Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 225 (2023) 3805–3812

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

A survey on pre-training requirements for deep learning models to detect obstructive sleep apnea events

Ángel Serrano Alarcón [a]*, Maksym Gaiduk[b]*, Natividad Martínez Madrid[a], Ralf Seepold[b]

[a]*Reutlingen University, Alteburgstraße 150, Reutlingen 72762, Germany*
[b]*HTWG Konstanz – University of Applied Sciences, Alfred-Wachtel-Str.8, Konstanz 78462, Germany*

Abstract

The development of automatic solutions for the detection of physiological events of interest is booming. Improvements in the collection and storage of large amounts of healthcare data allow access to these data faster and more efficiently. This fact means that the development of artificial intelligence models for the detection and monitoring of a large number of pathologies is becoming increasingly common in the medical field. In particular, developing deep learning models for detecting obstructive apnea (OSA) events is at the forefront. Numerous scientific studies focus on the architecture of the models and the results that these models can provide in terms of OSA classification and Apnea-Hypopnea-Index (AHI) calculation. However, little focus is put on other aspects of great relevance that are crucial for the training and performance of the models. Among these aspects can be found the set of physiological signals used and the preprocessing tasks prior to model training. This paper covers the essential requirements that must be considered before training the deep learning model for obstructive sleep apnea detection, in addition to covering solutions that currently exist in the scientific literature by analyzing the preprocessing tasks prior to training.

Keywords: sleep efficiency; sleep study; subjective sleep assessment

* Corresponding authors. Tel.: +49 7121 271 - 4117.
E-mail address: Angel.Serrano_alarcon@reutlingen-university.de

10.1016/j.procs.2023.10.376

1. Introduction

Artificial intelligence and data management are becoming part of every aspect of people's lives. It is becoming increasingly common to find intelligent algorithms that have been trained with data collected daily in everyday tasks. One of the fields where artificial intelligence can provide the most significant benefits is the field of medicine. The detection of physiological events of interest can be crucial for the diagnosis and monitoring of various diseases. Numerous solutions already exist today that use artificial intelligence to diagnose diseases [1]–[5].

One of the fields where the development of artificial intelligence models plays a significant role is sleep medicine. The problems derived from sleep can cause other pathologies and alter other physical conditions, putting people's health at risk, besides generating tiredness during the day in people who suffer from it. One of the common sleep disorders is obstructive sleep apnea. Polysomnography is the gold standard for detecting obstructive sleep apnea. However, despite its accuracy, it has numerous disadvantages that are widely known, such as the long waiting lists of patients to be monitored in sleep laboratories during the night to check whether they suffer from apnea or not [6]. This also implies an economic and personal cost since polysomnography tests require on-site personnel during the test and clinicians for the subsequent analysis of the results.

Because of the above, for some time now, ways have been developed to detect sleep apnea more quickly and in a less invasive way. One of these solutions is the use of portable devices that can be taken home by patients and the sleep test performed there [7], [8]. This would lead to the decongestion of waiting lists in sleep laboratories and provide the patient with greater comfort.

In general, new sleep apnea detection techniques involve intelligent or automatic algorithms that automatically recognize apnea events. The best-known and most widely used are machine learning models. Specifically, deep learning models allow us to avoid the feature-extracting process and thus avoid the high level of expertise required in this medical field [9]. Deep learning models can be trained with a large amount of data and provide results close to those of expert clinicians.

Although numerous solutions use deep learning to detect apnea events, there still needs to be a standard on the ideal methodology to develop these models for these classification tasks. One of the most commonly used models is convolutional neural networks (CNNs) or recurrent neural networks (RNNs) [10], [11]. However, the choice of model architecture is one of many things that need to be addressed when it comes to developing deep-learning models for apnea event detection. Many other factors can influence model development and performance. One such factor is the biomedical datasets to be used for sleep apnea detection. Deciding between the ideal number of signals and the ideal combination is a complex task. The features of the signals are also relevant, such as the signals' sampling rate, whether the whole signal is used to feed the algorithm, or whether the signal is divided into windows of a specific duration for training the algorithm [12].

