *27th Conference of the European Sleep Research Society (ESRS2024)*, Seville (Spain), September 24 - September 27, 2024, (Under review).

2. Verónica Barroso-García, Gonzalo C. Gutiérrez-Tobal, **Jorge Jiménez-García**, Clara García-Vicente, Daniel Álvarez, David Gozal, Roberto Hornero, "Automatic detection of paediatric sleep apnoea applying deep learning and explainable artificial intelligence techniques to airflow signals", *27th Conference of the European Sleep Research Society (ESRS2024)*, Seville (Spain), September 24 - September 27, 2024, (Under review).

A.1.4 National conferences

1. **Jorge Jiménez-García**, Roberto Romero-Oraá, María García, María I. López, Roberto Hornero, "Evaluación automática de la calidad en retinografías mediante clasificación de características globales de imágenes", *XXXVI Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2018)*, ISBN: 978-84-09-06253-9, pp. 29–32, Ciudad Real (Spain), November 21 - November 23, 2018.

2. Roberto Romero-Oraá, María García, **Jorge Jiménez-García**, María I. López, Roberto Hornero, "Clasificación de superpíxeles para la detección automática de lesiones rojizas en imágenes de fondo de ojo", *XXXVI Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2018)*, ISBN: 978-84-09-06253-9, pp. 25–28, Ciudad Real (Spain), November 21 - November 23, 2018.

3. **Jorge Jiménez-García**, Gonzalo C. Gutiérrez-Tobal, María García, Daniel Álvarez, **Verónica Barroso-García**, Fernando Vaquerizo-Villar, Adrián Martín-Montero, Félix del Campo, Leila Kheirandish-Gozal, David Gozal,

Roberto Hornero, "Evaluación de la información espectral de las señales de flujo aéreo y saturación de oxígeno en sangre para la ayuda al diagnóstico de la apnea del sueño infantil", *XXXVII Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2019)*, ISBN: 978-84-09-16707-4, pp. 25–28, Santander (Spain), November 27 - November 29, 2019.

4. Adrián Martín-Montero, Gonzalo C. Gutiérrez-Tobal, Daniel Álvarez, Fernando Vaquerizo-Villar, Verónica Barroso-García, **Jorge Jiménez-García**, Leila Kheirandish-Gozal, Félix del Campo, David Gozal, Roberto Hornero, "Utilidad de nuevas bandas espectrales en la señal de HRV para ayudar en el diagnóstico de la apnea del sueño infantil", *XXXVII Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2019)*, ISBN: 978-84-09-16707-4, pp. 295–298, Santander (Spain), November 27 - November 29, 2019.

5. **Jorge Jiménez-García**, Gonzalo C. Gutiérrez-Tobal, María García, Daniel Álvarez, Adrián Martín-Montero, Félix del Campo, Leila Kheirandish-Gozal, David Gozal, Roberto Hornero, "Análisis de flujo aéreo y saturación de oxígeno en sangre mediante transformada wavelet para la detección de la apnea obstructiva del sueño infantil", *XXXVIII Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2020)*, ISBN: 978-84-09-25491-0, pp. 315–318, Valladolid (Spain), November 25 - November 27, 2020.

6. Fernando Vaquerizo-Villar, Daniel Álvarez, Gonzalo C. Gutiérrez-Tobal, **Jorge Jiménez-García**, Carmen A. Arroyo, Félix del Campo, Roberto Hornero, "Modelo de deep learning basado en la combinación de redes neuronales convolucionales y recurrentes para clasificar eventos de apnea e hipopnea mediante la señal de oximetría", *XXXIX Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2021)*, ISBN: 978-84-09-36054-3, pp. 63–66, Madrid (Spain), November 25 - November 26, 2021.

7. **Jorge Jiménez-García**, Gonzalo C. Gutiérrez-Tobal, María García, Fernando Vaquerizo-Villar, Daniel Álvarez, Félix del Campo, Leila Kheirandish-Gozal, David Gozal, Roberto Hornero, "Combinación de redes neuronales convolucionales y recurrentes para la detección de la apnea obstructiva del sueño en niños empleando las señales de flujo aéreo y oximetría", *XL Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB*

2022), ISBN: 978-84-09-45972-8, pp. 427–430, Valladolid (Spain), November 23 - November 25, 2022.

8. Clara García-Vicente, Gonzalo C. Gutiérrez-Tobal, **Jorge Jiménez-García**, Adrián Martín-Montero, David Gozal, Roberto Hornero, "ECG-ENET: Red neuronal convolucional explicable para la ayuda en el diagnóstico de la apnea del sueño infantil", *XLI Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2023)*, ISBN: 978-84-17853-76-1, pp. 14–17, Cartagena (Spain), November 22 - November 24, 2023.

A.2 International internship

Three-month research internship at the Biomedical Data Processing research team (BIOMED), STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Leuven, Belgium.

i. **Purpose of the internship**
The doctoral student conducted formative and research activities closely related to his Doctoral Thesis' topic, focused on detecting pediatric obstructive sleep apnea (OSA) using biomedical signals such as respiratory airflow and oximetry. These activities addressed the research, development and application of advanced deep learning (DL) and explainable artificial intelligence (XAI) techniques in the context of pediatric OSA detection. In this sense, the student worked on the application of DL techniques such as convolutional neural networks (CNN) and recurrent neural networks (RNN), as well as their combination with XAI algorithms derived from the Shapley Additive Explanations (SHAP) technique. During the internship, the student collaborated with the research team members in the host institution on the design and implementation of a KernelSHAP-based XAI algorithm applicable to DL models such as CNN and the combination CNN+RNN developed during the thesis. The goal of this research is to dive into the explainability of these models and enhance the interpretation of the airflow and oximetry signals' patterns relevant for the automatic detection of OSA in children.

ii. **Methodological summary**
The study conducted during the research stay addressed the development of advanced XAI techniques applicable to biomedical time series data such as

airflow and oximetry signals. A review of the state-of-the-art focused on the previous works that have designed, developed, and applied XAI algorithms in the context of time series data classification, and with a particular applicability on the analysis of biomedical signals, was performed. Then, a novel XAI methodology aimed at computing Shapley values derived from the SHAP technique was developed. XAI techniques based on SHAP/Shapley values estimate which part of a model's output (a prediction) is attributable to each part of the input (in our case, the signals). In order to address this computation, some SHAP implementations were reviewed. KernelSHAP is a SHAP implementation aimed at estimating Shapley values specifically designed to handle high dimensional data such as those used by DL models (images, text, time series, etc.). Specifically, our developed DL models were aimed at processing and analyzing raw airflow and oximetry recordings. Therefore, KernelSHAP was specifically adapted to obtain SHAP-based explanatory attributions from short segments of the input signals that clarify their contribution to OSA detection. A public database composed of biomedical signals from children with suspicion of OSA, the Childhood Adenotonsillectomy Trial (CHAT) database, was used during the development and analysis of our proposal.

iii. Quality indicators of the institution

KU Leuven is a leading university in Belgium, and one of the most relevant and prestigious European universities. According to the "Times Higher Education World University Ranking" and "Shanghai Academic Ranking of World Universities", KU Leuven is ranked 42nd and 95th respectively, standing out in relevant areas in the field of doctoral thesis research, such as "Biomedical Engineering", "Medical Technology", "Biotechnology" or "Computer Science & Engineering". In the academic year 2021-2022 it had 65,100 students, of which 7,173 were PhD students and 21% came from outside Belgium. In addition, it has a total of 14,789 employees, of whom 1,876 are full professors or professors and 2,017 are postdoctoral researchers. Its scientific output amounted to 10,425 peer-reviewed publications (incl. journal articles, book chapters and conference papers). Finally, KU Leuven has 109 ERC projects (FP7 and H2020) and 11 ERC Horizon Europe projects. The BIOMED group belongs to the Department of Electrical Engineering (ESAT), where the STADIUS center is located. This center currently has 37 active research projects, of which 15 are funded by the European Union.

Led by Professors Dr. Alexander Bertrand and Dr. Maarten de Vos, the BIOMED group currently includes 2 postdoctoral researchers and about 25 PhD students. Its research lines are developed in close collaboration with the UZ Leuven University Hospital, and are mainly divided into 6 thematic areas: portable health monitoring, sleep monitoring, neonatal monitoring, signal processing for next generation neural implants, electroencephalography in daily life, and cancer diagnosis, all of them oriented towards medical monitoring and diagnostic support. Prof. Dr. Maarten de Vos is the co-director of the BIOMED group. He is currently Full Professor at the departments of Electrical Engineering (ESAT) and Medicine at KU Leuven, has lectured at the universities of Oxford (UK) and Oldenburg (Germany). He has supervised 10 PhD theses to date, and is currently director of more than 10 PhD theses within the group related to the application of artificial intelligence techniques in medicine. He has published 137 articles in journals indexed in Journal Citation Reports, 50 communications in congresses and 3 book chapters (Scopus), obtaining 7,509/11,890 citations and an h-index of 43/53 according to Scopus/Google Scholar. During his academic career he has obtained several awards related to research projects and scientific publications.

A.3 Grants

09/2020: **"Convocatoria 2019 de contratos predoctorales de la Universidad de Valladolid"**, grant from the Universidad de Valladolid and funded by the Banco Santander. Destination place: Grupo de Ingeniería Biomédica, Universidad de Valladolid, Valladolid, Spain. Duration: October 01, 2020 – September 30, 2024.

07/2023: **"Movilidad de Doctorandos. Ayudas para estancias breves en el desarrollo de tesis doctorales (convocatoria 2023)"**, grant from the University of Valladolid. Destination place: Biomedical Data Processing research team, KU Leuven, Leuven, Belgium. Duration: September 01, 2023 – December 01, 2023.

