

Article

Development of a Minimally Invasive Screening Tool to Identify Obese Pediatric Population at Risk of Obstructive Sleep Apnea/Hypopnea Syndrome

José Miguel Calderón ¹, Julio Álvarez-Pitti ² , Irene Cuenca ¹ , Francisco Ponce ^{2,3} and Pau Redon ^{2,3,*}

¹ Fundación Investigación Hospital Clínico (INCLIVA), Avda. Menedez Pelayo 4, 46010 Valencia, Spain; jocalte85@gmail.com (J.M.C.); irene.cuenca95@gmail.com (I.C.)

² Pediatric Department, Consorcio Hospital General Universitario de Valencia, Avda. Tres Cruces s/n, 46014 Valencia, Spain; japnago@gmail.com (J.Á.-P.); siscopz80@gmail.com (F.P.)

³ CIBEROBN, Health Institute Carlos III, Av. Monforte de Lemos, 3-5. Pavilion 11, 28029 Madrid, Spain

* Correspondence: paredlur@gmail.com

Received: 15 September 2020; Accepted: 17 October 2020; Published: 19 October 2020



Abstract: Obstructive sleep apnea syndrome is a reduction of the airflow during sleep which not only produces a reduction in sleep quality but also has major health consequences. The prevalence in the obese pediatric population can surpass 50%, and polysomnography is the current gold standard method for its diagnosis. Unfortunately, it is expensive, disturbing and time-consuming for experienced professionals. The objective is to develop a patient-friendly screening tool for the obese pediatric population to identify those children at higher risk of suffering from this syndrome. Three supervised learning classifier algorithms (i.e., logistic regression, support vector machine and AdaBoost) common in the field of machine learning were trained and tested on two very different datasets where oxygen saturation raw signal was recorded. The first dataset was the Childhood Adenotonsillectomy Trial (CHAT) consisting of 453 individuals, with ages between 5 and 9 years old and one-third of the patients being obese. Cross-validation was performed on the second dataset from an obesity assessment consult at the Pediatric Department of the Hospital General Universitario of Valencia. A total of 27 patients were recruited between 5 and 17 years old; 42% were girls and 63% were obese. The performance of each algorithm was evaluated based on key performance indicators (e.g., area under the curve, accuracy, recall, specificity and positive predicted value). The logistic regression algorithm outperformed (accuracy = 0.79, specificity = 0.96, area under the curve = 0.9, recall = 0.62 and positive predictive value = 0.94) the support vector machine and the AdaBoost algorithm when trained with the CHAT datasets. Cross-validation tests, using the Hospital General de Valencia (HG) dataset, confirmed the higher performance of the logistic regression algorithm in comparison with the others. In addition, only a minor loss of performance (accuracy = 0.75, specificity = 0.88, area under the curve = 0.85, recall = 0.62 and positive predictive value = 0.83) was observed despite the differences between the datasets. The proposed minimally invasive screening tool has shown promising performance when it comes to identifying children at risk of suffering obstructive sleep apnea syndrome. Moreover, it is ideal to be implemented in an outpatient consult in primary and secondary care.

Keywords: machine learning; oxygen saturation signal; obstructive sleep apnea syndrome; obese pediatric population

1. Introduction

The obstructive sleep apnea syndrome (OSAS) is a health problem characterized by recurrent episodes of reduction of airflow. These can be partial (hypopnea) or complete (apnea). According to the American Academy of Sleep Medicine (AASM), apnea is defined as an airflow reduction of at least 90% lasting for at least 10 s, while a hypopnea is a 30% reduction for at least 10 s or at least a 3% reduction in oxygen saturation (SpO₂) in comparison with the pre-event baseline or associated with arousal [1,2]. The frequency of these events in a fixed time interval, 1 h, is called the apnea/hypopnea index (AHI) and is used to classify the severity of the OSAS syndrome into three levels: mild ($5 \leq \text{AHI} < 15$), moderate ($15 \leq \text{AHI} < 30$) and severe ($30 \leq \text{AHI}$) [1,2]. OSAS can cause a deterioration of the sleep quality as well as other major negative consequences (e.g., behavior regulation, compliance, aggression, impulsivity, hyperactivity, anxiety, depressed mood, emotion regulation and neurocognitive deficits) [3,4]. Its prevalence in children ranges between 0.2% and 4%, increasing in the obese population up to over 50% [3].

Full night polysomnography (PSG) has been recognized by numerous professionals as the gold standard procedure to diagnose OSAS in children and adults [5–8]. Generally speaking, it consists of a minimum of 11 channels recording different signals like the electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG) and electrocardiogram (ECG). Unfortunately, despite the accuracy of the results, it is very disturbing for the patient, particularly in the pediatric population, and expensive because it requires special infrastructure and trained personnel. In an attempt to overcome these limitations, other methods have been proposed (e.g., respiratory polygraphy (RP), pediatric sleep questionnaire (PSQ), sleep clinical record (SCR), nocturnal oximetry studies (NOS)) [9].

Out of all the listed alternative methods, the authors are especially interested in those which can be performed in an outpatient context. This is especially relevant for screening and follow-up strategies in the obese pediatric population. The first of them is the RP, which reduces the invasiveness of the procedure by reducing the number of signals being recorded. Usually, the oronasal flow, chest and abdomen movements, heart rate and blood oxygen are the key parameters measured. Different tests have demonstrated that RP and home-based RP (HRP) are suitable for diagnosing OSAS in children and adolescents [10,11]. The second is the PSQ test. It consists of a set of questions related to 22 symptom items that ask about snoring frequency, loud snoring, observed apneas, difficulty breathing during sleep, daytime sleepiness, inattentive or hyperactive behavior and other pediatric OSAS features. It is oriented to children aged between 2 and 18 years old, and the sleep-related breathing disorder (SRBD) scale can predict the risk to an extent useful for research but not reliable enough for most individual patients [12,13]. The third is the NOS procedure, which measures the arterial oxygen saturation by the detection of pulsatile blood flow. The performance of continuous oximetry overnight can help to identify desaturation patterns useful to identify OSAS in children. As pointed out by Singh et al. [14], this method has been extensively used in the field of sleep medicine and has been classified as a type 4 monitoring device. It is cheap, available and easy to perform on an outpatient basis.

Despite the evident differences between all of these methods, they still require trained personnel to interpret the results. In some cases, this can be time-consuming even if a signal processing software is available with the hardware. Besides the increased physician workload that can be associated with such testing, the results are strongly dependent on the experience of the operator, and strong variabilities in the diagnosis might occur between experts.