As can be seen, there are several factors to be taken into account prior to training the model. There is no unanimity among the existing solutions in the scientific literature on the best selection of this set of factors. Usually, the well-known trial-and-error method is used. This can be problematic since the development of deep learning models should have an engineering approach and avoid this type of method as much as possible.

According to the literature, some solutions base the models' development on the selected architecture [13], [14]. However, the classification of apnea events is complex and involves the selection of several relevant factors, such as the type of classification to be performed by the model. Most models are based on binary classification, where a distinction is made between apneas and non-apneas without distinguishing between hypopneas. On the other hand, some models also base their modus operandi on the classification of different events. In this case, it would be apnea, non-apnea and hypopneas[15], [16].

The development of deep learning models for apnea event detection is associated with numerous prerequisites that need attention prior to model training. In this paper, we emphasize these requirements and cover various proposals in the literature to make a comparison between the different solutions and evaluate the selection of these factors.

2. Methods

This section presents the methodology used for the study of the prerequisites that are necessary for the development of deep learning models for OSA detection. First, the main requirements the authors consider necessary to address before training are identified. These requirements are:

- Database selection
- Set of signals definition
- Sampling rate
- Preprocessing
    - Filtering
    - Artefact Removal
    - Normalization
    - Dataset balancing
- Classification approach
    - Patient-based
    - Window-based

Based on these prerequisites, a search is done in the scientific literature on the development of deep learning models that address these issues. In order to search for relevant studies that fit our scope of study, searches in different databases such as PubMed, ScienceDirect and IEEE were performed. Since the authors of this work intend to analyse and expose recent solutions to address the preprocessing stage before model training as accurately as possible, the selected scientific papers are not older than 2017. The publications were included based on the information they contain in relation to the requirements outlined above.

After the scientific review, the main characteristics of the works found and selected will be presented for further analysis and assessment. It should be noted that the main objective of this work is not based on the model selection but on the selection of all the requirements listed above.

The results obtained by the selected scientific papers will also be presented and analyzed. However, good results in model performance are not the only relevant outcome. Other aspects of interest, such as the explainability of the model, should also be addressed. However, this aspect is outside the scope of this work.

One of the main aspects studied in this manuscript is the preprocessing or not of the signals since it is relevant to study whether a model can perform well by simply using the raw signals. Details of the implementation of signal pre-processing are also important and are part of the analysis.

3. Results

This section presents the results of the performed scientific literature search. After reviewing the scientific papers, thirteen publications were selected after considering that they met our requirements established in section 2. The data are presented in Tables 1 and 2. Table 1 shows the databases used by the scientific papers, the set of signals used, the sampling rate used for training the models (not the sampling rate during signal acquisition) and all the processing tasks prior to the training of the models.

Table 2 shows the architectures of the deep learning models used for the classification of apnea events. The results obtained based on the metrics used by the authors of the scientific papers are also presented. As a general rule, accuracy, sensitivity, and specificity are the most used to demonstrate the goodness of the model in the classification tasks.

Table 1. Main prerequisites prior to the training of the models by the solutions proposed after the review of the literature.

