# Automatic Sleep Staging in Children with Sleep Apnea using Photoplethysmography and Convolutional Neural Networks

Fernando Vaquerizo-Villar, Daniel Álvarez*, Jan F. Kraemer, Niels Wessel, Gonzalo C. Gutiérrez-Tobal, *Member, IEEE*, Eva Calvo, Félix del Campo, Leila Kheirandish-Gozal, David Gozal, Thomas Penzel, *Senior Member, IEEE*, Roberto Hornero, *Senior Member, IEEE*

***Abstract*. Sleep staging is of paramount importance in children with suspicion of pediatric obstructive sleep apnea (OSA). Complexity, cost, and intrusiveness of overnight polysomnography (PSG), the gold standard, have led to the search for alternative tests. In this sense, the photoplethysmography signal (PPG) carries useful information about the autonomous nervous activity associated to sleep stages and can be easily acquired in pediatric sleep apnea home tests with a pulse oximeter. In this study, we use the PPG signal along with convolutional neural networks (CNN), a deep-learning technique, for the automatic identification of the three main levels of sleep: wake (W), rapid eye movement (REM), and non-REM sleep. A database of 366 PPG recordings from pediatric OSA patients is involved in the study. A CNN architecture was trained using 30-s epochs from the PPG signal for three-stage sleep classification. This model showed a promising diagnostic performance in an independent test set, with 78.2% accuracy and 0.57 Cohen's kappa for W/NREM/REM classification. Furthermore, the percentage of time in wake stage obtained for each subject showed no statistically significant differences with the manually scored from PSG. These results were superior to the only state-of-the-art study focused on the analysis of the PPG signal in the automated detection of sleep stages in children suffering from OSA. This suggests that CNN can be used along with PPG recordings for sleep stages scoring in pediatric home sleep apnea tests.**

***Clinical Relevance*—This research establishes the usefulness of CNN to automatically score sleep stages in pediatric OSA patients using the PPG signal.**

## I. Introduction

Obstructive sleep apnea (OSA) is a highly prevalent sleep-related respiratory disorder in children (5%) [1]. OSA is characterized by recurrent apneas (breathing cessations) and hypopneas (airflow reductions), which derive in oxygen desaturations, arousals, and transitions between wakefulness and different sleep stages [1], [2]. As a consequence of sleep disturbance, the affected children suffer from behavioral and neurocognitive deficits [1], [2]. Hence, it is of the utmost importance to provide an early diagnosis of pediatric OSA that includes the characterization of the sleep architecture.

Nowadays, overnight polysomnography (PSG), together with the rules of the American Academy of Sleep Medicine (AASM), are used for the scoring of sleep stages and cardiorespiratory events [3]. PSG involves the recording of multiple biomedical signals, including electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), airflow, electrocardiogram (ECG), and blood oxygen saturation ($SpO_2$) from photoplethysmography (PPG) [3]. According to a visual inspection of EEG, EOG, and EMG, a sleep stage is assigned to each non-overlapping 30-s epoch [3]: rapid eye movement (REM) sleep, three levels of non-REM (NREM) sleep (N1, N2, and N3), and wake (W). Despite being considered as the gold standard for sleep assessment, PSG is complex and costly due to the necessary specialized equipment and trained staff [4]. PSG is also highly intrusive for children, which may derive in obtaining sleep recordings that are not representative of natural sleep, thus resulting in the need to repeat the diagnostic test [5].

To overcome these limitations, recent studies have proposed the automated analysis of EEG, ECG, EOG, PPG, actigraphy, and/or respiratory effort for sleep stage scoring in adult patients [6]. In this respect, the PPG signal is especially comfortable for children, as it measures blood volume changes in body tissues in a non-invasive way with a pulse oximeter probe [7]. The PPG signal can be used to estimate heart rate variability (HRV) [8], thus containing information of the autonomic nervous system activity associated to sleep stages [6]. In children, whose sleep architecture and cardiac activity

This work was supported by 'Ministerio de Ciencia, Innovación y Universidades - Agencia Estatal de Investigación' and 'European Regional Development Fund (FEDER)' under projects DPI2017-84280-R and RTC-2017-6516-1, by 'European Commission' and 'FEDER' under project 'POCTEP0702_MIGRAINEE_2_E', and in part by 'Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, (CIBER-BBN), Spain' through 'Instituto de Salud Carlos III (ISCII)' co-funded with FEDER funds, by Sociedad Española de Neumología y Cirugía Torácica under project 649/218, and by Sociedad Española de Sueño (SES) under project "Beca de Investigación SES 2019".