09/2023: **"Ayudas financieras destinadas a estudiantes o recién titulados de la Universidad de Valladolid para la realización de prácticas Erasmus+ en empresas extranjeras con sede en el espacio europeo de educación superior (EEES) y países asociados del pro-**

grama durante el curso académico 2023/2024", grant from the University of Valladolid and cofunded by European Funds. Destination place: Biomedical Data Processing research team, KU Leuven, Leuven, Belgium. Duration: September 01, 2023 – December 01, 2023.

09/2023: **"Acciones de Movilidad para estancias del personal CIBER-BBN en grupos externos. Año 2023"**, "Centro de Investigación Biomédica en Red en el área temática de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)". Destination place: Biomedical Data Processing research team, KU Leuven, Leuven, Belgium. Duration: September 01, 2023 – December 01, 2023.

Apéndice B

Resumen en castellano

B.1 Introducción

La apnea obstructiva del sueño (AOS) es un trastorno del sueño causado por obstrucciones o estrechamientos recurrentes y/o intermitentes de la vía aérea superior, y que provocan pausas totales y reducciones parciales de la respiración (apneas e hipopneas, respectivamente) durante el sueño (Dehlink and Tan, 2016; Moffa et al., 2020). Estas obstrucciones interrumpen la actividad respiratoria normal, se relacionan con un mayor esfuerzo respiratorio, interrumpen el desarrollo de las diferentes fases del sueño y provocan episodios de hipoxemia (Bitners and Arens, 2020; DelRosso, 2016). Los síntomas más frecuentes de AOS en niños son ronquidos, respiración dificultosa o pausas respiratorias e hipersomnolencia diurna. Sin embargo, estos síntomas pueden ser sutiles y no detectarse fácilmente (Tauman and Gozal, 2011). Las causas más frecuentes que predisponen a la aparición del síndrome de la AOS son la hipertrofia adenoamigdalar, la obesidad, anomalías craneofaciales y trastornos neuromusculares (Moffa et al., 2020).

La AOS pediátrica afecta a entre el 1 % y el 5 % de los niños, aunque algunos estudios han estimado una prevalencia del 5,7 % (DelRosso, 2016; Marcus et al., 2012). Sin embargo, la tasa real de niños que padecen AOS puede ser mayor debido en gran medida a la dificultad para identificar y diagnosticar a los niños que la padecen (Brockmann et al., 2018; Joosten et al., 2017). Asimismo, la AOS se asocia con diversas consecuencias negativas que van desde comorbilidades cardiometabólicas, como hipertensión y dislipidemia, hasta trastornos neuroconductuales, como déficits neurocognitivos y de atención, y también hiperactividad (Marcus et al., 2012; Tauman and Gozal, 2011). El diagnóstico precoz de los niños con riesgo

de AOS es importante para poder derivar a un posible tratamiento quirúrgico o iniciar tratamiento farmacológico, evitando las posibles consecuencias negativas y otras comorbilidades graves.

El diagnóstico de la AOS infantil se ve afectado por las complicaciones de los procedimientos diagnósticos más aceptados. El *gold standard* para diagnosticar la AOS es la polisomnografía nocturna (PSG) realizada en una unidad del sueño especializada. Esta prueba consiste en el registro de señales cardiorrespiratorias, neuronales, musculares, de posición y movimiento durante la noche mientras el paciente duerme (Berry et al., 2020; Jon, 2009). Después, las señales de la PSG se analizan para localizar y cuantificar los episodios de apnea o hipopnea (Berry et al., 2020). La Academia Americana de Medicina del Sueño (*American Academy of Sleep Medicine*, AASM) define las apneas como una reducción $\geq 90\,\%$ de la señal de flujo aéreo (FA) durante al menos dos periodos respiratorios. Asimismo, las hipopneas se definen como una reducción del FA $\geq 30\,\%$ durante al menos dos ciclos respiratorios asociada a una caída de la señal de saturación de oxígeno (SpO_2) $\geq 3\,\%$ (desaturación) o a un microdespertar o *arousal* en la señal de electroencefalograma. El índice de apnea-hipopnea (IAH) se define como la tasa de eventos de apnea o hipopnea por hora (e/h) de sueño, y es el principal indicador para diagnosticar la AOS pediátrica (Bitners and Arens, 2020; Moffa et al., 2020). La gravedad de la AOS se define en función del IAH: sin AOS (IAH<1 e/h), AOS leve (1≤IAH<5 e/h), AOS moderada (5≤IAH<10 e/h) y AOS grave (IAH≥10 e/h).

A pesar de que la PSG es el estándar diagnóstico de la AOS en niños, su disponibilidad es baja debido a la escasez de unidades de sueño especializadas, su alta complejidad y los costes asociados (Dehlink and Tan, 2016; Stowe and Afolabi-Brown, 2020). Estas razones dan lugar a largas listas de espera que retrasan el diagnóstico de los niños afectados por la enfermedad. Por lo tanto, es necesario simplificar el diagnóstico de la AOS para mejorar el acceso de los niños afectados a un diagnóstico temprano y a un posible tratamiento (Brockmann et al., 2018). Algunas alternativas a la PSG comprenden el análisis de menos señales, que también pueden registrarse fuera de los laboratorios de sueño (Tan et al., 2015). De este modo, la simplificación de la PSG permitiría reducir costes, aumentar la disponibilidad y la comodidad del paciente.

Esta tesis doctoral se ha enfocado hacia la simplificación del diagnóstico de la AOS pediátrica reduciendo el número de señales necesarias. La definición de los eventos de apnea e hipopnea según las reglas de la AASM recae sobre las señales de FA y SpO_2, por lo que su análisis permite obtener la información necesaria para detectar la AOS y evaluar su severidad (Berry et al., 2012). Este análisis de las

señales de FA y SpO_2 puede simplificarse aún más mediante algoritmos automáticos de procesamiento de señales (Bertoni and Isaiah, 2019; Mazzotti et al., 2018). Estos métodos pueden también añadir una etapa de reconocimiento de patrones para detectar automáticamente signos de patología y proporcionar un diagnóstico automático y simplificado (Gutiérrez-Tobal et al., 2022; Uddin et al., 2018). En esta tesis doctoral se han utilizado diferentes algoritmos avanzados de aprendizaje automático (*machine learning*, ML) como *ensemble learning* y *deep learning* (DL) combinados con métodos de inteligencia artificial explicable (*eXplainable artificial intelligence*, XAI).

Esta tesis doctoral se centró en el análisis automático de las señales de FA y SpO_2 para facilitar el diagnóstico de la AOS pediátrica. Se investigaron, desarrollaron y aplicaron varios algoritmos para procesar y caracterizar estas señales, extraer información útil relacionada con la enfermedad y por último detectar la AOS y estimar su severidad. Se abordaron diferentes enfoques, que cubrieron diversas técnicas de ML como ensemble learning, DL y XAI. Los resultados obtenidos en los diferentes estudios dieron lugar a la publicación de tres artículos en revistas, todas ellas indexadas en Journal Citation Reports (JCR) de Web of Science™(WOS). Así, esta tesis doctoral se presenta como un compendio de publicaciones.

B.2 Hipótesis y objetivos

Las alternativas a la PSG destinadas a diagnosticar la AOS mediante un número reducido de señales han incluido habitualmente el FA y la SpO_2, ya que proporcionan información suficiente para localizar las apneas, hipopneas y desaturaciones asociadas a las mismas (Alonso-Álvarez et al., 2015; Kaditis et al., 2016a). Los modelos de ML podrían por lo tanto combinar ambas fuentes de información para mejorar su capacidad diagnóstica y además reducir el número de señales a analizar. Se ha considerado en estudios previos que la información del FA puede complementarse con el ODI 3 % (Barroso-García et al., 2020). En este trabajo se asumió que *la información complementaria del FA y la SpO_2 es suficiente para desarrollar soluciones basadas en ML para ayudar a diagnosticar la AOS pediátrica.*

Los métodos de *ensemble learning* no se han probado ampliamente en el contexto de la AOS infantil, y los enfoques existentes están limitados al análisis de características clínicas, derivadas de la oximetría, o de la actigrafía (Bertoni et al., 2020; Calderón et al., 2020). Además, modelos de *ensemble learning* como AdaBoost se probaron con éxito en adultos (Gutierrez-Tobal et al., 2016; Gutiérrez-Tobal et al., 2019). La investigación llevada a cabo en esta tesis asumió que *los*

algoritmos de ensemble learning pueden ayudar a mejorar la precisión diagnóstica de los enfoques actuales basados en ML en la AOS pediátrica.

A pesar del potencial mostrado por los modelos de ML basados en características, están limitados por la capacidad de los expertos para obtener y analizar descriptores útiles. Los algoritmos de DL pueden superar esta limitación, ya que aprenden características complejas con un alto nivel de abstracción directamente a partir de los datos en crudo (Lecun et al., 2015). Esta investigación se llevó a cabo asumiendo que *los enfoques basados en DL pueden aprender la información necesaria para detectar AOS pediátrica directamente a partir de las señales de FA y SpO_2 en crudo.*

El uso de métodos de XAI es cada vez más común en el ámbito médico para encontrar los signos de patología aprendidos por los modelos e interpretar estos patrones de las señales (Loh et al., 2022). Esto contribuye a aumentar la confianza de los usuarios en estas ayudas diagnósticas basadas en inteligencia artificial (*artificial intelligence*, AI). Las arquitecturas de DL propuestas en esta tesis se interpretaron considerando que *los métodos de XAI pueden ayudar a identificar patrones relevantes vinculados a la presencia de AOS en las señales de FA y SpO_2 de pacientes pediátricos.*

La investigación llevada a cabo a lo largo de esta tesis doctoral se realizó asumiendo la hipótesis general de que *el análisis automático de las señales nocturnas de FA y SpO_2 mediante técnicas avanzadas de ML como ensemble learning, DL y XAI puede ayudar a simplificar el diagnóstico de la AOS infantil.*

El objetivo principal de esta tesis doctoral fue *estudiar, desarrollar y validar métodos avanzados de ML como ensemble learning o DL junto con nuevas técnicas de XAI en el contexto del análisis automático de señales de FA y SpO_2, de forma que estos métodos puedan ser utilizados para ayudar al diagnóstico de la AOS pediátrica.* Para alcanzar este objetivo principal, se plantearon los siguientes objetivos específicos:

I. Elaborar y analizar una base de datos de registros nocturnos de FA y SpO_2 procedentes de PSG realizadas a sujetos pediátricos con sospecha de AOS, incluyendo sus datos sociodemográficos y clínicos relacionados con la presencia y severidad de la enfermedad.