In the past two decades, multiple solutions have been proposed to reduce or mitigate such dependencies in adults and pediatric populations (see Table 1). Generally speaking, these are characterized by a set of features extracted from one or several recorded biosignals and a classifier for automatic diagnosis of OSAS severity (see Table 1). Mostafa et al. [15] recently reviewed the solutions available for the adult population. Initially, univariate analyses were employed, and now machine learning (ML) and deep learning (DL) algorithms are proliferating. These can either identify events directly from a register or based on a global analysis of the data.

For the pediatric population, there seem to coexist ML- and DL-based solutions with similar performance between them. The threshold values used to classify severity have a greater impact on performance in this cohort than in adults.

Table 1 presents a summary of the performance of each these tools, including the results regarding sensitivity, specificity and accuracy of solutions focusing on the SpO2 signal. Even though the differences in datasets, sample sizes and methods do not allow for direct comparison between them, this information is useful to benchmark the performance of new tools. Accuracy is similar between both populations, while greater specificity and sensitivity are observed in the adult population.

Table 1. Summary of the performance of oxygen saturation (SpO2)-based tools to diagnose obstructive sleep apnea (OSA) syndrome.

Reference	Cohort	Type of Classifier	Sample Size	Sensitivity	Specificity	Accuracy	Year	Home-Based
[16]	A	Multivariate adaptive regression splines	793	83	54	NA	1999	N
[17]	A	Linear regression	148	91	83	89	2009	N
[18]	A	Univariate	475	96	67	87	2012	Y
[19]	A	Baggin ReTree	25	78	84	83	2012	N
[20]	A	Artificial Neural Network	93	88	100	93	2012	N
[21]	A	Univariate	996	84	86	NA	2014	Y
[22]	A	Linear discriminant analysis	302	97	50	93	2017	Y
[23]	A	Deep belief networks	33	60	92	85	2017	N
[24]	A	Long-short term memory	8	93	NA	96	2017	N
[25]	A	Convolutional neural networks	23	NA	NA	80	2018	N
[26]	A	Recurrent and convolutional neural network	15,804	NA	NA	88	2018	N
[27]	A	Common Bayesian Network	32	NA	NA	85	2017	N
[28]	P	Neural network	176	NA	NA	84.7–85.8	2015	N
[29]	P	Logistic regression	298	79.1	84.1	81.9	2017	N
[30]	P	Neural network	4191	84.0–68.7	53–94	75.2–90	2017	N
[31]	P	Logistic regression, QDA, LDA	176	NA	NA	84.3–82.7	2018	N
[32]	P	Convolutional neural network	298	NA	NA	81.3–85.3	2018	N
[33]	P	Convolutional neural network	779	40–54	98.6–99.6	74.8–95.1	2020	N
[34]	P	AdaBoost	974	91–41	22.7–98.1	78.2–85.9	2020	N

A, adult population; P, pediatric population; NA, not available; QDA, quadratic discriminant analysis; LDA, linear discriminant analysis; Y, yes; N, no.

Unfortunately, the development of this type of solution is not accessible to everyone and in some cases might require a continuous update. In this context, the main objective of this paper is to develop a patient-friendly screening tool oriented to the pediatric population and feasible to be implemented on an outpatient basis. Moreover, this solution should be easy to run, minimizing the interaction of the physicians throughout the whole process and keeping them updated without the need for domain-specific skills. This will allow the popularizing of this type of diagnostic test, which can be especially relevant for specific pediatric cohorts like the obese population.

2. Materials and Methods

2.1. Datasets

Two datasets were used in this study. The first one was the Childhood Adenotonsillectomy Trial (CHAT), obtained from National Sleep Research Resource (NSRR), [6,35–37] with 453 children with ages ranging between 5 and 10 years old. Fifty-two percent of the registries corresponded to girls, one-third of the population was obese (defined as BMI percentile at or above 95%) and the predominant race was black (55%). All participants underwent polysomnography in the sleep laboratory of the hospital in which the minimum period of lights off was 7 h. A total of 43% of them suffered severe OSAS.

Out of all the variables available from this dataset, only those related to oxygen saturation (SpO2) were considered in the present study. The main reason is that apneas and hypopneas are both defined by this parameter. This signal was recorded in the CHAT with a pulse-oximeter, Nonin Model 8000J or comparable, at a sampling rate of 10 Hz or higher. The extracted features from the SpO2 signal are

listed and defined in Table 2 [6,35–37]. It is worth noting that the oxygen desaturation index (ODI) has been considered as the hourly average number of desaturation episodes, defined as at events lasting >10 s with at least a 3% decrease in saturation from the average saturation in the preceding 120 s [18]. Based on the ODI, the desaturation periods were tagged as hypopnea or apnea. All subjects were classified according to AHI (ahi_a0h3a). Patients with an AHI < 5 were labeled as healthy, whereas patients with an AHI \geq 5 were considered at risk of suffering OSAS syndrome.

The second dataset consisted of 27 subjects, between 5 and 17 years of age and recruited from those attending an obesity assessment consult at the Pediatric Department of Hospital General of Valencia. All were Caucasian, 42% were girls and 63% were obese (defined as BMI percentile at or above 95%). All participants underwent an HRP using the Philips Alice Respironics device with a minimum of 6 h of time in bed (TIB) and 3 h of sleep time [38]. The recorded signals were airflow, thoracic and abdominal movement, SpO₂, body position, ECG and snore. Only the SpO₂ raw signal was processed to extract features and calculate the variables of interest according to the definitions shown in Table 2. To do so, MATLAB software was employed.

Finally, to allow direct comparison of the variables from both datasets, a standardization process was performed. This consisted of calculating the Z-score value for each of them.

Table 2. Extracted features from the raw SpO₂ signal used in the Childhood Adenotonsillectomy Trial (CHAT) [6,35–37].