Publication	Database	Set of signals	Sampling Rate (Hz)	Preprocessing
Chang H. et al. 2020 [17]	PhysioNet Apnea-ECG Database 35 subjects	Single lead ECG	100	- Band-pass filtering (0.5 Hz to 15 Hz) - Standardization
Dey D. et al. 2017 [16]	PhysioNet Apnea-ECG Database 35 subjects	Single lead ECG	100	-
Drzazga J. et al. 2021 [15]	SHHS-1 PhysioNet Sleep	Oronasal airflow (thermocouple) Thoracic respiratory effort (RIP band) Abdominal respiratory effort (RIP band)	-	- Offset removal by subtracting mean value of the signal from each sample - Signals are scaled to even out differences between patients - Filtering with 1.25 Hz low-pass filter; phase alignment.
Haidar R. et al. 2018 [18]	MESA	Nasal airflow Thoracic and Abdominal	32	-Data normalization
Kristiansen S. et al. 2021[7]	A3 Study (over 7400 hours from 579 patients)	Nasal cannula SpO2 RIP chest RIP abdomen	1	- Artefacts removal - Data partioned into non-overlapping 60-seconds periods - Dataset balancing
Kwon H. et al. 2022[19]	Proprietary dataset 36 recordings	Impulse-radio ultra-wideband (IR-UWB) radar	20 fps (radar)	- Set windows of 20 seconds duration - Downsizing images to 80 x 300 pixels;
Zhang J. et al. 2021 [20]	Physionet Apnea-ECG dataset 35 subjects	Single lead ECG	100	- Chebyshev type-II band-pass filter (5-11 Hz) - Set windows of 10 seconds duration
Pathinarupothi R. et al. 2017 [21]	Physionet apnoea-ECG dataset 35 subjects	SpO2 IHR	100	- Set windows of 60 seconds duration
Urtnasan E. et al. 2018 [22]	Standard full-night PSG data 82 subjects	Single lead ECG	200	- Band-pass filtering (5 Hz to 11 Hz) - Set windows of 10 seconds duration
Nikkonen S. et al. 2021 [23]	In-lab PSG 887 subjects	SpO2 Thermistor Airflow Nasal pressure-airflow Thorax respiratory effort	4	- Lowpass filtered (2Hz cutoff frecuency) - Downsampling - Set windows of 30 seconds duration
Zhang H. et al. 2022 [24]	Physionet dataset 994 recordings SHHS-1 1000 recordings	Multichannel (12)	200	-
Van Steenkiste T. et al. 2019 [25]	SHHS-1	Abdores Thorres ECG Derived Respiration (EDR)	5	- Fourth-order low-pass zero-phase-shift Butterworth filter witha cut-off frequency of 0.7 Hz. - Motion artifacts removed by subtracting a moving average filtered signal with a width of 4 seconds from the original signal. - Downsampling - Set windows of 30 seconds duration
Urtnasan E. et al 2020 [26]	SMC Sleep Apnea Dataset	Single lead ECG	200	- Set windows of 30 seconds duration

As can be seen in Table 1, there is diversity in the databases used. The most used databases are the PhysioNet Apnea-ECG Database and SHHS-1, used in 4 and 3 scientific publications, respectively. Another dataset that is also commonly used is MESA, which has been used by [18]. Usually, the most significant difference in these databases is the number of patients. PhysioNet Apnea-ECG Database consists of only 35 patients, while other databases used by [23] consist of 887 patients. This fact may not seem remarkable, but a more significant number of patients implies greater variability between the signals used and allows a more significant data set to train, validate and test the model.

Regarding the set of signals, a significant disparity can also be seen between the solutions analyzed. Only in the case of those papers that use a single lead ECG is there a certain similarity between the proposed methodologies. The rest of the solutions use different combinations of signals, being [24] the one that used the most significant number of signals with a total of 12. The number of physiological signals used has different readings. On the one hand, the fewer signals used, the greater the comfort for the patient can be achieved and the faster the training of the deep learning models can be performed. However, it must be borne in mind that a reduced number of signals is only sometimes a good indication as it can lead to a decrease in the quality of the outcome. Since, for example, for the detection of hypopneas according to the established rules, it is necessary to have SpO2, and some solutions do not use it or an alternative to it. Another critical factor that affects the set of signals used is the sampling rate. Here we have to distinguish between the sampling rate used for signal acquisition and the sampling rate used for model training. In this work, we focus on the sampling rate used to develop the models and the one chosen after the processing tasks. Therefore, Table 1 shows the sampling rate of the signals that feed the deep-learning models. Generally, when it comes to ECGs, a 100 Hz or 200 Hz sampling rate is mainly used. Considering the table as a whole, the disparity between the sampling rate used can be observed since there are solutions that use 1Hz [7] and other solutions that use signals at 200 Hz [22,24,26].