II. Evaluar la complementariedad de la información extraída de las señales de FA y SpO_2 mediante métodos de *feature engineering*, así como DL para mejorar su rendimiento diagnóstico individual, utilizando métodos de reconocimiento

de patrones de clasificación y regresión.

III. Evaluar la capacidad diagnóstica de los métodos de ensemble y DL entrenados con información relevante y no redundante de FA y SpO_2, así como con señales de FA y SpO_2 en crudo, respectivamente, todos ellos dirigidos a estimar el IAH y clasificar la gravedad de la AOS a partir de estos registros nocturnos.

IV. Identificar los patrones de FA y SpO_2 más relevantes que los métodos de DL relacionan con la presencia de apneas y/o hipopneas, y utilizan para detectar la AOS, mediante técnicas de XAI.

B.3 Materiales y métodos

B.3.1 Bases de datos

Para llevar a cabo esta investigación, se han utilizado dos bases de datos de señales nocturnas de FA y SpO_2 procedentes de un total de 2.612 estudios del sueño. La primera de ellas fue proporcionada por el *Comer Children's Hospital*, *University of Chicago* (UofC) *School of Medicine* (Chicago, IL, EE.UU.). Esta base de datos contenía las PSGs de 974 sujetos de hasta 13 años con síntomas de la AOS (ronquidos, pausas respiratorias durante el sueño, despertares durante la noche, hipersomnolencia, etc.) que fueron derivados a la unidad del sueño de este hospital (Hornero et al., 2017). Los estudios de sueño se realizaron con un equipo de PSG Polysmith® (Nihon Kohden America Inc., Irvine, CA, USA) y fueron analizados de acuerdo con las reglas establecidas por la AASM para diagnosticar los sujetos (Berry et al., 2012). Los 974 registros fueron separados de manera aleatoria en dos conjuntos independientes de entrenamiento (584) y test (390). Esta base de datos se utilizó en los tres artículos que componen esta tesis doctoral.

La segunda base de datos empleada en esta tesis doctoral se obtuvo del *Childhood Adenotonsillectomy Trial* (CHAT). Esta base de datos pública y multicéntrica proporcionada por el *National Sleep Research Resource* contenía un total de 1.638 estudios del sueño realizados a niños entre 5 y 10 años de edad con síntomas de AOS. Estos estudios fueron utilizados para llevar a cabo un estudio aleatorizado para comprobar la efectividad de la adenoamigdalectomía frente a un tratamiento conservador (Marcus et al., 2013; Redline et al., 2011). Seis centros hospitalarios distintos en EE.UU. estuvieron involucrados en la obtención de datos, y las PSG fueron analizadas de acuerdo al protocolo descrito por Marcus

et al. (2013). De los 453 sujetos incluidos inicialmente en el estudio aleatorizado (*baseline*), 406 fueron evaluados de nuevo 7 meses después para comprobar su evolución (*follow-up*). Otros 779 sujetos fueron evaluados, pero no formaron parte del estudio aleatorizado (*non randomized*). Los sujetos de esta base de datos fueron distribuidos aleatoriamente en grupos de entrenamiento (1006), validación (326) y test (306). Esta base de datos se utilizó en el segundo y tercer artículo del compendio de publicaciones, los que presentan las dos arquitecturas de DL empleadas en esta tesis doctoral.

Las señales de FA se registraron como parte de las señales que componían las PSGs por medio de un termistor, con frecuencias de muestreo entre 20 y 512 Hz. Por otra parte, las señales de SpO_2 también formaban parte de las PSGs y fueron registradas mediante un pulsioxímetro colocado en un dedo del paciente a tasas de muestreo entre 1 y 512 Hz.

B.3.2 Metodología

La metodología global aplicada en esta tesis doctoral se componía de 5 etapas: (i) preprocesado de las señales, (ii) caracterización, (iii) selección de características, (iv) clasificación de características y (v) aplicación de modelos de DL y XAI sobre las señales mínimamente preprocesadas. La etapa de preprocesado (i) se implementó como paso previo al análisis de las señales. Después, la metodología desplegada se dividió en dos ramas principales: *feature engineering* (ii, iii) junto a *ensemble learning* (iv), y DL con XAI (v). Todas las metodologías propuestas en esta tesis doctoral fueron evaluadas mediante su capacidad diagnóstica, es decir, la precisión y fiabilidad de los modelos automáticos de *ensemble learning* y DL para detectar la presencia y la severidad de la AOS pediátrica.

Las señales de FA y SpO_2 se preprocesaron mínimamente para remuestrear cada registro nocturno a una frecuencia de muestreo común, reducir ruido mediante filtrado, normalizar su amplitud y eliminar artefactos. Este ultimo paso no fue aplicado en los estudios abarcan técnicas de DL. Concretamente, las señales de FA fueron filtradas con un filtro paso bajo entre 0 y 1.5 Hz y después se aplicó un método de normalización adaptativa de la amplitud (Várady et al., 2002). Con respecto a las señales de SpO_2, sus valores de amplitud fueron normalizados sustrayendo la media y dividiendo el resultado por la desviación estándar como paso previo a la aplicación de métodos de DL.

La rama de la metodología referente al enfoque de *feature engineering* abarcó etapas de extracción, selección y clasificación de características para estimar la

presencia de AOS y su gravedad a partir de la información más relevante y complementaria extraída del FA y la SpO_2 (Jiménez-García et al., 2020). Se calcularon parámetros de las señales como los momentos estadísticos en el dominio temporal, así como características espectrales a partir de la densidad espectral de potencia de las mismas. Así mismo, se extrajeron características de las bandas de interés de cada señal (0.134–0.176 Hz en el FA y 0.020–0.044 Hz en la SpO_2) (Jiménez-García et al., 2020). También se calcularon parámetros no lineales como la medida de tendencia central (*central tendency measure*, CTM), la complejidad de Lempel-Ziv (*Lempel-Ziv complexity*, LZC) y la entropía muestral (*sample entropy*, SampEn) a partir del FA y la SpO_2, y finalmente se incluyó en los análisis el índice de desaturaciones del 3 % (*oxygen desaturation index*, ODI 3 %) (Jiménez-García et al., 2020). A continuación, la fase de selección de características estaba orientada a obtener subconjuntos de características relevantes y no redundantes de cada una de las señales y de manera conjunta (Jiménez-García et al., 2020). Para lograrlo se aplicó el algoritmo *Fast Correlation-Based Filter* (FCBF) a las características extraídas del FA y la SpO_2 por separado y de manera conjunta, diferenciando también entre incluir el ODI 3 % y excluirlo como predictor.

La etapa de clasificación se basó en la implementación de modelos de *ensemble learning* para estimar el nivel de severidad de la AOS a partir de los subconjuntos características relevantes y complementarias de las diferentes señales (Jiménez-García et al., 2020). Para ello se empleó AdaBoost, un algoritmo de clasificación de tipo *boosting* (Freund and Schapire, 1997). AdaBoost asigna al patrón descriptivo de cada sujeto a un nivel de severidad de la AOS mediante el voto por mayoría ponderado de una gran cantidad de clasificadores de tipo *Linear Discriminant Analysis* (LDA). Cada uno de estos sencillos modelos fueron ajustados secuencialmente con distintas representaciones de los datos de entrenamiento, dando mayor importancia a las instancias falladas en iteraciones pasadas (Freund and Schapire, 1997).

La otra rama en la que se dividió la metodología propuesta en esta tesis doctoral incluía el desarrollo de diferentes arquitecturas de DL destinadas a detectar y categorizar la severidad de la AOS mediante la estimación del IAH. Los algoritmos de DL son capaces de aprender representaciones con un elevado nivel de abstracción directamente sobre datos en crudo, prescindiendo así del enfoque basado en la caracterización típico de los métodos clásicos de ML (Lecun et al., 2015). Este tipo de arquitecturas han superado la capacidad diagnóstica mostrada por los algoritmos clásicos de ML en contextos como el análisis de señales biomédicas, y más concretamente, la detección de la AOS (Faust et al., 2018; Mostafa et al.,

2019). En esta investigación, se han propuesto dos arquitecturas de DL para analizar las señales de FA y SpO_2 simultáneamente y estimar el IAH (Jiménez-García et al., 2022, 2024). La primera de ellas consistió en una red neuronal convolucional (*convolutional neural network*, CNN), formada por sucesivas capas de convolución 2D, normalización, activación, reducción de dimensionalidad (*pooling*) y regularización mediante *dropout* (Goodfellow et al., 2016). La estimación final se obtuvo por medio de una capa de tipo *fully-connected*, que proporciona una estimación del número total de episodios de apnea detectados por cada época de 5 minutos de ambas señales (Jiménez-García et al., 2022). El segundo modelo surgió de la combinación del modelo CNN propuesto anteriormente con una red neuronal recurrente (*recurrent neural network*, RNN), destinado también a detectar la AOS y estimar su severidad por medio del IAH (Jiménez-García et al., 2024). Las capas del modelo CNN original fueron trasladadas al modelo CNN+RNN mediante un enfoque de *transfer learning* (Jiménez-García et al., 2024). Después, una RNN fue implementada mediante una capa de tipo *Bidirectional Gated Recurrent Unit* (Bi-GRU) para analizar la estructura temporal de la información de las señales procesada en las capas de la CNN y estimar el IAH a partir del total de eventos detectados por cada segmento de 30 minutos de las señales de FA y SpO_2.