The work of Daniel Álvarez was supported by a "Ramón y Cajal" grant (RYC2019-028566-I) from the 'Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación' co-funded by the European Social Fund. F. Vaquerizo-Villar was in receipt of a 'Ayuda para contratos predoctorales para la Formación de Profesorado Universitario (FPU)' grant from the Ministerio de Educación, Cultura y Deporte (FPU16/02938). L. Kheirandish-Gozal and D. Gozal were supported by National Institutes of Health (NIH) grant HL130984, HL140548, and AG061824.

F. Vaquerizo-Villar, D. Álvarez, G. C. Gutiérrez-Tobal, E. Calvo, and R. Hornero are with the Biomedical Engineering Group, Universidad de Valladolid (e-mail: fernando.vaquerizo@gib.tel.uva.es) and CIBER-BBN (ISCIII), Spain.

J. F. Kraemer and N. Wessel are with the Department of Physics, Humboldt Universität zu Berlin, Germany (e-mail: wessel@physik.hu-berlin.de).

D. Álvarez, and F. del Campo are with the Hospital Universitario Río Hortega of Valladolid, Spain (e-mail: fsas@telefonica.net) and CIBER-BBN (ISCIII), Spain.

L. Kheirandish-Gozal and D. Gozal are with the Department of Child Health, The University of Missouri School of Medicine, Columbia, Missouri, USA (email: gozald@health.missouri.edu).

T. Penzel is with the Interdisciplinary Center of Sleep Medicine, Charité-Universitätsmedizin Berlin, Germany (e-mail: thomas.penzel@charite.de).

differ from adults [3], [9], only Dehkordi *et al*. [10] have conducted sleep staging using HRV features derived from PPG. However, PPG also contains information related to changes in cortical and respiratory activity during sleep stages that are not reflected in the HRV [6], [11], [12], which demands the application of additional methods.

In the present study, a convolutional neural network (CNN) is proposed to analyze raw PPG data during wake, NREM, and REM sleep. As a deep-learning algorithm, CNNs have the ability to automatically learn complex features from raw data [13]. In contrast to recurrent and fully connected networks, however, CNNs have a lower computational cost [13], which facilitates its integration in portable sleep monitoring devices. We hypothesized that CNNs could help to automatically extract the relevant information from the raw PPG signal associated to sleep stages in pediatric OSA patients. Therefore, our main objective is to evaluate the usefulness of a CNN architecture to detect wake, NREM, and REM sleep stages from the PPG signal in children suffering from pediatric OSA.

## II. Material and Methods

### A. Subjects and signals

The baseline dataset from the public multicenter Childhood Adenotonsillectomy Trial (CHAT) database was employed in this study [14], [15]. The entire protocol of the CHAT database is available in the supplementary material of Marcus *et al*. [14]. This dataset is composed of PSG studies of 453 children ranging from 5 to 10 years of age who were randomized for pediatric OSA treatment [15]. Cardiorespiratory events and sleep stages were scored in compliance with the AASM 2007 rules [3]. Accordingly, the apnea-hypopnea index (AHI) is calculated to establish pediatric OSA diagnosis.

Complete PPG signals from PSG-derived pulse oximetry were obtained for 366 subjects showing different sampling rates: 16, 100, 200, 256, and 512 Hz. In order to homogenize these, PPG signals were resampled to a common sampling frequency of 64 Hz [11]. Then, PPG signals were divided into 30-s non-overlapping epochs, being each epoch labelled as W, NREM, or REM using the sleep stages annotations scored by the technicians [14]. This dataset was split into three sets: training set (first 219 subjects, 60%), employed to train the CNN; validation set (73 following subjects, 20%), used to design the hyperparameters of the CNN architecture; and test set (last 74 subjects, 20%), used for performance assessment in an independent group. Table I shows clinical and polysomnographic data from the subjects under study.