Regarding the preprocessing task, it is observed that most of the tasks are based on the elimination of artefacts, noise or the standardization of the signal. Most of the exposed solutions are based on the classification by events or windows. The signal is divided into windows of a specific duration where the apnea events are found. The duration of the windows is also highly controversial, and, as seen in Table 1, their duration ranges from 10 seconds to 60 seconds, passing through 30 seconds. The duration of the windows implies one of the most crucial aspects when it comes to the classification of freediving events. This is due to the fact that a concise window may not include the entire apnea event and therefore lose information. The same happens if the window is very long since windows of a duration greater than 60 can include more than one apnea event and lead to an underestimation of the total number of events, since counting the windows as apnea events, despite being able to find more than one event in the window, these events would be counted as one. This would significantly affect the AHI calculation, being of of the significant outputs.

If we look at Table 2, CNN models predominate over RNN. The results are generally very good for OSA event classification tasks, with accuracies ranging from 80% to 100% in most cases. When looking at the results, it is essential to differentiate between different metrics since one of the main problems when it comes to binary classification occurs here: balanced datasets. If we look exclusively at accuracy, a good result may not be representative if the dataset is unbalanced. In this case, numerous metrics, such as sensitivity and specificity, are displayed in all the proposed solutions, which give a much more accurate representation of the model's performance. The balancing of the dataset is something that must be covered prior to training the models. There are also techniques to reduce their influence in the event that it is not possible to use a balanced dataset.

Based on the results, it could be stated that all the models presented in Tables 1 and 2 generate good results, regardless of the signals used, sampling rate or preprocessing.

Table 2.Models and results obtained from scientific papers after reviewing the literature. CNN: convolutional neural network. LSTM: Long Short-Term Memory.

Publication	Model	Metrics – OSA Detection
Chang H. et al. 2020 [17]	1D-CNN	*Per-minute apnea detection:* Sen: 81.1% Spe: 92% Acc: 87.9% AUC: 0.94 *Per-recording classification:* Sen: 95.7% Spe: 100% Acc: 97.1% Corr.: 0.865
Dey D. et al. 2017 [16]	CNN	Sen: 97.8% Spe: 99.2% Acc: 98.9%
Drzazga J. et al. 2021 [15]	LSTM	*Acc. for 3 classes:* *No event, hypopnea, apnea* 86.42% 58.28% 69.50%
Haidar R. et al. 2018 [18]	CNN	Acc: 83.5%
Kristiansen S. et al. 2021 [7]	Bi-directional LSTM With Attention (BIWALSTM)	Acc: 89.4% Cohen's kappa: 0.79
Kwon H. et al. 2022 [19]	Hybrid CNN-LSTM	*Apnoe + hypopnoe and not apnea:* Cohen's kappa: 0.728 Sen: 78.1% Spe: 95.6% Acc: 93% AHI Pearson's coefficient: 0.97
Zhang J. et al. 2021 [20]	1D-CNN	Sen: 96.1 % Spe: 96.2 %
Pathinarupothi R. et al. 2017 [21]	LSTM-RNN	Acc: 92.1 % Precision: 99.5% Recall: 84.7% AUC: 0.98
Urtnasan E. et al. 2018 [22]	CNN	Acc: 96% Sen: 96% Spe: 96%
Nikkonen S. et al. 2021 [23]	LSTM	Cohen's kappa: 0.728
Zhang H. et al. 2022 [24]	U-NET	Acc: 94% F1: 63% Sen: 55% Spe: 98%
Van Steenkiste T. et al. 2019 [25]	LSTM	*Abdores* Acc: 77.2 % Sen: 62.3% Spe: 80.3% Precision: 39.9 *Thorres* Acc: 75 % Sen: 67.8% Spe: 76.5% Precision: 37.7 *Abdores* Acc: 60.1 % Sen: 52.1% Spe: 61.8% Precision: 22.1
Urtnasan E. et al 2020 [26]	CNN	Acc: 99% F1: 98%

4. Conclusions and outlook

A literature review has been carried out to study those requirements that are essential prior to training deep learning models for OSA detection. The database used, the set of signals, the sampling rate or the processing tasks are as crucial in the development of the models as the selection of the architecture. However, there is still no standard methodology for this stage of development.