Además, la explicabilidad del modelo CNN+RNN desarrollado se abordó con el algoritmo Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). Este es un método de XAI destinado a obtener explicaciones sobre las predicciones de los modelos de DL en forma de mapas de calor (o *heatmaps*) asociados a las predicciones del modelo, señalando los patrones más relevantes de las señales según el modelo para detectar signos de la AOS (Jiménez-García et al., 2024). De esta forma se interpretaron las estimaciones, tanto correctas como erróneas, que el modelo CNN+RNN proporciona sobre de la cantidad de eventos de apnea/hipopnea presentes en las señales y se identificaron los patrones más relevantes del FA y la SpO_2 relacionados con la AOS (Jiménez-García et al., 2024).

Por último, se aplicaron diversas técnicas y análisis estadísticos para evaluar los resultados producidos en cada una de estas etapas. Se emplearon tests estadísticos para evaluar diferencias entre poblaciones de sujetos, y coeficientes de correlación para evaluar la asociación de las características extraídas con el IAH. Se usaron métricas de concordancia, así como diferentes medidas de rendimiento diagnóstico para evaluar el desempeño de los métodos propuestos para detectar la AOS y establecer su severidad, empleando para ello diferentes estrategias de validación.

B.4 Resultados y discusión

Los resultados obtenidos empleando las diferentes metodologías presentadas en esta tesis doctoral permitieron evaluar de manera precisa la presencia y severidad de la AOS infantil. Mediante el enfoque de *feature engineering* se reveló que las señales de FA y SpO_2 presentaban características relevantes y complementarias según los resultados obtenidos utilizando el algoritmo de selección de características FCBF. El ODI 3 % obtenido de la SpO_2 fue la característica más relevante y dominante, mientras que la CTM calculada a partir del FA también resultó relevante y no redundante con el ODI 3 % (Jiménez-García et al., 2020). Estos resultados sugieren que las señales de FA y SpO_2 son complementarias y podrían ayudar a diagnosticar la AOS pediátrica de manera conjunta mediante reconocimiento de patrones. De hecho, esta combinación de señales, junto al clasificador AdaBoost, obtuvieron la mayor concordancia al estimar la severidad de la AOS en términos de precisión de 4 clases ($Acc_4 = 57,95\,\%$) y kappa de Cohen ($k = 0,3984$) (Jiménez-García et al., 2020). La combinación del FA con el ODI 3 % obtuvo resultados muy similares ($Acc_4 = 57,95\,\%$, $k = 0,3930$). Estas métricas logradas de manera conjunta fueron superiores a las obtenidas con los enfoques que solo incluían la información de una de estas señales. No obstante, el ODI 3 % fue la característica que más contribuyó a mejorar la detección de la AOS mediante AdaBoost. La combinación de FA con ODI 3 % alcanzó el máximo rendimiento para diagnosticar la AOS pediátrica en términos de exactitud (Acc), sensibilidad (Se) y especificidad (Sp) en los umbrales de IAH de 1 e/h ($Acc = 81,28\,\%$, $Se = 92,06\,\%$, $Sp = 36,00\,\%$), 5e/h ($Acc = 82,05\,\%$, $Se = 76,03\,\%$, $Sp = 85,66\,\%$) y 10e/h ($Acc = 90,26\,\%$, $Se = 62,65\,\%$, $Sp = 97,72\,\%$) (Jiménez-García et al., 2020). En general, los modelos AdaBoost multiclase alcanzaron un rendimiento diagnóstico notable en comparación con otros enfoques que también combinaban FA con ODI 3 % (Barroso-García et al., 2020, 2021a,b). Estos resultados refuerzan la idea de que las señales de FA y SpO_2 son complementarias y útiles para detectar la AOS pediátrica.

En lo que respecta a los enfoques basados en DL, estas arquitecturas superaron a los métodos clásicos de ML que se habían aplicado previamente. Además, el uso de las dos señales también mejoró el rendimiento diagnóstico con respecto a un modelo de CNN que solo empleaba la señal de SpO_2 (Vaquerizo-Villar et al., 2021). De esta forma, se confirma que un enfoque CNN de doble canal usando el FA y la SpO_2 puede ser ventajoso frente a utilizar solo una de ellas (Jiménez-García et al., 2022). Además, este modelo CNN de doble canal se completó con una

RNN que permitió obtener una arquitectura CNN+RNN más sofisticada gracias al enfoque de *transfer learning*. El modelo CNN+RNN resultante ha permitido aprovechar la capacidad de reconocimiento de patrones de la CNN anterior y la capacidad de modelar la estructura temporal de los datos de entrada, formados por largas secuencias de 30 minutos de ambas señales (Jiménez-García et al., 2024). Además, gracias al uso de *transfer learning* se optimizó el proceso de entrenamiento y validación del modelo CNN+RNN. Hasta donde tenemos conocimiento, este tipo de arquitectura no se había propuesto hasta ahora en el contexto de la AOS infantil.

El modelo CNN+RNN se combinó con un algoritmo de XAI basado en Grad-CAM que permitió resaltar los patrones más relevantes de ambas señales relacionados con la detección de la AOS (Jiménez-García et al., 2024). Esto contribuyó a aumentar la explicabilidad de las predicciones realizadas por este modelo, así como a descubrir e interpretar dichos patrones en las señales. Los cambios instantáneos en la amplitud de la señal de FA y las interrupciones del patrón cíclico de la respiración son algunas de las características señaladas por el método de Grad-CAM, así como variaciones repentinas que podrían coincidir con *arousals*. Con respecto a la señal de SpO_2, las caídas de los niveles de saturación como consecuencia de los eventos apneicos, así como las posteriores recuperaciones son los patrones que el modelo CNN+RNN más frecuentemente asocia a la presencia de apneas/hipopneas. Todos estos patrones son útiles para comprender y verificar el funcionamiento de los modelos de DL, así como reforzar la confianza de los usuarios en este tipo de algoritmos. Sin embargo, también se ha observado que a menudo no se señalan hipopneas no asociadas a desaturaciones, sino a *arousals*, lo que podría explicar la infraestimación del modelo CNN+RNN de este tipo de eventos respiratorios (Jiménez-García et al., 2024). También se ha observado que a veces Grad-CAM señala ciertos artefactos en ambas señales como patrones que el modelo CNN+RNN identificó como un signo de la AOS. Finalmente, también se identificaron mediante Grad-CAM algunas desaturaciones que no se asociaron a eventos de apnea por la eventual falta de calidad de la señal de FA. En estos casos, los heatmaps explicativos pueden ayudar a los especialistas a revisar las señales para confirmar la presencia de signos de patología (Jiménez-García et al., 2024). El análisis de estos *heatmaps* además puede ayudar a descubrir información nueva sobre la AOS más allá de las reducciones del FA y las desaturaciones. Hasta donde sabemos, este enfoque de XAI no se había aplicado hasta ahora en el contexto de la AOS infantil.

Con respecto a la concordancia y el rendimiento diagnóstico de los modelos de DL, estas arquitecturas superaron claramente a los anteriores enfoques de

feature engineering (Jiménez-García et al., 2022, 2024). La concordancia entre el IAH real y estimado, calculada mediante el coeficiente de correlación intraclase (ICC), resultó ser muy elevada en los conjuntos de test de ambas bases de datos. Se alcanzaron ICC más altos en los datos de CHAT ($ICC = 0,9546$ con CNN; $ICC = 0,9465$ con CNN+RNN) que en los UofC ($ICC = 0,8821$ con CNN; $ICC = 0,9004$ con CNN+RNN). Estos elevados resultados también se vieron reflejados en los valores de Acc_4 y k. La concordancia fue mayor en CHAT ($Acc_4 = 72,55\,\%$, $k = 0,6011$ con CNN; $Acc_4 = 74,51\,\%$, $k = 0,6231$ con CNN+RNN), mientras que estos resultados superaron a los alcanzados con los modelos AdaBoost al compararlos en la base de datos UofC ($Acc_4 = 61,79\,\%$, $k = 0,4469$ con CNN; $Acc_4 = 62,31\,\%$, $k = 0,4495$ con CNN+RNN). Comparando estos modelos, se comprobó que CNN+RNN superó a CNN tanto en concordancia como en rendimiento diagnóstico. Se obtuvo una capacidad superior para diagnosticar la AOS en todos los puntos de corte del IAH usando CNN+RNN, con unos resultados diagnósticos muy altos en 1 e/h ($Acc = 87,25\,\%$, $Se = 87,03\,\%$, $Sp = 88,06\,\%$), 5 e/h ($Acc = 93,46\,\%$, $Se = 80,22\,\%$, $Sp = 99,07\,\%$) y 10 e/h ($Acc = 93,46\,\%$, $Se = 71,43\,\%$, $Sp = 96,97\,\%$) en el conjunto de test de CHAT. Por otra parte, estas métricas también fueron elevadas en el conjunto de test UofC en 1 e/h ($Acc = 84,10\,\%$, $Se = 96,83\,\%$, $Sp = 30,67\,\%$), 5 e/h ($Acc = 84,62\,\%$, $Se = 82,88\,\%$, $Sp = 85,66\,\%$) y 10 e/h ($Acc = 90,51\,\%$, $Se = 78,31\,\%$, $Sp = 93,81\,\%$).