Variable	Description
ahi_a0h3a	Apnea/hypopnea index (AHI) \geq 3% oxygen desaturation per hour of sleep
odi3	Oxygen desaturation index \geq 3% during sleep time
odi4	Oxygen desaturation index \geq 4% during sleep time
ndes2ph	Number of desaturations with \geq 2% desaturation
ndes3ph	Number of desaturations with \geq 3% desaturation
ndes4ph	Number of desaturations with \geq 4% desaturation
ndes5ph	Number of desaturations with \geq 5% desaturation
pctle90	Percentage of time \leq 90% oxygen saturation
pctle92	Percentage of time \leq 92% oxygen saturation

2.2. Machine Learning Algorithms

Three popular ML algorithms specially oriented to supervised learning were used in this study. The first of them is the logistic regression (LR) model, which uses a weighted least square algorithm to predict the regression line that best fits the data points by minimizing the weighted sum of the square distances to the fitted regression line. It is simple and easy to implement and can relate one dependent variable with one or several independent variables. The second is the support vector machine (SVM), which tries to model the input variables by finding the separating boundary (i.e., hyperplane) to reach the classification of the input variables [39]. The third is the AdaBoost (AB) model, which is a common ensemble method that combines in series multiple weak classifiers to generate a strong one. These algorithms from the Scikit-learn library [40] were trained and afterward tested in terms of their capability to classify the subjects. Each classification algorithm has its own set of parameters that can be modified to obtain a better performance. SVM was implemented with an 'rbg' kernel, a penalty parameter C from 0.01 to 100 and a gamma parameter from 0.001 to 100. LR classifier was run with L1 penalty and a C, inverse of regularization strength, from 0.01 to 100. AdaBoost was implemented with Decision Tree as a base estimator, a learning rate parameter from 0.01 to 1 and several estimators from 50 to 1000 with an interval of 10. The machine learning algorithms were implemented in Python [41,42].

The CHAT dataset was used to train and test each of the classifiers according to the extracted features from the SpO₂ raw signal. A 15 k-fold approach was used. Balanced datasets were used. On the other hand, cross-validation was performed using the Hospital General de Valencia (HG) dataset, which is only composed of an obese population ranging from 9 to 17 years of age. The same features as those in the CHAT dataset were calculated from the raw SpO₂ signal using MATLAB software.

Balanced datasets were used to evaluate the performance of each model based on the following parameters: recall or sensitivity, precision or positive predictive value (PPV), specificity, negative predictive value (NPV), accuracy, F1 score and area under the curve (AUC). The recall or sensitivity and precision or PPV are two very important parameters for assessing the performance in identifying unhealthy patients (i.e., denoted as positive subjects in the current manuscript). The former is the portion of real positive cases that are correctly predicted positive, while the latter denotes the portion of predicted positive cases that are really positive. Mathematically, they are defined according to Equations (1) and (2), respectively.

$$\text{Sensitivity} = \text{True Positive}/(\text{True Positive} + \text{False Negative}) \quad (1)$$

$$\text{PPV} = \text{True Positive}/(\text{True Positive} + \text{False Positive}) \quad (2)$$

The specificity and the negative predictive value (NPV) are respectively homologous to the two previously described parameters but assess the performance of the model regarding its capability to identify the negative values or healthy subjects. They are mathematically defined by Equations (3) and (4).

$$\text{Specificity} = \text{True Negative}/(\text{True Negative} + \text{False Positive}) \quad (3)$$

$$\text{NPV} = \text{True Negative}/(\text{True Negative} + \text{False Negative}) \quad (4)$$

Accuracy and F1 score are two parameters frequently used to assess overall performance. Accuracy is the portion of correctly identified cases, independently of being positive or negative, from the total number of samples (see Equation (5)). The F1 score is the harmonic mean between PPV and sensitivity. Out of the three different types of averages that can be calculated (i.e., arithmetic, geometric and harmonic), the harmonic average is the most conservative of them all; in other words, it is the one that yields the lowest value. It is defined by Equation (6).

$$\text{Accuracy} = (\text{True Positive} + \text{True Negative})/(\text{Sample size}) \quad (5)$$

$$\text{F1 score} = 2 \times \text{True Positive}/(2 \times \text{True Positive} + \text{False Positive} + \text{False Negative}) \quad (6)$$

Finally, the area under the curve (AUC) parameter was also considered. This value is calculated after performing a receiver operating characteristic curve (ROC) analysis. It can also be assessed graphically by plotting sensitivity against 1 – specificity and determining the area under the resulting curve. A higher AUC indicates that the model has a better capability to distinguish between healthy and unhealthy subjects. Figure 1 summarizes the methodology used in this manuscript.

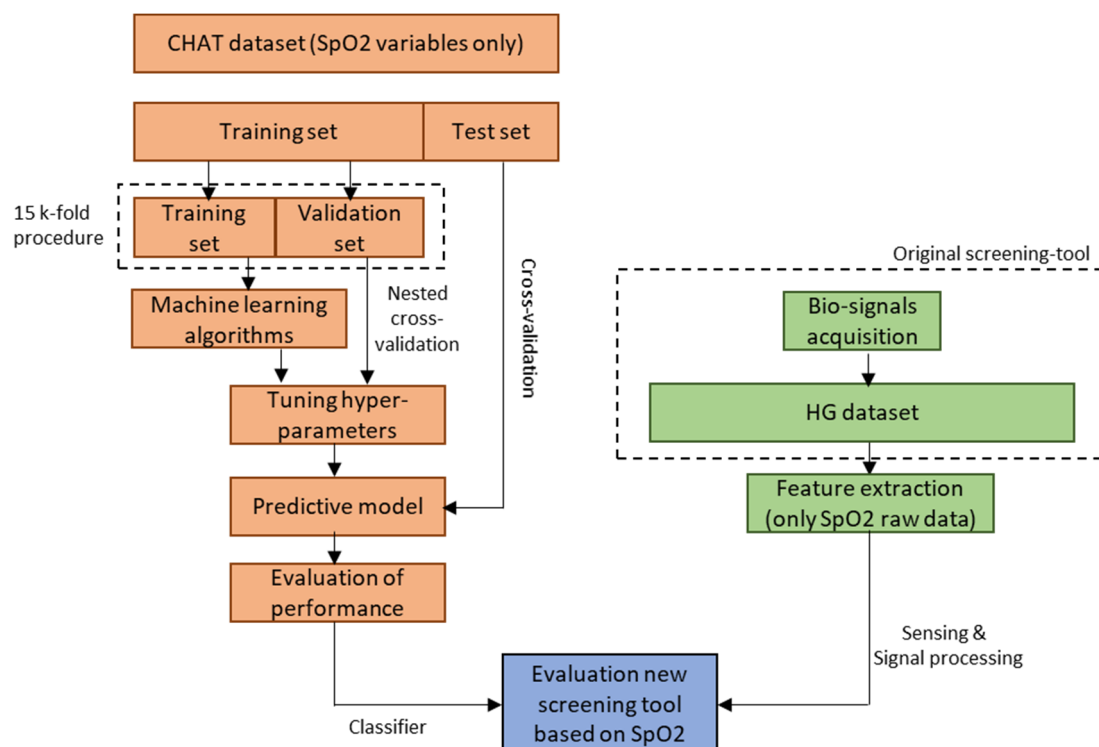


Figure 1. Schematic representation of the workflow followed to generate a screening tool based exclusively on pulse oximetry measurements and machine learning algorithms. This tool is specially oriented to screen asymptomatic obese pediatric population in search of subjects at risk of suffering OSA syndrome. The manuscript only focuses on the generation of this new tool, emphasizing the development of the classifier.