Several solutions use different signal sets or models and obtain promising results in terms of apnea event classification. However, there is no evidence of why some methodologies work better than others. Numerous scientific papers that have developed automated OSA event detection solutions have been presented in this paper.

It is important to note that not everything is based on the models' performance, but other aspects are also relevant, such as providing a tool that is useful for doctors so that they can clearly study the models' decisions.

In conclusion, although there are scientific papers that agree on the set of signals to be used or the model used, there are usually variations in the preprocessing tasks or in the database used, which can significantly affect the performance of the model. This fact should be studied in the future with a broader literature review. Aspects such as AHI calculation and model explainability should also be addressed.

Acknowledgments

This research was partially funded by the German Federal Ministry for Economic Affairs and Energy, ZiM project "Sleep Lab at Home" (SLaH) grant: ZF4825301AW9 and Carl Zeiss Foundation: Project "Non-invasive system for measuring parameters relevant to sleep quality" (project number: P2019-03-003).

References

[1] S. Biswal *et al.*, "Expert-level sleep scoring with deep neural networks," *Journal of the American Medical Informatics Association*, vol. 25, no. 12, pp. 1643–1650, Dec. 2018, doi: 10.1093/jamia/ocy131.

[2] S. H. Choi *et al.*, "Real-time apnea-hypopnea event detection during sleep by convolutional neural networks," *Comput Biol Med*, vol. 100, pp. 123–131, Sep. 2018, doi: 10.1016/J.COMPBIOMED.2018.06.028.

[3] D. B. Kirsch, "PRO: sliding into home: portable sleep testing is effective for diagnosis of obstructive sleep apnea," *J Clin Sleep Med*, vol. 9, no. 1, pp. 5–7, 2013, doi: 10.5664/JCSM.2324.

[4] L. Cen, Z. L. Yu, T. Kluge, and W. Ser, "Automatic System for Obstructive Sleep Apnea Events Detection Using Convolutional Neural Network," *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2018, pp. 3975–3978, Oct. 2018, doi: 10.1109/EMBC.2018.8513363.

[5] D. Kim, J. Lee, Y. Woo, J. Jeong, C. Kim, and D. K. Kim, "Deep Learning Application to Clinical Decision Support System in Sleep Stage Classification," *J Pers Med*, vol. 12, no. 2, Feb. 2022, doi: 10.3390/JPM12020136.

[6] V. Thorey, A. B. Hernandez, P. J. Arnal, and E. H. During, "AI vs Humans for the diagnosis of sleep apnea," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 1596–1600, Jun. 2019, doi: 10.1109/EMBC.2019.8856877.

[7] S. Kristiansen *et al.*, "Machine Learning for Sleep Apnea Detection with Unattended Sleep Monitoring at Home," *ACM Trans Comput Healthc*, vol. 2, no. 2, pp. 1–25, Feb. 2021, doi: 10.1145/3433987.

[8] A. Manoni *et al.*, "A new wearable system for home sleep apnea testing, screening, and classification," *Sensors (Switzerland)*, vol. 20, no. 24, pp. 1–26, Dec. 2020, doi: 10.3390/s20247014.

[9] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, "Deep learning for time series classification: a review," *Data Min Knowl Discov*, vol. 33, no. 4, pp. 917–963, 2019, doi: 10.1007/s10618-019-00619-1.

[10] A. Ramachandran and A. Karuppiah, "A Survey on Recent Advances in Machine Learning Based Sleep Apnea Detection Systems," *Healthcare 2021, Vol. 9, Page 914*, vol. 9, no. 7, p. 914, Jul. 2021, doi: 10.3390/HEALTHCARE9070914.

[11] M. Piorecky *et al.*, "Apnea Detection in Polysomnographic Recordings Using Machine Learning Techniques," *Diagnostics (Basel)*, vol. 11, no. 12, Dec. 2021, doi: 10.3390/DIAGNOSTICS11122302.

[12] R. Zemouri, N. Zerhouni, and D. Racoceanu, "Deep Learning in the Biomedical Applications: Recent and Future Status," *Applied Sciences 2019, Vol. 9, Page 1526*, vol. 9, no. 8, p. 1526, Apr. 2019, doi: 10.3390/APP9081526.