TABLE I. Demographic and Polysomnographic Data of the Subjects Under Study

| | All | Training set | Validation set | Test set |
|---|---|---|---|---|
| **Subjects (n)** | 366 | 219 | 73 | 74 |
| **Age (years)** | 6 [5, 8] | 6 [5, 8] | 6 [5, 8] | 6 [5, 7] |
| **Males (n)** | 178 (48.6%) | 108 (49.3%) | 31 (42.5%) | 39 (52.7%) |
| **BMI (kg/m2)** | 17.2 [15.2, 21.7] | 17.3 [15.5, 21.4] | 19.2 [15.2, 22.9] | 16.0 [15.0, 19.9] |
| **AHI (e/h)** | 4.8 [2.7, 8.7] | 4.8 [2.7, 8.7] | 4.7 [2.5, 8.6] | 4.9 [3.2, 8.7] |
| **Wake (n)** | 113315 (25.3%) | 67963 (25.2%) | 23132 (26.2%) | 2220 (24.9%) |
| **NREM (n)** | 271556 (60.7%) | 164164 (60.8%) | 53183 (60.3%) | 54209 (60.9%) |
| **REM (n)** | 62434 (14.0%) | 37829 (14.0%) | 11943 (13.5%) | 12662 (14.2%) |
| **TRT (min)** | 591 [552, 660] | 595 [553, 662] | 588 [540, 646] | 592 [552, 648] |
| **TST (min)** | 464 [428, 494] | 471 [438, 496] | 451 [425, 482] | 458 [421, 500] |

Data are presented as median [interquartile range], n or %. BMI: Body Mass Index; AHI: Apnea-Hypopnea Index; e/h: events per hour; REM: Rapid Eye Movement; NREM: Non-REM; TRT: Total Recording Time; TST: Total Sleep Time

### B. Proposed CNN architecture

CNNs have shown its usefulness to analyze time series data during sleep [6], due to its multi-layer architecture with weights sharing, sparse connections, and dimensionality reduction elements [13]. Fig 1 shows the main components of the CNN architecture employed in this work.

The input section of the network consists of the PPG signal for the 30-s epoch (1920 samples) to be classified, concatenated with the PPG signal for the five preceding and the four following epochs, thus having a 19200x1 input vector.

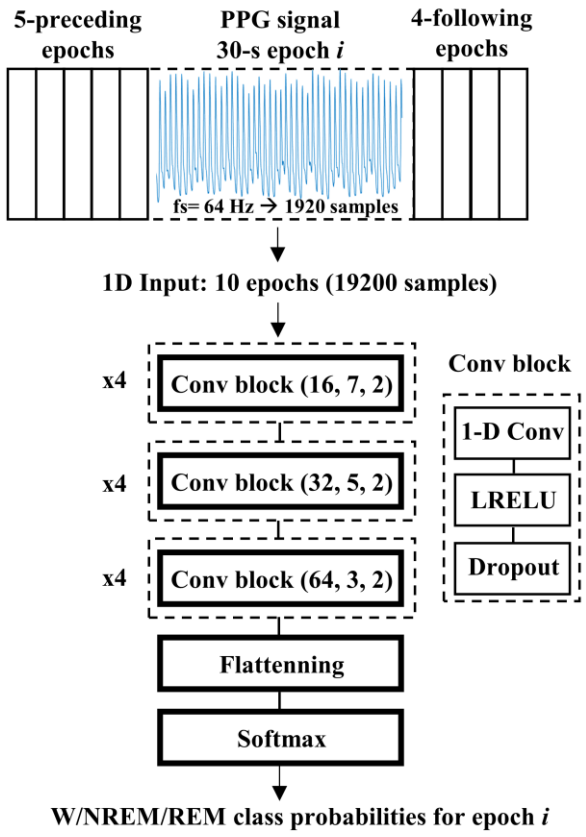


Figure 1. Overview of the proposed CNN architecture. Each convolutional block (conv block) includes a 1-D convolution (1-D conv), a LRELU activation function, and dropout.

This approach allows for better modeling of the temporal context used for sleep stages scoring [16].