2.3. Statistical Analyses

Statistical analyses were performed on both datasets using Python [40,41,43,44]. Initially, a histogram was plotted to visualize the frequency distribution regarding the AHI values of both datasets. Afterward, a Shapiro–Wilk test was performed on the extracted features to determine if they had a normal distribution. If not, the Mann–Whitney U test was performed to determine if significant differences were present between healthy and unhealthy patients in each of the datasets. Lastly, the ROC analysis was applied to evaluate the performance of the different models by calculating the AUC parameter. A p -value ≤ 0.05 was considered to be significant.

To build the model based on the extracted features, the first step was to perform a cross-correlation matrix to determine the relationship between each pair of features as well as with AHI. The authors arbitrarily defined a strong correlation if values were greater than 0.9. In this case, dimensionality reduction was applied. This is crucial for obtaining an effective algorithm by avoiding the incorporation of repeated information into the model. This dimensionality reduction was achieved in LR using the L1 penalty term and by the mutual information measure (MI) in the SVM and AdaBoost procedures. In other words, features with lower MI were eliminated.

The selection of the best predictive model was done using a nested cross-validation method for adjusting the model parameters and estimating the error.

3. Results

3.1. Preprocessing

Initially, the distribution of AHI for each dataset was plotted (see Figure 2). As expected, the CHAT dataset has patients with AHI ranging between 1 and 27 approximately. In contrast, the HG dataset is predominantly composed of subjects under the established threshold of 5 or with high AHI values, ≥ 15 .

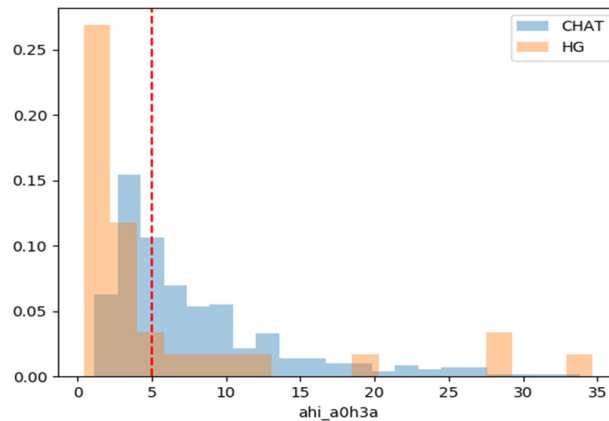


Figure 2. Histogram of apnea/hypopnea index of the CHAT (blue) and the Hospital General de Valencia (HG) (orange) datasets. The discontinuous red line depicts the threshold value used in the present paper, AHI = 5. Individuals with AHI ≤ 5 were considered as healthy.

Afterward, Shapiro–Wilk analysis was applied to the CHAT dataset to determine if all variables had a normal distribution. As expected, this test confirmed that all variables were not normally distributed. Their p -values were lower than 0.05 (see Table 3). Accordingly, Mann–Whitney U analysis was performed to determine if significant differences were present between the healthy (AHI ≤ 5) and unhealthy groups (AHI > 5). Correlation analyses were performed to check the relationship between each pair of features (see Figure 3). The results reveal that strong correlations are present between odi3 and odi4 as well as ndes2ph, ndes3ph, ndes4ph and ndes5ph. Consequently, a dimensionality reduction procedure was applied to these variables before testing the three algorithms. The variables odi4 and odi3 were found to correlate the most with AHI.

Table 3. Result of applying inferential statistics test on extracted features from the SpO2 raw signal.

Feature	Healthy, n = 197 (Mean \pm std)	At Risk, n = 256 (Mean \pm std)	Shapiro–Wilk	Mann Whitney U
			p -Value	p -Value
ndes2ph	82.91 \pm 58.31	189.94 \pm 107.56	$< 1 \times 10^{-15}$	$< 1 \times 10^{-30}$
ndes3ph	28.13 \pm 20.45	91.67 \pm 63.06	$< 1 \times 10^{-20}$	$< 1 \times 10^{-40}$
ndes4ph	10.08 \pm 8.44	47.17 \pm 40.32	$< 1 \times 10^{-20}$	$< 1 \times 10^{-40}$
ndes5ph	4.41 \pm 4.49	26.51 \pm 26.97	$< 1 \times 10^{-25}$	$< 1 \times 10^{-40}$
odi3	2.79 \pm 2.08	10.53 \pm 7.38	$< 1 \times 10^{-20}$	$< 1 \times 10^{-45}$
odi4	0.98 \pm 0.83	5.53 \pm 4.81	$< 1 \times 10^{-25}$	$< 1 \times 10^{-45}$
pctle90	0.06 \pm 0.69	0.29 \pm 0.51	$< 1 \times 10^{-35}$	$< 1 \times 10^{-25}$
pctle92	0.38 \pm 3.49	0.81 \pm 1.37	$< 1 \times 10^{-35}$	$< 1 \times 10^{-25}$

ndes2ph, number of desaturations $\geq 2\%$ per hour; ndes3ph, number of desaturations $\geq 3\%$ per hour; ndes4ph, number of desaturations $\geq 4\%$ per hour; ndes5ph, number of desaturations $\geq 5\%$ per hour; odi3, oxygen desaturation index $\geq 3\%$ during sleep time; odi4, oxygen desaturation index $\geq 4\%$ during sleep time; pctle90, percentage of time desaturation was $\leq 90\%$; pctle92, percentage of time desaturation was $\leq 92\%$.