Estos resultados indican que el modelo CNN+RNN es el más preciso entre todos los enfoques abordados en esta tesis doctoral. Por su parte, el modelo CNN entrenado con ambas señales superó a un enfoque anterior muy similar empleando únicamente la SpO_2 como fuente de datos (Jiménez-García et al., 2022; Vaquerizo-Villar et al., 2021). Al compararlos, se observó que este enfoque de doble canal obtuvo mejores resultados al diagnosticar la AOS en los umbrales de 1 y 5 e/h, lo que indicaría que el FA contribuye de forma notable a mejorar la utilidad de un modelo basado en CNN. Por su parte, la arquitectura CNN+RNN no solo superó al modelo CNN del que deriva, sino a todos los enfoques anteriores centrados en la detección de la AOS pediátrica (Jiménez-García et al., 2024). Esto pone de manifiesto la utilidad de una arquitectura de DL que combina diferentes técnicas de análisis para aprender automáticamente las particularidades de la AOS infantil a través de señales de FA y SpO_2 y que además es interpretable. A la vista de estos resultados, los métodos propuestos en esta tesis doctoral podrían utilizarse para desarrollar un protocolo de cribado de la AOS pediátrica en entornos en los que la PSG hospitalaria no tiene una alta disponibilidad. Este protocolo consistiría en

analizar únicamente los registros nocturnos de FA y SpO_2 registradas en la casa del paciente por medio de la arquitectura CNN+RNN. Dependiendo del diagnóstico preliminar proporcionado por el modelo, se podría:

1. Si el modelo predice IAH<1 e/h: no se necesitaría realizar una PSG para confirmar el diagnóstico, y se podría descartar la cirugía. Solo el 1,63 % de niños que obtuvieron IAH<1 e/h según el modelo automático realmente tenían IAH≥5 e/h, que es el umbral para considerar un tratamiento quirúrgico. No obstante, se recomendaría a los padres o tutores legales vigilar los síntomas y ponerlos en conocimiento del personal médico para una posible revaluación.

2. Si el modelo predice 1≤IAH<5 e/h: iniciar un tratamiento no quirúrgico de los síntomas de la AOS infantil (p.ej., reducción de peso, tratamiento antiinflamatorio, etc.). Además, realizar un seguimiento de los pacientes y revaluar los síntomas periódicamente. La cirugía se podría descartar inicialmente, ya que solo el 11,99 % de los niños con 1≤IAH<5 según el modelo realmente tenían AOS moderada o severa. Si los síntomas persistieran, recomendar la realización de una PSG en el hospital para confirmar la necesidad de tratamiento quirúrgico.

3. Si el modelo predice 5≤IAH<10: recomendar la realización de una PSG para confirmar el diagnóstico preliminar de AOS moderada, ya que el 69,72 % de los pacientes con este diagnóstico preliminar realmente tenían AOS moderada o severa. Dependiendo del resultado de la PSG, considerar el tratamiento quirúrgico.

4. Si el algoritmo predice IAH≥10 e/h: considerar la posibilidad de tratamiento quirúrgico directamente sin realizar la PSG, ya que solo el 3,28 % de los pacientes que obtuvieron IAH≥10 e/h según el modelo en realidad tenían IAH<5 e/h. Si los cirujanos no consideran el caso como adecuado operar, proponer un tratamiento farmacológico.

Según los resultados diagnósticos del modelo CNN+RNN, el uso de este protocolo de cribado evitaría la realización de hasta el 78,45 % de las PSG realizadas a pacientes con síntomas de AOS infantil. De esta manera se reduciría drásticamente la carga de trabajo de los médicos a la hora de analizar las señales, así como las listas de espera de las unidades del sueño pediátricas.

B.5 Conclusiones

A la vista de los resultados obtenidos en la investigación llevada a cabo, se pueden extraer las siguientes conclusiones:

1) El análisis conjunto de las señales de FA y SpO_2 es útil para detectar automáticamente la AOS pediátrica. En este sentido, las características relevantes y complementarias de ambas señales resultaron útiles para obtener modelos de clasificación automática. El CTM obtenido del FA y el ODI 3 % de la SpO_2 fueron las características más útiles de estas señales, que demostraron su relevancia para detectar la AOS pediátrica y su no redundancia entre sí.

2) Los clasificadores multiclase AdaBoost fueron capaces de detectar con precisión la gravedad de la AOS pediátrica a partir del FA y la SpO_2. Los clasificadores lograron un notable rendimiento diagnóstico a partir de un conjunto de características complementarias de las señales de FA y SpO_2.

3) CTM y ODI 3 % son las características más útiles de FA y SpO_2, respectivamente. La combinación de estas características mediante AdaBoost mejoró y maximizó la capacidad diagnóstica de este algoritmo de clasificación en comparación con otras combinaciones que implicaban el uso de una sola señal.

4) Un enfoque de doble canal para detectar la AOS pediátrica que incluya señales de FA y SpO_2 resulta ventajoso frente a otros enfoques de un solo canal. Tanto los modelos de ensemble learning como de DL lograron un mayor rendimiento diagnóstico con 2 señales en comparación con otros enfoques que sólo incluían una señal.

5) Una arquitectura CNN 2D fue útil para procesar y analizar las señales de FA y SpO_2 y para estimar el IAH a partir de estos registros nocturnos. El enfoque propuesto basado en CNN logró un notable rendimiento diagnóstico para detectar la AOS pediátrica, especialmente al diferenciar entre sujetos de control y sujetos con AOS leve, así como al discriminar entre pacientes con AOS leve y moderada. La CNN propuesta basada en FA y SpO_2 superó el rendimiento diagnóstico de los modelos de DL anteriores, orientados a analizar las señales de SpO_2 únicamente.

6) La combinación de CNN con RNN fue útil para detectar la AOS pediátrica a partir del FA y la SpO_2. En este sentido, el uso de un enfoque de *transfer*

learning para desarrollar un modelo CNN+RNN más sofisticado demostró su utilidad. Esta novedosa arquitectura superó también a modelos anteriores basados en CNN.

7) La arquitectura CNN+RNN orientada a analizar el FA y la SpO_2 es el modelo de DL más preciso para estimar la severidad de la AOS mediante el IAH. Hasta la fecha, el rendimiento de este algoritmo es el más avanzado entre los métodos de detección automática de la AOS pediátrica mediante un conjunto reducido de señales biomédicas, ya que alcanzó la mayor capacidad diagnóstica en comparación con enfoques similares dirigidos a la misma población.

8) El uso de métodos XAI como Grad-CAM permitió explorar patrones relacionados con la AOS en las señales de FA y SpO_2. Los *heatmaps* explicativos resaltaron partes específicas de las señales de entrada que fueron relevantes para que el modelo CNN+RNN realizara sus predicciones, y permitieron interpretar el funcionamiento de esta compleja arquitectura.

9) Los *heatmaps* sobre las señales obtenidos de la CNN+RNN mediante Grad-CAM mostraron la capacidad de justificar por qué el modelo CNN+RNN predecía la presencia de apneas o hipopneas en las señales de entrada. Por lo tanto, Grad-CAM contribuyó a mejorar su veracidad diagnóstica.

10) Los *heatmaps* explicativos revelaron patrones relevantes relacionados con la AOS aprendidos por el algoritmo CNN+RNN. Los patrones destacados en las señales de FA y SpO_2 mediante Grad-CAM estaban relacionados principalmente con desaturaciones y cambios repentinos en la amplitud de las ondas respiratorias. Estos *heatmaps* explicativos podrían ser útiles para que los médicos analicen e interpreten estas señales con el objetivo de simplificar el diagnóstico de la AOS pediátrica.

La conclusión global de esta tesis doctoral es que el procesamiento y análisis automático de señales de FA y SpO_2 basado en métodos de ensemble learning y DL combinados con XAI propuesto en esta investigación tiene una gran utilidad diagnóstica, y pueden utilizarse para implementar métodos de cribado alternativos, sencillos, fiables y veraces que sirvan de ayuda en el diagnóstico de la AOS en niños.

Bibliography

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160.

Aggarwal, C. C., 2021. An introduction to artificial intelligence. In: Artificial Intelligence: A Textbook. Springer International Publishing, Cham, Ch. 1, pp. 1–34.

Allen, J., 2007. Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement 28 (3).

Alonso-Álvarez, M. L., Terán-Santos, J., Ordax Carbajo, E., Cordero-Guevara, J. A., Navazo-Egüia, A. I., Kheirandish-Gozal, L., Gozal, D., apr 2015. Reliability of Home Respiratory Polygraphy for the Diagnosis of Sleep Apnea in Children. Chest 147 (4), 1020–1028.

Álvarez, D., Hornero, R., Marcos, J. V., del Campo, F., dec 2010. Multivariate Analysis of Blood Oxygen Saturation Recordings in Obstructive Sleep Apnea Diagnosis. IEEE Transactions on Biomedical Engineering 57 (12), 2816–2824.

Álvarez, D., Crespo, A., Vaquerizo-Villar, F., Gutiérrez-Tobal, G. C., Cerezo-Hernández, A., Barroso-García, V., Ansermino, J. M., Dumont, G. A., Hornero, R., del Campo, F., Garde, A., oct 2018. Symbolic dynamics to enhance diagnostic ability of portable oximetry from the Phone Oximeter in the detection of paediatric sleep apnoea. Physiological Measurement 39 (10), 104002.

Álvarez, D., Cerezo-Hernández, A., Crespo, A., Gutiérrez-Tobal, G. C., Vaquerizo-Villar, F., Barroso-García, V., Moreno, F., Arroyo, C. A., Ruiz, T., Hornero, R., del Campo, F., 2020. A machine learning-based test for adult sleep apnoea screening at home using oximetry and airflow. Scientific Reports 10 (5332), 1–12.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58 (October 2019), 82–115.