The input is processed using 12 convolutional blocks, each one composed of a 1-D convolutional layer, an activation layer, and a dropout layer:

- 1-D convolutional layer (1-D conv). In this layer, feature maps are extracted from the data by using convolutional filters, named kernels [13]. In this study, the convolutional layer of the first four convolutional blocks had 16 filters with a kernel size of 7 (7x1), the following four blocks used layers with 32 filters of size 5x1, and the last four blocks used 64 filters of size 3x1. Increasing the number of filters and decreasing the kernel size with network depth is a common approach in CNNs. All the layers also included a stride of 2 to reduce dimensionality.
- Activation layer. In this layer, a leaky rectified linear unit (LReLU) activation function with a negative slope of 0.1 is applied to introduce nonlinearity to the extracted feature maps, which allows approximations to any universal function [13].
- Dropout layer. As the last layer of each convolutional block, node connections were randomly removed with a probability $p$=0.1, which is known as dropout, in order to minimize overfitting during the training phase [13].

After the last convolutional block, a flattening layer is firstly used to transform the 2-D feature maps into 1-D data. Then, a softmax activation function is applied to obtain the output of the network, i.e., the probability of belonging to each class (W/NREM/REM) for the input 30-s PPG epoch.

The implementation of the CNN architecture was programmed in Python using the Keras library with TensorFlow backend. Network weights were initialized using He-normal method, Adam algorithm was used with an initial learning rate of 0.0001 to update network weights at each iteration, and categorical cross entropy was employed as the loss function [13]. Using 50 reading queues [16], training data were fed in random order from different patients to the network using a batch size of 100, which allows for improving the convergence of the Adam algorithm [13].

### C. Statistical analysis

The overall agreement between the sleep stages predicted by the CNN architecture and those manually scored by the technicians was evaluated by means of confusion matrices, which were used to compute the Cohen's kappa index (kappa) and the 3-class accuracy (Acc). The performance for each individual class was measured by means of precision (positive predictive value, proportion of epochs assigned to the class that are true positives), recall (sensitivity, percentage of epochs belonging to the class rightly classified), and F1-score (harmonic mean of the precision and recall). In addition, the percentage of time in each sleep stage was obtained for each subject and compared with those from the standard PSG using intra-class correlation coefficient (ICC) and Wilcoxon signed-rank test, considering $p$-value<0.001 as significant.

## III. Results

### A. CNN model performance

Figure 2 shows the confusion matrix of the CNN model in the test set for the three-stage classification (W/NREM/REM). This model rightly classified 78.2% of the 30-s PPG epochs in the test set, with a kappa value of 0.57. For each individual sleep stage, the precision/recall/F1-score were 0.81/0.69/0.74 for wake, 0.79/0.91/0.85 for NREM sleep, and 0.64/0.41/0.50 for REM sleep, which were derived from the confusion matrix. Notice that higher performance metrics were obtained for NREM stages than for wake and REM stages.

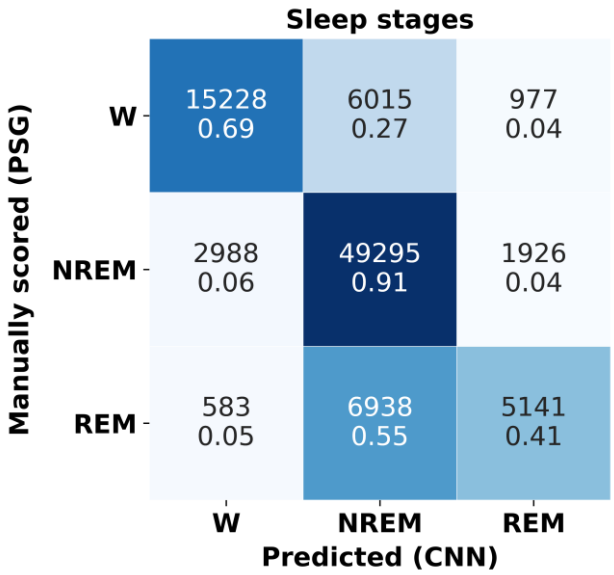


Figure 2. Confusion matrix of the CNN architecture in the test set. This matrix compares the sleep stages from standard PSG with the corresponding assignation using the CNN model.

### B. Estimation of polysomnographic parameters

Table II shows the comparison of the percentage of time in W/NREM/REM stages with those obtained during PSG. Notice that Wake (%) estimated by our CNN-based approach reached high agreement with Wake (%) manually obtained during PSG, as derived from the median [interquartile range] error (-3.4 [8.8] %), ICC (0.59), and $p$-value (0.002).