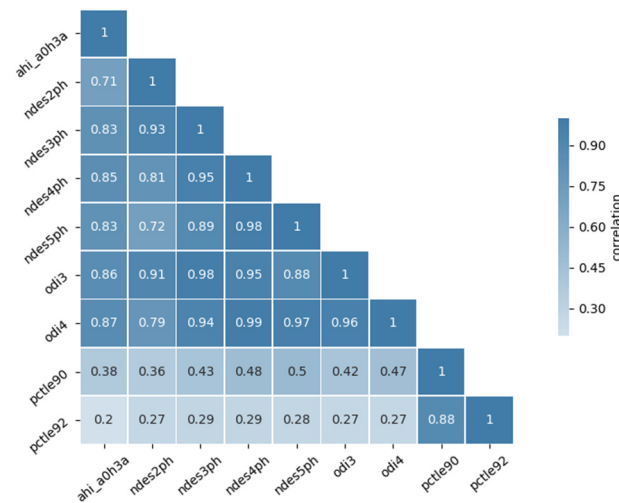


Figure 3. Cross-correlation matrix of the SpO2 signal features.

Finally, the authors calculated the number of samples required to minimize the overfitting of the model. From Figure 4 it is deduced that the overfitting problems are negligible beyond 120 samples. The training (blue) and the testing (green) datasets yield the same or similar accuracy as the number of samples increases. Considering that the CHAT has more than 400 registries, the overfitting of the model was mitigated during the training phase.

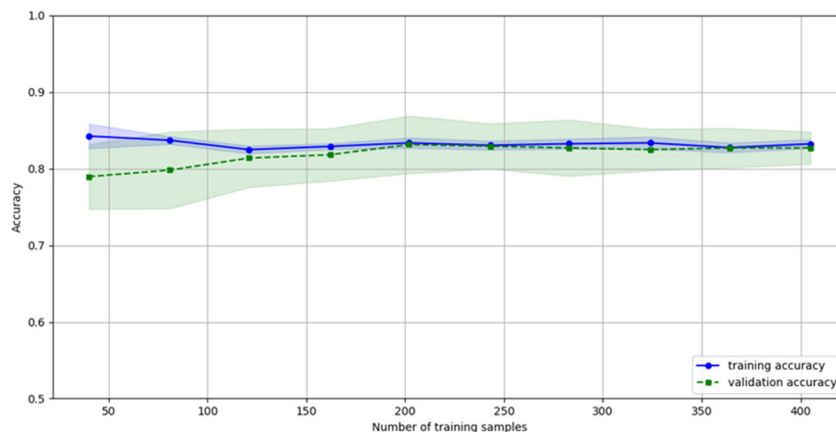


Figure 4. Impact of the number of samples in the overfitting of the model. The overlapping of the blue and green lines as well as the corresponding shaded areas reflect the lack of overfitting. Considering more than 120 samples mitigates overfitting.

3.2. Evaluating the Performance of the Tested Models

The performance of the three supervised learning classifiers was evaluated according to AUC, accuracy, recall, specificity and PPV. The results of the 15-fold process are summarized in Table 4. Focusing on the CHAT dataset, all three supervised methods yield similar AUC and accuracy. However, LR outperforms the other two regarding specificity and PPV. When applying them to the HG dataset, LR confirms its higher performance as well as higher replicability. The mean values of the different performance indicators are within the range of those yielded by more complex solutions (i.e., more than one biosignal being fed into the model and/or the use of DL algorithms to build the classifier, with the latter especially requiring domain-specific skills to design the architecture of the neural networks and to employ specific software packages) even though the accuracy is slightly lower. The fact that only a minor loss of performance is observed is very promising, especially considering the differences in age,

race and SpO₂ measuring devices. It is worth noting that the std value is equal to zero due to the small sample size of the HG dataset.

Table 4. Performance results for each algorithm in the CHAT and the HG datasets.

Dataset	Algorithm	AUC (Mean ± std)	Accuracy (Mean ± std)	Sensitivity (Mean ± std)	Specificity (Mean ± std)	PPV (Mean ± std)
CHAT	SVM	89.2 ± 7.7	82.9 ± 9.9	78.3 ± 13.5	87.4 ± 13.5	87.7 ± 12.2
	LR	90.2 ± 6.9	79.0 ± 7.2	62.0 ± 13.2	96.0 ± 5.4	94.3 ± 7.2
	AB	89.0 ± 6.7	82.1 ± 6.7	73.2 ± 11.8	90.9 ± 9.3	90.2 ± 9.8
HG	SVM	68.3 ± 4.3	66.7 ± 4.9	80.8 ± 13.6	52.5 ± 6.8	62.8 ± 4.2
	LR	85.2 ± 0.0	75.0 ± 0.0	62.5 ± 0.0	87.5 ± 0.0	83.3 ± 0.0
	AB	79.9 ± 1.3	74.6 ± 2.8	86.7 ± 3.1	62.5 ± 4.6	69.9 ± 2.7

STD, standard deviation; CHAT, Childhood Adenotonsillectomy Trial; HG, Hospital General de Valencia; SVM, support vector machine; LR, logistic regression; AB, AdaBoost; AUC, area under the curve; PPV, positive predictive value.

4. Discussion

The main objective of this study was to develop a screening tool, composed of a commercially available measuring device and an ML-based classifier, capable of identifying children at risk of suffering OSAS and feasible to be applied on an outpatient basis in the asymptomatic obese pediatric population. The results reveal that the proposed solution based on exclusively measuring the SpO₂ signal, calculating the ODI and the NDES features and applying an LR-based classifier outperforms SVM and AdaBoost-based solutions. The performance is within the range of results achieved by methods recently developed in symptomatic adult and pediatric population and where DL-based solutions are predominant (see Table 1). Interestingly, only a minor loss of performance is observed when the LR-based classifier is applied to a second independent dataset where the participants underwent unattended NOS at home.

The prevalence of OSAS in the general pediatric population can range between 0.2–4%, but it can reach more than 50% in the obese pediatric population. Regrettably, this syndrome is not limited to a deterioration of the sleep quality of an individual, as it can cause severe health consequences (e.g., behavior regulation, compliance, aggression, impulsivity, hyperactivity, anxiety, depressed mood, emotion regulation and neurocognitive deficits) [3,4]. The current methods for diagnosing this syndrome are complex and disturbing, especially for the pediatric population, compromising the quality and reproducibility of the results. Out of all of them, full-night PSG remains the gold standard method. Besides being a very disturbing procedure, it is expensive because it requires special equipment and trained personnel to correctly detect, identify and evaluate the events from the recorded signals. Even if specialized software is available with the hardware, experts are still required to validate the results.