Barroso-García, V., Gutiérrez-Tobal, G., Kheirandish-Gozal, L., Álvarez, D., Vaquerizo-Villar, F., Crespo, A., del Campo, F., Gozal, D., Hornero, R., aug 2017. Irregularity and Variability Analysis of Airflow Recordings to Facilitate the Diagnosis of Paediatric Sleep Apnoea-Hypopnoea Syndrome. Entropy 19 (9), 447.

Barroso-García, V., Gutiérrez-Tobal, G. C., Kheirandish-Gozal, L., Álvarez, D., Vaquerizo-Villar, F., Núñez, P., del Campo, F., Gozal, D., Hornero, R., 2020. Usefulness of recurrence plots from airflow recordings to aid in paediatric sleep apnoea diagnosis. Computer Methods and Programs in Biomedicine 183, 105083.

Barroso-García, V., Gutiérrez-Tobal, G. C., Kheirandish-Gozal, L., Vaquerizo-Villar, F., Álvarez, D., del Campo, F., Gozal, D., Hornero, R., 2021a. Bispectral analysis of overnight airflow to improve the pediatric sleep apnea diagnosis. Computers in Biology and Medicine 129 (August 2020).

Barroso-García, V., Gutiérrez-Tobal, G. C., Gozal, D., Vaquerizo-Villar, F., Álvarez, D., Del Campo, F., Kheirandish-Gozal, L., Hornero, R., 2021b. Wavelet analysis of overnight airflow to detect obstructive sleep apnea in children. Sensors 21 (4), 1–19.

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., Others, Redline, S., Strohl, K. P., Ward, S. L. D., Tangredi, M. M., oct 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 (5), 597–619.

Berry, R. B., Quan, S. F., Abreu, A., Medicine, e. a. f. t. A. A. o. S., 2020. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6. American Academy of Sleep Medicine, Darien, IL.

Bertoni, D., Isaiah, A., 2019. Towards Patient-centered Diagnosis of Pediatric Obstructive Sleep Apnea—A Review of Biomedical Engineering Strategies. Expert Review of Medical Devices 16 (7), 617–629.

Bertoni, D., Sterni, L. M., Pereira, K. D., Das, G., Isaiah, A., 2020. Predicting polysomnographic severity thresholds in children using machine learning. Pediatric Research 88 (3), 404–411.

Biswal, S., Sun, H., Goparaju, B., Brandon Westover, M., Sun, J., Bianchi, M. T., 2018. Expert-level sleep scoring with deep neural networks. Journal of the American Medical Informatics Association 25 (12), 1643–1650.

Bitners, A. C., Arens, R., apr 2020. Evaluation and Management of Children with Obstructive Sleep Apnea Syndrome. Lung 198 (2), 257–270.

Bland, J. M., Altman, D. G., feb 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 327 (8476), 307–310.

Blechner, M., Williamson, A. A., 2016. Consequences of Obstructive Sleep Apnea in Children. Current Problems in Pediatric and Adolescent Health Care 46 (1), 19–26.

Borrelli, M., Corcione, A., Cimbalo, C., Annunziata, A., Basilicata, S., Fiorentino, G., Santamaria, F., aug 2023. Diagnosis of Paediatric Obstructive Sleep-Disordered Breathing beyond Polysomnography. Children 10 (8), 1331.

Brockmann, P. E., Alonso-Álvarez, M. L., Gozal, D., jun 2018. Diagnóstico del síndrome de apnea hipopnea del sueño en niños: pasado, presente y futuro. Archivos de Bronconeumología 54 (6), 303–305.

Brodsky, L., dec 1989. Modern Assessment of Tonsils and Adenoids. Pediatric Clinics of North America 36 (6), 1551–1569.

Bronzino, J. D., 2006. The biomedical engineering handbook. CRC Press - Taylor & Francis Group.

Brouillette, R. T., Morielli, A., Leimanis, A., Waters, K. A., Luciano, R., Ducharme, F. M., feb 2000. Nocturnal Pulse Oximetry as an Abbreviated Testing Modality for Pediatric Obstructive Sleep Apnea. Pediatrics 105 (2), 405–412.

Calderón, J. M., Álvarez-Pitti, J., Cuenca, I., Ponce, F., Redon, P., 2020. Development of a minimally invasive screening tool to identify obese Pediatric population at risk of obstructive sleep Apnea/Hypopnea syndrome. Bioengineering 7 (4), 1–13.

Callén Blecua, M. T., 2017. Amígdalas grandes ¿hay que operarlas?

Chan, E. D., Chan, M. M., Chan, M. M., jun 2013. Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations. Respiratory Medicine 107 (6), 789–799.

Chang, L., Wu, J., Cao, L., mar 2013. Combination of symptoms and oxygen desaturation index in predicting childhood obstructive sleep apnea. International Journal of Pediatric Otorhinolaryngology 77 (3), 365–371.

Chen, C.-C., Barnhart, H. X., dec 2008. Comparison of ICC and CCC for assessing agreement for data without and with replications. Computational Statistics & Data Analysis 53 (2), 554–564.

Chiner, E., Cánovas, C., Molina, V., Sancho-Chust, J. N., Vañes, S., Pastor, E., Martinez-Garcia, M. A., 2020. Home Respiratory Polygraphy is Useful in the Diagnosis of Childhood Obstructive Sleep Apnea Syndrome. Journal of Clinical Medicine 9 (7), 2067.

Choi, S. H., Yoon, H., Kim, H. S., Kim, H. B., Kwon, H. B., Oh, S. M., Lee, Y. J., Park, K. S., 2018. Real-time apnea-hypopnea event detection during sleep by convolutional neural networks. Computers in Biology and Medicine 100 (June), 123–131.

Cohen, J., apr 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20 (1), 37–46.

Cohen, M., Hudson, D., Deedwania, P., 1996. Applying continuous chaotic modeling to cardiac signal analysis. IEEE Engineering in Medicine and Biology Magazine 15 (5), 97–102.

Collop, N. A., 2002. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Medicine 3 (1), 43–47.

Crespo, A., Álvarez, D., Kheirandish-Gozal, L., Gutiérrez-Tobal, G. C., Cerezo-Hernández, A., Gozal, D., Hornero, R., del Campo, F., dec 2018. Assessment of oximetry-based statistical classifiers as simplified screening tools in the management of childhood obstructive sleep apnea. Sleep and Breathing 22 (4), 1063–1073.

Crowson, M. G., Gipson, K. S., Kadosh, O. K., Hartnick, E., Grealish, E., Keamy, D. G., Kinane, T. B., Hartnick, C. J., 2023. Paediatric sleep apnea event prediction using nasal air pressure and machine learning. Journal of Sleep Research 32 (4), 1–7.

Deeks, J. J., Altman, D. G., jul 2004. Diagnostic tests 4: likelihood ratios. BMJ 329 (7458), 168–169.

Dehkordi, P., Garde, A., Karlen, W., Petersen, C. L., Wensley, D., Dumont, G. A., Mark Ansermino, J., feb 2016. Evaluation of cardiac modulation in children in response to apnea/hypopnea using the Phone Oximeter ™. Physiological Measurement 37 (2), 187–202.

Dehlink, E., Tan, H.-L., 2016. Update on paediatric obstructive sleep apnoea. J Thorac Dis 8 (2), 224–235.

del Campo, F., Crespo, A., Cerezo-Hernández, A., Gutiérrez-Tobal, G. C., Hornero, R., Álvarez, D., 2018. Oximetry use in obstructive sleep apnea. Expert Review of Respiratory Medicine 12 (8), 665–681.

DelRosso, L. M., 2016. Epidemiology and Diagnosis of Pediatric Obstructive Sleep Apnea. Current Problems in Pediatric and Adolescent Health Care 46 (1), 2–6.

Dutt, M., Redhu, S., Goodwin, M., Omlin, C. W., 2022. SleepXAI: An explainable deep learning approach for multi-class sleep stage identification. Applied Intelligence.

Erdenebayar, U., Kim, Y. J., Park, J. U., Joo, E. Y., Lee, K. J., 2019. Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. Computer Methods and Programs in Biomedicine 180, 105001.

Ertel, W., 2017. Introduction. In: Introduction to Artificial Intelligence. Springer International Publishing, Cham, Ch. 1, pp. 1–21.

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., Acharya, U. R., jul 2018. Deep learning for healthcare applications based on physiological signals: A review. Computer Methods and Programs in Biomedicine 161, 1–13.

Fawcett, T., jun 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861–874.

Freund, Y., Schapire, R. E., aug 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 55 (1), 119–139.

García-Vicente, C., Gutiérrez-Tobal, G. C., Jiménez-García, J., Martín-Montero, A., Gozal, D., Hornero, R., dec 2023. ECG-based convolutional neural network in pediatric obstructive sleep apnea diagnosis. Computers in Biology and Medicine 167 (December), 107628.

Garde, A., Dehkordi, P., Karlen, W., Wensley, D., Ansermino, J. M., Dumont, G. A., nov 2014. Development of a Screening Tool for Sleep Disordered Breathing in Children Using the Phone Oximeter™. PLoS ONE 9 (11), e112959.

Garde, A., Hoppenbrouwer, X., Dehkordi, P., Zhou, G., Rollinson, A. U., Wensley, D., Dumont, G. A., Ansermino, J. M., aug 2019. Pediatric pulse oximetry-based OSA screening at different thresholds of the apnea-hypopnea index with an expression of uncertainty for inconclusive classifications. Sleep Medicine 60, 45–52.

Gil, E., Bailon, R., Vergara, J. M., Laguna, P., may 2010. PTT Variability for Discrimination of Sleep Apnea Related Decreases in the Amplitude Fluctuations of PPG Signal in Children. IEEE Transactions on Biomedical Engineering 57 (5), 1079–1088.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Gutiérrez-Tobal, G. C., Álvarez, D., Marcos, J. V., del Campo, F., Hornero, R., dec 2013. Pattern recognition in airflow recordings to assist in the sleep apnoea–hypopnoea syndrome diagnosis. Medical & Biological Engineering & Computing 51 (12), 1367–1380.