## IV. Discussion

In this preliminary study, we propose a CNN architecture to automatically detect wake, NREM, and REM sleep stages from PPG in pediatric OSA patients. To our knowledge, the application of deep-learning techniques is novel to detect sleep stages in pediatric patients.

Our proposal reached a high performance, with 78.2% Acc and 0.57 kappa for W/NREM/REM classification. Specifically, the obtained kappa value (in the range 0.41-0.60) indicates that there is a moderate agreement between our automatic CNN-based PPG scoring and manual scoring from PSG [17]. According to this agreement, our approach could be used to analyze NREM and REM characteristics in at-home simplified tests for pediatric OSA diagnosis [4], such as

TABLE II. Estimation of the percentage of time in W, NREM, and REM stages

| | CNN | Error (CNN-PSG) | ICC | $p$-value |
|---|---|---|---|---|
| **Wake (%)** | 19.3 [12.5, 27.2] | -3.4 [-8.1,0.7] | 0.59 | 0.002 |
| **NREM (%)** | 72.3 [61.8-79.3] | 9.8 [2.1-15.3] | 0.31 | <0.001 |
| **REM (%)** | 8.4 [5.3, 12.6] | -5.8 [-8.9,.2.5] | 0.24 | <0.001 |

Data are presented as median [interquartile range] or n. CNN: Convolutional Neural Network; PSG: Polysomnography; ICC: Intra-class Correlation Coefficient; REM: Rapid Eye Movement; NREM: Non-REM.

polygraphy and oximetry [18], [19], which do not include EEG. As aforementioned, the values of precision/recall/F1-score were higher in the NREM class than in the W and REM classes. The lower performance metrics in W and REM classes may be explained by the slight trend of the CNN to assign W and REM epochs to the NREM class, as also observed in the CNN-PSG differences in terms of percentage of time in each sleep stage shown in Table II. The Wake (%) estimated by our proposal also showed a high agreement with PSG, as there were no statistically significant differences with the manually scored Wake (%) from PSG. Accordingly, our CNN-based approach could be used to determine the total sleep time (TST) in polygraphy and oximetry tests [18], [19], as it can be derived from the Wake (%).

Previous studies shown the usefulness of deep-learning approaches to automatically score sleep stages from raw physiological signals in adult patients, outperforming feature-engineering approaches [6]. From these studies, it is of note the work done by Korkalainen *et al*. [11], who implemented a CNN combined with a recurrent neural network (RNN) for PPG-based sleep stage classification in the adult context, reaching 80.1% Acc and 0.65 kappa for W/NREM/REM classification. The TST estimated by Korkalainen *et al*. [11] also reached a high agreement with PSG (*p*-value of 0.03). In contrast to these studies, our study applied CNN to detect the three main levels of sleep (W/NREM/REM) in children. In this respect, sleep staging is more challenging in children, as sleep architecture and cardiorespiratory activity change during the childhood [3], [9]. In addition, our CNN model is easier to integrate in a low-cost portable oximeter than RNN-based architectures since it has a lower computational load.

In pediatric patients, Dehkordi *et al*. [10] trained two support vector machine-based classifiers with common HRV time and spectral features extracted from the PPG signal to differentiate between wake and sleep segments (W/Sleep) and between NREM and REM segments (NREM/REM), reaching 77% Acc (W/Sleep) and 80% Acc (NREM/REM). In contrast, our current study showed a higher diagnostic performance with a CNN trained with raw PPG data: 78% Acc for 3-class (W/NREM/REM) classification, 88% Acc to differentiate wake from NREM and REM sleep epochs (W/Sleep), 80% Acc to differentiate NREM from wake and REM, and 88% to differentiate REM from wake and NREM.

This study presents some limitations. First, the dataset employed in this study does not include no-OSA children (AHI<1 e/h). The inclusion of these subjects would be useful to compare the performance of the CNN model in non-pathological sleep. Further evaluation would be also required to assess the effect of demographic factors, such as age, sex, and BMI. Additionally, the use of RNN and attention mechanisms may help improve the automatic sleep staging at the cost of higher computational complexity. Finally, it would be useful to validate this proposal using PPG signals acquired at patient's home.