[13] X. Qian *et al.*, "A review of methods for sleep arousal detection using polysomnographic signals," *Brain Sci*, vol. 11, no. 10, 2021, doi: 10.3390/brainsci11101274.

[14] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2017-May, pp. 1578–1585, 2017, doi: 10.1109/IJCNN.2017.7966039.

[15] J. Drzazga and B. Cyganek, "An lstm network for apnea and hypopnea episodes detection in respiratory signals," *Sensors*, vol. 21, no. 17, 2021, doi: 10.3390/s21175858.

[16] D. Dey, S. Chaudhuri, and S. Munshi, "Obstructive sleep apnoea detection using convolutional neural network based deep learning framework," *Biomed Eng Lett*, vol. 8, no. 1, pp. 95–100, 2018, doi: 10.1007/s13534-017-0055-y.

[17] H. Y. Chang, C. Y. Yeh, C. Te Lee, and C. C. Lin, "A sleep apnea detection system based on a one-dimensional deep convolution neural network model using single-lead electrocardiogram," *Sensors (Switzerland)*, vol. 20, no. 15, pp. 1–15, 2020, doi: 10.3390/s20154157.

[18] R. Haidar, S. McCloskey, I. Koprinska, and B. Jeffries, "Convolutional Neural Networks on Multiple Respiratory Channels to Detect Hypopnea and Obstructive Apnea Events," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2018-July, Oct. 2018, doi: 10.1109/IJCNN.2018.8489248.

[19] H. Bin Kwon *et al.*, "Hybrid CNN-LSTM Network for Real-Time Apnea-Hypopnea Event Detection Based on IR-UWB Radar," *IEEE Access*, vol. 10, pp. 17556–17564, 2022, doi: 10.1109/ACCESS.2021.3081747.

[20] J. Zhang *et al.*, "Automatic detection of obstructive sleep apnea events using a deep CNN-LSTM model," *Comput Intell Neurosci*, vol. 2021, 2021, doi: 10.1155/2021/5594733.

[21] R. K. Pathinarupothi, J. Dhara Prathap, E. S. Rangan, A. E. Gopalakrishnan, R. Vinaykumar, and K. P. Soman, "Single Sensor Techniques for Sleep Apnea Diagnosis Using Deep Learning," *Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017*, pp. 524–529, 2017, doi: 10.1109/ICHI.2017.37.

[22] E. Urtnasan, J. U. Park, E. Y. Joo, and K. J. Lee, "Automated Detection of Obstructive Sleep Apnea Events from a Single-Lead Electrocardiogram Using a Convolutional Neural Network," *J Med Syst*, vol. 42, no. 6, 2018, doi: 10.1007/s10916-018-0963-0.

[23] S. Nikkonen *et al.*, "Automatic Respiratory Event Scoring in Obstructive Sleep Apnea Using a Long Short-Term Memory Neural Network," *IEEE J Biomed Health Inform*, vol. 25, no. 8, pp. 2917–2927, Aug. 2021, doi: 10.1109/JBHI.2021.3064694.

[24] H. Zhang, X. Wang, H. Li, S. Mehendale, and Y. Guan, "Auto-annotating sleep stages based on polysomnographic data," *Patterns*, vol. 3, no. 1, p. 100371, 2022, doi: 10.1016/j.patter.2021.100371.

[25] T. Van Steenkiste, W. Groenendaal, Di. Deschrijver, and T. Dhaene, "Automated Sleep Apnea Detection in Raw Respiratory Signals Using Long Short-Term Memory Neural Networks," *IEEE J Biomed Health Inform*, vol. 23, no. 6, pp. 2354–2364, 2019, doi: 10.1109/JBHI.2018.2886064.

[26] E. Urtnasan, J. U. Park, E. Y. Joo, and K. J. Lee, "Identification of Sleep Apnea Severity Based on Deep Learning from a Short-term Normal ECG," *J Korean Med Sci*, vol. 35, no. 47, pp. 1–11, 2020, doi: 10.3346/jkms.2020.35.e399.