Alternative methods have been proposed to make these tests more patient-friendly and to reduce the workload and the diagnostic discrepancies between experts. One of the most promising due to its ease of use and low cost is the NOS, which mainly focuses on measuring the SpO₂ signal, coupled with an ML- or DL-based classifier. Table 1 presents a summary of a number of these tools recently developed for symptomatic adults [16–27] and the pediatric population [28–34].

One major limitation of developing automated interpreting systems is choosing the most clinically meaningful biosignal. Multiple studies have tested different signal combinations to improve the performance of the classifier in detecting the OSAS severity [17,19,24–26,34]. Even the ECG signal, which is a commonly used parameter, has recently been questioned in terms of its continuous superior performance when compared with others [15,24]. In this context, an interesting study performed by Pathinarupothi et al. [24] showed that SpO₂ outperformed ECG when a DL-based classifier was applied to each of them. The second major limitation is extracting the most distinguishable features. For example, despite the clinical relevance of ODI and of cumulative time of desaturation index (CT),

not all the studies employ them [16,18,20,22,28,30,32,34]. Even when performing similar calculations, the extracted features are different or are used differently. For example, several papers used time, frequency and nonlinear domain calculations, but few or even none of them match [22,29–31,34]. There is also disparity when performing spectral calculations. Two recently published studies, one in adults and the other in the pediatric population, have developed tools based on images (i.e., spectrogram) [26] or based on the resulting statistic parameters [28], respectively.

Unfortunately, it is very common that these tools present a strong dataset dependency reflected by a drastic reduction of their performance when applied to a second independent dataset. Consequently, this makes it complex to compare the performance between tools and to widely implement them. DL-based solutions can overcome these limitations to an extent by using unsupervised learning algorithms to perform pattern recognition and data interpretation. However, the need for big datasets as well as domain-specific skills and software to develop and maintain these classifiers makes this approach more expensive, which limits its implementation outside clinical research where the budget of healthcare providers is already tight. In this context, the authors have focused on the development of an affordable screening tool that can be easily implementable in an obesity assessment outpatient consult to identify those asymptomatic children and adolescents who are at higher risk of suffering OSA. This is especially relevant not only for prescribing personalized physical activity strategies to fight against obesity but also for following up. This tool is composed of a commercially available device for measuring transcutaneous oxygen levels in the blood (SpO₂) and an ML-based classifier for data interpretation based on the extracted features. Out of the three tested ML algorithms (i.e., LR, SVM and AB) the LR has shown better performance on the CHAT test set and after cross-validation in a second independent dataset (i.e., HG dataset). The LR outperforms the other two regarding specificity and PPV and underperforms regarding sensitivity, while all three have similar accuracy and AUC (see Table 4). This performance is promising, especially when considering that they are within the range of similar tools developed in symptomatic adults and pediatric population and using DL algorithms (e.g., neural networks). For those where an LR classifier was used [29,31], the accuracy levels are similar. However, the proposed tool has substantially higher specificity in contrast with a significantly lower sensitivity. In other words, the proposed tool is very reliable in identifying healthy patients in contrast with the others [29,31], which are better in identifying subjects at risk of suffering OSA. The findings were expected according to the purpose with which each tool was built. While [29,31] focus on determining the OSA severity of symptomatic pediatric patients, the proposed tool is reliable in identifying the healthy subjects which can immediately be prescribed with personalized physical activity-based strategies. Those which might be at risk can be confirmed or finally discarded during follow up.

Cross-validation of this tool was performed with a second independent dataset (i.e., HG). Interestingly, only a minor loss of performance was observed in the LR-based solution despite the significant differences between both datasets (e.g., age, race, the prevalence of obesity and the SpO₂ measuring device). It is also relevant to point out that while the readings in CHAT were hospital-based, those in HG were home-based. Even though this loss is observed in all five indicators, specificity and PPV are the indicators that present greater reductions; this is in contrast with accuracy, recall and AUC, which only suffer a slight decrease (see Table 4). It is worth noting that the small sample size causes the prediction of all the models derived from the 15-fold process to yield identical results and therefore explains why the standard deviation is equal to 0. Out of those listed in Table 1, only the study performed by Vaquerizo-Villar et al. [32] performed cross-validation on an independent dataset. The accuracy level was 76% despite using a DL-based algorithm that was trained using power spectrum density, odi3, age and sex and targeted symptomatic children and adolescents.

Despite the need to increase the sample size of the HG dataset, these results are promising and indicate that it is worth continuing the development and optimization of the developed tool, which is expected to benefit clinicians as well as the pediatric population and their families. The former will benefit by having a tool that does not increase their workload; is easy to use and maintain; is cheap,

robust and reliable; and reduces diagnosis variability between experts. The latter will benefit from a minimally disturbing tool capable of being utilized at home in an unattended manner.

5. Conclusions

In the recent past, alternative methods based on SpO₂ measurements have been proposed to diagnose OSAS in symptomatic adults and the pediatric population. However, none of them have targeted the asymptomatic obese population even though OSAS prevalence can reach up to over 50%.

The proposed tool has shown promising results even when applied to a second independent dataset where the tests were performed on a very distinct cohort and in a home-based setting instead of a hospital-based one. Additionally, major differences were also present regarding the SpO₂ measuring device. The performance yield suggests that the methodology employed can be implementable abroad. Clinicians and, in particular, the pediatric population will benefit from a tool like the one developed in this manuscript.

Author Contributions: J.M.C. worked on feature extraction of SpO₂ raw signal as well as in the training, testing and validation of the classifier. J.Á.-P. recruited the patients from the Hospital General Universitario of Valencia, performed the respiratory polygraphy and gave the clinical diagnosis. I.C. worked on feature extraction of SpO₂ raw signal as well as in the training, testing and validation of the classifier. F.P. contributed to the performance of the respiratory polygraphy and building the dataset of the Hospital General Universitario of Valencia. P.R. supervised and contributed to the biosignal processing and the interpretation of the results applying machine learning algorithms. All authors contributed to writing the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Generalitat Valenciana “Ayudas Prometeo para grupos de investigación de excelencia”, grant number 2016/084”

Acknowledgments: The authors will like to acknowledge the collaboration received from Empar Lurbe, (MD) which is currently the Head of Pediatric Department at Consorcio Hospital General Universitario of Valencia.