In summary, a CNN-based deep-learning architecture has shown usefulness to automatically identify wake, NREM, and REM sleep stages from raw PPG data in pediatric OSA patients. In addition, the Wake (%) estimated by our CNN-based approach also showed high agreement with PSG. Therefore, we can conclude that CNN-based PPG approaches could be potentially used to automatically score sleep stages in pediatric sleep apnea testing.

## REFERENCES


[1] S. J. Hunter, D. Gozal, D. L. Smith, M. F. Philby, J. Kaylegian, and L. Kheirandish-Gozal, "Effect of sleep-disordered breathing severity on cognitive performance measures in a large community cohort of young school-aged children," *Am. J. Respir. Crit. Care Med.*, vol. 194, no. 6, pp. 739–747, 2016.

[2] O. S. Capdevila, L. Kheirandish-Gozal, E. Dayyat, and D. Gozal, "Pediatric obstructive sleep apnea: Complications, management, and long-term outcomes," *Proc. Am. Thorac. Soc.*, vol. 5, no. 2, pp. 274–282, 2008.

[3] C. Iber, S. Ancoli-Israel, A. Chesson, and S. F. Quan, "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specification," *J. Clin. Sleep Med.*, vol. 3, no. 7, p. 752, 2007.

[4] H. L. Tan, L. Kheirandish-Gozal, and D. Gozal, "Pediatric home sleep apnea testing slowly getting th ere!," *Chest*, vol. 148, no. 6, pp. 1382–1395, 2015.

[5] E. S. Katz, R. B. Mitchell, and C. M. D. Ambrosio, "Obstructive Sleep Apnea in Infants," *Am. J. Respir. Crit. Care Med.*, vol. 185, no. 8, pp. 805–816, 2012.

[6] O. Faust, H. Razaghi, R. Barika, E. J. Ciaccio, and U. R. Acharya, "A review of automated sleep stage scoring based on physiological signals for the new millennia," *Comput. Methods Programs Biomed.*, vol. 176, pp. 81–91, 2019.

[7] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, 2007.

[8] S. Lu *et al.*, "Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information?," *J. Clin. Monit. Comput.*, vol. 22, no. 1, pp. 23–29, 2008.

[9] D. Y. T. Goh, P. Galster, and C. L. Marcus, "Sleep architecture and respiratory disturbances in children with obstructive sleep apnea," *Am. J. Respir. Crit. Care Med.*, vol. 162, no. 2 I, pp. 682–686, 2000.

[10] P. Dehkordi, A. Garde, W. Karlen, D. Wensley, J. M. Ansermino, and G. A. Dumont, "Sleep stage classification in children using photo plethysmogram pulse rate variability," *Comput. Cardiol. (2010).*, vol. 41, no. January, pp. 297–300, 2014.

[11] H. Korkalainen *et al.*, "Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea," *Sleep*, vol. 43, no. 11, pp. 1–10, 2020.

[12] G. B. Papini, P. Fonseca, M. M. V. Gilst, J. W. M. Bergmans, R. Vullings, and S. Overeem, "Respiratory activity extracted from wrist-worn reflective photoplethysmography in a sleep-disordered population," *Physiol. Meas.*, vol. 41, no. 6, 2020.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[14] C. L. Marcus *et al.*, "A Randomized Trial of Adenotonsillectomy for Childhood Sleep Apnea," *N. Engl. J. Med.*, vol. 368, no. 25, pp. 2366–2376, 2013.

[15] S. Redline *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*, vol. 34, no. 11, pp. 1509–1517, 2011.

[16] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J. F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," *Biomed. Signal Process. Control*, vol. 42, pp. 107–114, 2018.

[17] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[18] M. L. A. Álvarez *et al.*, "Reliability of Respiratory Polygraphy for the Diagnosis of Sleep Apnea-Hypopnea Syndrome in Children," *Arch. Bronconeumol. ((English Ed.*, vol. 44, no. 6, pp. 318–323, 2008.

[19] F. del Campo, A. Crespo, A. Cerezo-Hernández, G. C. Gutiérrez-Tobal, R. Hornero, and D. Álvarez, "Oximetry use in obstructive sleep apnea," *Expert Rev. Respir. Med.*, vol. 12, no. 8, pp. 665–681, 2018.