Gutiérrez-Tobal, G. C., Alonso-Álvarez, M. L., Álvarez, D., del Campo, F., Terán-Santos, J., Hornero, R., apr 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.

Gutierrez-Tobal, G. C., Alvarez, D., del Campo, F., Hornero, R., mar 2016. Utility of AdaBoost to Detect Sleep Apnea-Hypopnea Syndrome From Single-Channel Airflow. IEEE Transactions on Biomedical Engineering 63 (3), 636–646.

Gutierrez-Tobal, G. C., Kheirandish-Gozal, L., Vaquerizo-Villar, F., Alvarez, D., Barroso-Garcia, V., Crespo, A., Campo, F. D., Gozal, D., Hornero, R., jul 2018. Bispectral Analysis to Enhance Oximetry as a Simplified Alternative for Pediatric Sleep Apnea Diagnosis. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Vol. 2018-July. IEEE, pp. 175–178.

Gutiérrez-Tobal, G. C., Álvarez, D., Crespo, A., Del Campo, F., Hornero, R., mar 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.

Gutiérrez-Tobal, G. C., Álvarez, D., Kheirandish-Gozal, L., del Campo, F., Gozal, D., Hornero, R., aug 2022. Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis. Pediatric Pulmonology 57 (8), 1931–1943.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Hornero, R., Kheirandish-Gozal, L., Gutiérrez-Tobal, G. C., Philby, M. F., Alonso-Álvarez, M. L., Álvarez, D., Dayyat, E. A., Xu, Z., Huang, Y.-S., Tamae Kakazu, M., Li, A. M., Van Eyck, A., Brockmann, P. E., Ehsan, Z., Simakajornboon, N., Kaditis, A. G., Vaquerizo-Villar, F., Crespo Sedano, A., Sans Capdevila, O., von Lukowicz, M., Terán-Santos, J., Del Campo, F., Poets, C. F., Ferreira, R., Bertran, K., Zhang, Y., Schuen, J., Verhulst, S., Gozal, D., dec 2017. Nocturnal Oximetry–based Evaluation of Habitually Snoring Children. American Journal of Respiratory and Critical Care Medicine 196 (12), 1591–1598.

Iber, C., Ancoli-Israel, S., Chesson, A. L., Quan, S. F., 2007. The AASM manual for the scoring of sleep and associated events: Rules Terminology and Technical Specification. American academy of sleep medicine, Westchester, IL.

Jiménez-García, J., Gutiérrez-Tobal, G. C., García, M., Kheirandish-Gozal, L., Martín-Montero, A., Álvarez, D., del Campo, F., Gozal, D., Hornero, R., jun 2020. Assessment of Airflow

and Oximetry Signals to Detect Pediatric Sleep Apnea-Hypopnea Syndrome Using AdaBoost. Entropy 22 (6), 670.

Jiménez-García, J., García, M., Gutiérrez-Tobal, G. C., Kheirandish-Gozal, L., Vaquerizo-Villar, F., Álvarez, D., del Campo, F., Gozal, D., Hornero, R., aug 2022. A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry. Computers in Biology and Medicine 147 (August), 105784.

Jiménez-García, J., García, M., Gutiérrez-Tobal, G. C., Kheirandish-Gozal, L., Vaquerizo-Villar, F., Álvarez, D., del Campo, F., Gozal, D., Hornero, R., jan 2024. An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals. Biomedical Signal Processing and Control 87 (September 2023), 105490.

Jon, C., 2009. Polysomnography in Children. In: Mitchell, R., Pereira, K. (Eds.), Pediatric Otolaryngology for the Clinician. Humana Press, Totowa, NJ, pp. 35–47.

Joosten, K. F., Larramona, H., Miano, S., Van Waardenburg, D., Kaditis, A. G., Vandenbussche, N., Ersu, R., feb 2017. How do we recognize the child with OSAS? Pediatric Pulmonology 52 (2), 260–271.

Kaditis, A., Kheirandish-Gozal, L., Gozal, D., jun 2016a. Pediatric OSAS: Oximetry can provide answers when polysomnography is not available. Sleep Medicine Reviews 27, 96–105.

Kaditis, A. G., Alonso Alvarez, M. L., Boudewyns, A., Alexopoulos, E. I., Ersu, R., Joosten, K., Larramona, H., Miano, S., Narang, I., Trang, H., Tsaoussoglou, M., Vandenbussche, N., Villa, M. P., Van Waardenburg, D., Weber, S., Verhulst, S., jan 2016b. Obstructive sleep disordered breathing in 2- to 18-year-old children: diagnosis and management. European Respiratory Journal 47 (1), 69–94.

Korkalainen, H., Aakko, J., Nikkonen, S., Kainulainen, S., Leino, A., Duce, B., Afara, I. O., Myllymaa, S., Toyras, J., Leppanen, T., 2020a. Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea. IEEE Journal of Biomedical and Health Informatics 24 (7), 2073–2081.

Korkalainen, H., Aakko, J., Duce, B., Kainulainen, S., Leino, A., Nikkonen, S., Afara, I. O., Myllymaa, S., Töyräs, J., Leppänen, T., 2020b. Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea. Sleep 43 (11), 1–10.

Kuncheva, L. I., 2014. Combining Pattern Classifiers: Methods and Algorithms. Wiley.

Kuo, C.-E., Chen, G.-T., Liao, P.-Y., sep 2021. An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge. Biomedical Signal Processing and Control 70 (July), 102981.

Lazaro, J., Gil, E., Vergara, J. M., Laguna, P., jan 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 18 (1), 240–246.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Leino, A., Nikkonen, S., Kainulainen, S., Korkalainen, H., Töyräs, J., Myllymaa, S., Leppänen, T., Ylä-Herttuala, S., Westeren-Punnonen, S., Muraja-Murro, A., Jäkälä, P., Mervaala, E., Myllymaa, K., 2021. Neural network analysis of nocturnal SpO2 signal enables easy screening of sleep apnea in patients with acute cerebrovascular disease. Sleep Medicine 79, 71–78.

Lempel, A., Ziv, J., jan 1976. On the Complexity of Finite Sequences. IEEE Transactions on Information Theory 22 (1), 75–81.

Li, A. M., Au, C. T., So, H. K., Lau, J., Ng, P. C., Wing, Y. K., sep 2010. Prevalence and Risk Factors of Habitual Snoring in Primary School Children. Chest 138 (3), 519–527.

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., Acharya, U. R., 2022. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine 226, 107161.

Magalang, U. J., Dmochowski, J., Veeramachaneni, S., Draw, A., Mador, M. J., El-Solh, A., Grant, B. J., 2003. Prediction of the Apnea-Hypopnea Index from Overnight Pulse Oximetry. Chest 124 (5), 1694–1701.

Marcos, J. V., Hornero, R., Álvarez, D., Aboy, M., Del Campo, F., jan 2012. Automated Prediction of the Apnea-Hypopnea Index from Nocturnal Oximetry Recordings. IEEE Transactions on Biomedical Engineering 59 (1), 141–149.

Marcus, C. L., Brooks, L. J., Ward, S. D., Draper, K. A., Gozal, D., Halbower, A. C., Jones, J., Lehmann, C., Schechter, M. S., Sheldon, S., Shiffman, R. N., Spruyt, K., sep 2012. Diagnosis and Management of Childhood Obstructive Sleep Apnea Syndrome. Pediatrics 130 (3), e714–e755.

Marcus, C. L., Moore, R. H., Rosen, C. L., Giordani, B., Garetz, S. L., Taylor, H. G., Mitchell, R. B., Amin, R., Katz, E. S., Arens, R., Paruthi, S., Muzumdar, H., Gozal, D., Thomas, N. H., Ware, J., Beebe, D., Snyder, K., Elden, L., Sprecher, R. C., Willging, P., Jones, D., Bent, J. P., Hoban, T., Chervin, R. D., Ellenberg, S. S., Redline, S., 2013. A Randomized Trial of Adenotonsillectomy for Childhood Sleep Apnea. New England Journal of Medicine 368 (25), 2366–2376.

Martín-Montero, A., Gutiérrez-Tobal, G. C., Gozal, D., Barroso-García, V., Álvarez, D., Del Campo, F., Kheirandish-Gozal, L., Hornero, R., 2021a. Bispectral analysis of heart rate variability to characterize and help diagnose pediatric sleep apnea. Entropy 23 (8).

Martín-Montero, A., Gutiérrez-Tobal, G. C., Kheirandish-Gozal, L., Jiménez-García, J., Álvarez, D., del Campo, F., Gozal, D., Hornero, R., may 2021b. Heart rate variability spectrum characteristics in children with sleep apnea. Pediatric Research 89 (7), 1771–1779.

Martín-Montero, A., Armañac-Julián, P., Gil, E., Kheirandish-Gozal, L., Álvarez, D., Lázaro, J., Bailón, R., Gozal, D., Laguna, P., Hornero, R., Gutiérrez-Tobal, G. C., mar 2023. Pediatric sleep apnea: Characterization of apneic events and sleep stages using heart rate variability. Computers in Biology and Medicine 154 (September 2022), 106549.

Mazzotti, D. R., Lim, D. C., Sutherland, K., Bittencourt, L., Mindel, J. W., Magalang, U., Pack, A. I., de Chazal, P., Penzel, T., sep 2018. Opportunities for utilizing polysomnography signals

to characterize obstructive sleep apnea subtypes and severity. Physiological Measurement 39 (9), 09TR01.

Meltzer, L. J., Paisley, C., jun 2023. Beyond Polysomnography. Sleep Medicine Clinics 18 (2), 147–160.