Conflicts of Interest: The authors have no conflict of interest.

References

- Berry, R.B.; Budhiraja, R.; Gottlieb, D.J.; Gozal, D.; Iber, C.; Kapur, V.K.; Marcus, C.L.; Mehra, R.; Parthasarathy, S.; Quan, S.F.; et al. Rules for scoring respiratory events in sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. *J. Clin. Sleep Med.* **2012**, *8*, 597–619. [[CrossRef](#)] [[PubMed](#)]
- Grigg-Damberger, M.M. The AASM Scoring Manual Four Years Later. *J. Clin. Sleep Med.* **2012**, *8*, 323–332. [[CrossRef](#)] [[PubMed](#)]
- Blechner, M.; Williamson, A.A. Consequences of Obstructive Sleep Apnea in Children. *Curr. Probl. Pediatr. Adolesc. Health Care* **2016**, *46*, 19–26. [[CrossRef](#)] [[PubMed](#)]
- Tauman, R.; Gozal, D. Obesity and obstructive sleep apnea in children. *Paediatr. Respir. Rev.* **2006**, *7*, 247–259. [[CrossRef](#)]
- Aurora, R.N.; Zak, R.S.; Karippot, A.; Lamm, C.I.; Morgenthaler, T.I.; Auerbach, S.H.; Bista, S.R.; Casey, K.R.; Chowdhuri, S.; Kristo, D.A.; et al. Practice Parameters for the Respiratory Indications for Polysomnography in Children. *Sleep* **2011**, *34*, 379–388. [[CrossRef](#)]
- Marcus, C.L.; Brooks, L.J.; Draper, K.A.; Gozal, D.; Halbower, A.C.; Jones, J.; Schechter, M.S.; Sheldon, S.H.; Spruyt, K.; Ward, S.D.; et al. Diagnosis and management of childhood obstructive sleep apnea syndrome. *Pediatrics* **2012**, *130*, 576–584. [[CrossRef](#)]
- Pulmonology, S.o.P.; Syndrome, S.o.O.S.A. Clinical Practice Guideline: Diagnosis and Management of Childhood Obstructive Sleep Apnea Syndrome. *Pediatrics* **2002**, *109*, 704–712.
- Wise, M.S.; Nichols, C.D.; Grigg-Damberger, M.M.; Marcus, C.L.; Witmans, M.B.; Kirk, V.G.; D’Andrea, L.A.; Hoban, T.F. Executive Summary of Respiratory Indications for Polysomnography in Children: An Evidence-Based Review. *Sleep* **2011**, *34*, 389–398. [[CrossRef](#)]
- Dehlink, E.; Tan, H.-L. Update on paediatric obstructive sleep apnoea. *J. Thorac. Dis.* **2016**, *8*, 224–235. [[CrossRef](#)]

10. Alonso-Álvarez, M.L.; Terán-Santos, J.; Ordax Carbajo, E.; Cordero-Guevara, J.A.; Navazo-Egüia, A.I.; Kheirandish-Gozal, L.; Gozal, D. Reliability of home respiratory polygraphy for the diagnosis of sleep apnea in children. *Chest* **2015**, *147*, 1020–1028. [[CrossRef](#)]
11. Portable Monitoring Task Force of the American Academy of Sleep Medicine. Clinical Guidelines for the Use of Unattended Portable Monitors in the Diagnosis of Obstructive Sleep Apnea in Adult Patients. *J. Clin. Sleep Med.* **2007**, *3*, 737–747. [[CrossRef](#)]
12. Chervin, R.D.; Hedger, K.; Dillon, J.E.; Pituch, K.J. Pediatric sleep questionnaire (PSQ): Validity and reliability of scales for sleep-disordered breathing, snoring, sleepiness, and behavioral problems. *Sleep Med.* **2000**, *1*, 21–32. [[CrossRef](#)]
13. Chervin, R.D.; Weatherly, R.A.; Garetz, S.L.; Ruzicka, D.L.; Giordani, B.J.; Hodges, E.K.; Dillon, J.E.; Guire, K.E. Pediatric Sleep Questionnaire: Prediction of Sleep Apnea and Outcomes. *Arch. Otolaryngol. Head Neck Surg.* **2007**, *133*, 216–222. [[CrossRef](#)] [[PubMed](#)]
14. Singh, S.; Khan, S.Z.; Singh, D.; Verma, S.; Talwar, A. The uses of overnight pulse oximetry. *Lung India* **2020**, *37*, 151–157. [[CrossRef](#)]
15. Mostafa, S.S.; Mendonça, F.; Ravelo-García, A.G.; Morgado-Dias, F. A Systematic Review of Detecting Sleep Apnea Using Deep Learning. *Sensors* **2019**, *19*, 4934. [[CrossRef](#)]
16. Lg, O.; Ambrogetti, A.; Sg, G. Prediction of sleep-disordered breathing by unattended overnight oximetry. *J. Sleep Res.* **1999**, *8*, 51–55. [[CrossRef](#)]
17. Alvarez, D.; Hornero, R.; Marcos, J.; Del Campo, F.; Lopez, M. Spectral analysis of electroencephalogram and oximetric signals in obstructive sleep apnea diagnosis. In Proceedings of the Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 3–6 September 2009; Volume 2009, pp. 400–403. [[CrossRef](#)]
18. Chung, F.; Liao, P.; Elsaid, H.; Islam, S.; Shapiro, C.M.; Sun, Y. Oxygen Desaturation Index from Nocturnal Oximetry: A Sensitive and Specific Tool to Detect Sleep-Disordered Breathing in Surgical Patients. *Anesth. Analg.* **2012**, *114*, 993–1000. [[CrossRef](#)] [[PubMed](#)]
19. Xie, B.; Minn, H. Real-time sleep apnea detection by classifier combination. *IEEE Trans. Inf. Technol. Biomed.* **2012**, *16*, 469–477. [[CrossRef](#)] [[PubMed](#)]
20. Almazaydeh, L.; Faezipour, M.; Elleithy, K. A Neural Network System for Detection of Obstructive Sleep Apnea Through SpO2 Signal Features. *IJACSA* **2012**, *3*. [[CrossRef](#)]
21. Schlotthauer, G.; Di Persia, L.E.; Larrateguy, L.D.; Milone, D.H. Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry. *Med Eng. Phys.* **2014**, *36*, 1074–1080. [[CrossRef](#)] [[PubMed](#)]
22. Gutiérrez-Tobal, G.C.; Álvarez, D.; Crespo, A.; del Campo, F.; Hornero, R. Evaluation of Machine-Learning Approaches to Estimate Sleep Apnea Severity From At-Home Oximetry Recordings. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 882–892. [[CrossRef](#)]
23. Mostafa, S.S.; Mendonça, F.; Morgado-Dias, F.; Ravelo-Garcia, A. SpO2 based Sleep Apnea Detection using Deep Learning. In Proceedings of the 2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES), Larnaca, Cyprus, 20–23 October 2017; pp. 91–96.
24. Pathinarupothi, R.K.; DharaPrathap, J.; Rangan, E.; Gopalakrishnan, E.; Vinaykumar, R.; Somank, P. Single Sensor Techniques for Sleep Apnea Diagnosis Using Deep Learning. In Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, USA, 23–26 August 2017; pp. 524–529.
25. Cen, L.; Yu, Z.L.; Kluge, T.; Ser, W. Automatic System for Obstructive Sleep Apnea Events Detection Using Convolutional Neural Network. In Proceedings of the Conference of the IEEE Engineering in Medicine and Biology Society, Honolulu, HI, USA, 18–21 July 2018; Volume 2018, pp. 3975–3978. [[CrossRef](#)]
26. Biswal, S.; Sun, H.; Goparaju, B.; Westover, M.B.; Sun, J.; Bianchi, M.T. Expert-level sleep scoring with deep neural networks. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1643–1650. [[CrossRef](#)]
27. Mendonça, F.; Mostafa, S.S.; Morgado-Dias, F.; Ravelo-García, A.G. An Oximetry Based Wireless Device for Sleep Apnea Detection. *Sensors* **2020**, *20*, 888. [[CrossRef](#)]
28. Gutierrez-Tobal, G.C.; Kheirandish-Gozal, L.; Alvarez, D.; Crespo, A.; Philby, M.F.; Mohammadi, M.; Del Campo, F.; Gozal, D.; Hornero, R. Analysis and classification of oximetry recordings to predict obstructive sleep apnea severity in children. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, 25–29 August 2015; Volume 2015, pp. 4540–4543. [[CrossRef](#)]