Mendonca, F., Mostafa, S. S., Ravelo-Garcia, A. G., Morgado-Dias, F., Penzel, T., mar 2019. A Review of Obstructive Sleep Apnea Detection Approaches. IEEE Journal of Biomedical and Health Informatics 23 (2), 825–837.

Moffa, A., Rinaldi, V., Costantino, A., Cassano, M., Gelardi, M., Fiore, V., Lopez, M. A., Baptista, P., Campisi, P., Casale, M., 2020. Childhood Obstructive Sleep Apnea: from Diagnosis to Therapy—an Update. Current Sleep Medicine Reports 6 (3), 157–162.

Mostafa, S. S., Mendonça, F., Ravelo-García, A. G., Morgado-Dias, F., 2019. A systematic review of detecting sleep apnea using deep learning. Sensors (Switzerland) 19 (22), 1–26.

Mostafa, S. S., Baptista, D., Ravelo-García, A. G., Juliá-Serdá, G., Morgado-Dias, F., 2020. Greedy based convolutional neural network optimization for detecting apnea. Computer Methods and Programs in Biomedicine 197, 105640.

Nikkonen, S., Korkalainen, H., Leino, A., Myllymaa, S., Duce, B., Leppanen, T., Toyras, J., aug 2021. Automatic Respiratory Event Scoring in Obstructive Sleep Apnea Using a Long Short-Term Memory Neural Network. IEEE Journal of Biomedical and Health Informatics 25 (8), 2917–2927.

Oceja, E., Rodríguez, P., Jurado, M. J., Alonso, M. L., Del Río, G., Villar, M. Á., Mediano, O., Martínez, M., Juarros, S., Merino, M., Corral, J., Luna, C., Kheirandish-Gozal, L., Gozal, D., Durán-Cantolla, J., 2021. Validity and cost-effectiveness of pediatric home respiratory polygraphy for the diagnosis of obstructive sleep apnea in children: Rationale, study design, and methodology. Methods and Protocols 4 (1), 1–14.

Rangayyan, R. M., 2015. Biomedical signal analysis, 2nd Edition. IEEE, Hoboken, New Jersey.

Redline, S., Amin, R., Beebe, D., Chervin, R. D., Garetz, S. L., Giordani, B., Marcus, C. L., Moore, R. H., Rosen, C. L., Arens, R., Gozal, D., Katz, E. S., Mitchell, R. B., Muzumdar, H., Taylor, H. G., Thomas, N., Ellenberg, S., 2011. The Childhood Adenotonsillectomy Trial (CHAT): Rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population. Sleep 34 (11), 1509–1517.

Richman, J. S., Moorman, J. R., jun 2000. Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology-Heart and Circulatory Physiology 278 (6), H2039–H2049.

Riha, R. L., Celmina, M., Cooper, B., Hamutcu-Ersu, R., Kaditis, A., Morley, A., Pataka, A., Penzel, T., Roberti, L., Ruehland, W., Testelmans, D., van Eyck, A., Grundström, G., Verbraecken, J., Randerath, W., jan 2023. ERS technical standards for using type III devices (limited channel studies) in the diagnosis of sleep disordered breathing in adults and children. European Respiratory Journal 61 (1), 2200422.

Roebuck, A., Monasterio, V., Gederi, E., Osipov, M., Behar, J., Malhotra, A., Penzel, T., Clifford, G. D., 2014. A review of signals used in sleep analysis. Physiological Measurement 35 (1).

Rosen, C. L., D'andrea, L., Haddad, G. G., nov 1992. Adult Criteria for Obstructive Sleep Apnea Do Not Identify Children with Serious Obstruction. American Review of Respiratory Disease 146 (5_pt_1), 1231–1234.

Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23 (19), 2507–2517.

Sagi, O., Rokach, L., jul 2018. Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery 8 (4), 1–18.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision 2017-Octob, 618–626.

Serrano Alarcón, Á., Martínez Madrid, N., Seepold, R., Ortega, J. A., 2023. Obstructive sleep apnea event detection using explainable deep learning models for a portable monitor. Frontiers in Neuroscience 17 (July), 1–19.

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. Physiological Measurement 38 (9), R204–R252.

Shouldice, R. B., O'Brien, L. M., O'Brien, C., de Chazal, P., Gozal, D., Heneghan, C., jun 2004. Detection of Obstructive Sleep Apnea in Pediatric Subjects using Surface Lead Electrocardiogram Features. Sleep 27 (4), 784–792.

Skotko, B. G., Macklin, E. A., Muselli, M., Voelz, L., McDonough, M. E., Davidson, E., Allareddy, V., Jayaratne, Y. S. N., Bruun, R., Ching, N., Weintraub, G., Gozal, D., Rosen, D., apr 2017. A predictive model for obstructive sleep apnea and Down syndrome. American Journal of Medical Genetics Part A 173 (4), 889–896.

Stowe, R. C., Afolabi-Brown, O., apr 2020. Pediatric polysomnography—A review of indications, technical aspects, and interpretation. Paediatric Respiratory Reviews 34, 9–17.

Tan, H.-L., Bandla, H. P. R., Ramirez, H. M., Gozal, D., Kheirandish-Gozal, L., 2014. Overnight Polysomnography versus Respiratory Polygraphy in the Diagnosis of Pediatric Obstructive Sleep Apnea. Sleep 37 (2), 255–260.

Tan, H.-L., Kheirandish-Gozal, L., Gozal, D., dec 2015. Pediatric Home Sleep Apnea Testing. Chest 148 (6), 1382–1395.

Tauman, R., Gozal, D., jun 2011. Obstructive sleep apnea syndrome in children. Expert Review of Respiratory Medicine 5 (3), 425–440.

Theissler, A., Spinnato, F., Schlegel, U., Guidotti, R., 2022. Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions. IEEE Access 10 (August), 100700–100724.

Uddin, M. B., Chow, C. M., Su, S. W., 2018. Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: A systematic review. Physiological Measurement 39 (3).

Van Eyck, A., Verhulst, S. L., mar 2018. Improving the diagnosis of obstructive sleep apnea in children with nocturnal oximetry-based evaluations. Expert Review of Respiratory Medicine 12 (3), 165–167.

Vaquerizo-Villar, F., Álvarez, D., Kheirandish-Gozal, L., Gutiérrez-Tobal, G. C., Barroso-García, V., Crespo, A., del Campo, F., Gozal, D., Hornero, R., nov 2018a. Detrended fluctuation analysis of the oximetry signal to assist in paediatric sleep apnoea–hypopnoea syndrome diagnosis. Physiological Measurement 39 (11), 114006.

Vaquerizo-Villar, F., Álvarez, D., Kheirandish-Gozal, L., Gutiérrez-Tobal, G. C., Barroso-García, V., Crespo, A., del Campo, F., Gozal, D., Hornero, R., mar 2018b. Utility of bispectrum in the screening of pediatric sleep apnea-hypopnea syndrome using oximetry recordings. Computer Methods and Programs in Biomedicine 156, 141–149.

Vaquerizo-Villar, F., Álvarez, D., Kheirandish-Gozal, L., Gutiérrez-Tobal, G. C., Barroso-García, V., Crespo, A., del Campo, F., Gozal, D., Hornero, R., dec 2018c. Wavelet analysis of oximetry recordings to assist in the automated detection of moderate-to-severe pediatric sleep apnea-hypopnea syndrome. PLoS ONE 13 (12), e0208502.

Vaquerizo-Villar, F., Alvarez, D., Kheirandish-Gozal, L., Gutierrez-Tobal, G. C., Barroso-Garcia, V., Santamaria-Vazquez, E., del Campo, F., Gozal, D., Hornero, R., aug 2021. A Convolutional Neural Network Architecture to Enhance Oximetry Ability to Diagnose Pediatric Obstructive Sleep Apnea. IEEE Journal of Biomedical and Health Informatics 25 (8), 2906–2916.

Vaquerizo-Villar, F., Gutiérrez-Tobal, G. C., Calvo, E., Álvarez, D., Kheirandish-Gozal, L., del Campo, F., Gozal, D., Hornero, R., 2023. An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea. Computers in Biology and Medicine 165 (August).

Várady, P., Micsik, T., Benedek, S., Benyó, Z., sep 2002. A novel method for the detection of apnea and hypopnea events in respiration signals. IEEE Transactions on Biomedical Engineering 49 (9), 936–942.

Welch, P. D., 1967. 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 AU-15 (2), 70–73.

Witten, I. H., Frank, E., Hall, M. A., 2011. Data mining: practical machine learning tools and techniques, 3rd Edition. Morgan Kaufmann/Elsevier, Burlington.

Wu, D., Li, X., Guo, X., Qin, J., Li, S., apr 2017. A simple diagnostic scale based on the analysis and screening of clinical parameters in paediatric obstructive sleep apnoea hypopnea syndrome. The Journal of Laryngology & Otology 131 (04), 363–367.

Xu, Z., Gutiérrez-Tobal, G. C., Wu, Y., Kheirandish-Gozal, L., Ni, X., Hornero, R., Gozal, D., feb 2019. Cloud algorithm-driven oximetry-based diagnosis of obstructive sleep apnoea in symptomatic habitually snoring children. European Respiratory Journal 53 (2), 1801788.

Ye, P., Qin, H., Zhan, X., Wang, Z., Liu, C., Song, B., Kong, Y., Jia, X., Qi, Y., Ji, J., Chang, L., Ni, X., Tai, J., 2022. Diagnosis of obstructive sleep apnea in children based on the XGBoost algorithm using nocturnal heart rate and blood oxygen feature. American Journal of Otolaryngology 44 (2), 103714.

Yu, L., Liu, H., 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. J. Mach. Learn. Res. 5, 1205–1224.

Index