29. Vaquerizo-Villar, F.; Alvarez, D.; Gutierrez-Tobal, G.C.; Barroso-Garcia, V.; Kheirandish-Gozal, L.; Crespo, A.; Del Campo, F.; Gozal, D.; Hornero, R. Usefulness of discrete wavelet transform in the analysis of oximetry signals to assist in childhood sleep apnea-hypopnea syndrome diagnosis. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Jeju Island, Korea, 11–15 July 2017; Volume 2017, pp. 3753–3756. [CrossRef]
30. 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.; et al. Nocturnal Oximetry-based Evaluation of Habitually Snoring Children. *Am. J. Respir. Crit. Care Med.* **2017**, *196*, 1591–1598. [CrossRef]
31. Crespo, A.; Álvarez, D.; Kheirandish-Gozal, L.; Gutiérrez-Tobal, G.C.; Cerezo-Hernández, A.; Gozal, D.; Hornero, R.; Del Campo, F. Assessment of oximetry-based statistical classifiers as simplified screening tools in the management of childhood obstructive sleep apnea. *Sleep Breath. Schlaf Atm.* **2018**, *22*, 1063–1073. [CrossRef]
32. Vaquerizo-Villar, F.; Álvarez, D.; Kheirandish-Gozal, L.; Gutiérrez-Tobal, G.C.; Barroso-García, V.; Crespo, A.; del Campo, F.; Gozal, D.; Hornero, R. Utility of bispectrum in the screening of pediatric sleep apnea-hypopnea syndrome using oximetry recordings. *Comput. Methods Programs Biomed.* **2018**, *156*, 141–149. [CrossRef]
33. Vaquerizo-Villar, F.; Álvarez, D.; Kheirandish-Gozal, L.; Gutiérrez-Tobal, G.C.; Gómez-Pilar, J.; Crespo, A.; Del Campo, F.; Gozal, D.; Hornero, R. Automatic Assessment of Pediatric Sleep Apnea Severity Using Overnight Oximetry and Convolutional Neural Networks. In Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 633–636. [CrossRef]
34. Jiménez-García, J.; Gutiérrez-Tobal, G.C.; García, M.; Kheirandish-Gozal, L.; Martín-Montero, A.; Álvarez, D.; Del Campo, F.; Gozal, D.; Hornero, R. Assessment of Airflow and Oximetry Signals to Detect Pediatric Sleep Apnea-Hypopnea Syndrome Using AdaBoost. *Entropy* **2020**, *22*, 670. [CrossRef]
35. Dean, S.N.; Shriver-Lake, L.C.; Stenger, D.A.; Erickson, J.S.; Golden, J.P.; Trammell, S.A. Machine Learning Techniques for Chemical Identification Using Cyclic Square Wave Voltammetry. *Sensors* **2019**, *19*, 2392. [CrossRef]
36. 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.; et al. The Childhood Adenotonsillectomy Trial (CHAT): Rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population. *Sleep* **2011**, *34*, 1509–1517. [CrossRef]
37. Zhang, G.-Q.; Cui, L.; Mueller, R.; Tao, S.; Kim, M.; Rueschman, M.; Mariani, S.; Mobley, D.; Redline, S. The National Sleep Research Resource: Towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1351–1358. [CrossRef]
38. Morante-Vélez, F.; Ordax-Carbajo, E. *Procedimientos en Trastornos Respiratorios del Sueño*; RESPIRA-Fundación Española del Pulmon-SEPAR: Madrid, Española, 2010.
39. Verplancke, T.; Van Looy, S.; Benoit, D.; Vansteelandt, S.; Depuydt, P.; De Turck, F.; Decruyenaere, J. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Med. Inform. Decis. Mak.* **2008**, *8*, 56. [CrossRef]
40. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
41. Millman, K.J.; Aivazis, M. Python for Scientists and Engineers. *Comput. Sci. Eng.* **2011**, *13*, 9–12. [CrossRef]
42. Oliphant, T.E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9*, 10–20. [CrossRef]
43. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 56–61. [CrossRef]
44. SciPy: Open Source Scientific Tools for Python—ScienceOpen. Available online: <https://www.scienceopen.com/document?vid=ab12905a-8a5b-43d8-a2bb-defc771410b9> (accessed on 7 September 2020).

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).