Sleep apnea events classification from a dual accelerometry system using deep learning models

Hugo Lafaye de Micheaux, Julie Fontecave-Jallon, Aurélien Bricout-Serrurier and Pierre-Yves Guméry
Univ. Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC
38000 Grenoble, France
hugo.lafaye-de-micheaux@univ-grenoble-alpes.fr and julie.fontecave@univ-grenoble-alpes.fr

***Abstract*—Sleep apnea syndrome (SAS) is a very common chronic disease characterized by the repetition of abnormal and frequent respiratory events, including central and obstructive apneas or hypopneas. Facing this public health problem, the challenge is twofold: improving early detection with minimally intrusive devices, and helping the diagnosis with an automatic scoring of recordings. In this context, we propose to exploit a dual thoracic and abdominal accelerometry system combined with electrocardiography (ECG) to classify sleep apnea events. Several common deep learning architectures were applied and performances were compared against several configurations of signals including nasal airflow from reference polysomnography (PSG), and according to manual expert annotations from PSG recordings. A Gated Recurrent Unit network was found to be the most efficient model and the configuration including ECG and abdominal and thoracic respiratory efforts from accelerometers as input channels provided very promising classification performances of normal and abnormal respiratory events. Thus, the simple technological proposition of dual accelerometry offers the possibility of an automatic identification of SAS events, very close to the expert annotations established from PSG multi-sensors.**

***Index Terms*—classification, accelerometry, deep learning, sleep apnea syndrome**

I. Introduction

Sleep Apnea Syndrome (SAS) is one of the most common chronic diseases, with one billion affected people worldwide [1]. It is a sleep-associated pathology that manifests itself by the repetition of abnormal, frequent and abnormally distributed respiratory events. These events may correspond to total (apnea) or partial (hypopnea) obstruction of the upper airways and are characterized by episodes of cessation or significant reduction of respiratory airflow for at least 10 seconds [2]. The events' origin is either obstructive or central. Obstructive events are due to pharyngeal collapse, and in that case, thoracic and abdominal movements are preserved. Central apneas or hypopneas are characterized by reduced control of the respiratory centers and are associated with the absence of thoracic and abdominal movements [3].

Polysomnography (PSG) is today the standard method for the diagnosis of sleep disorders, especially SAS. It consists in the simultaneous recording of several signals including electroencephalography, electrocardiography (ECG), airflow measured by a nasal cannula, oxygen saturation, and respiratory movements measured by respiratory inductance plethysmography. PSG recordings are then manually annotated in order to establish the diagnosis of SAS, based on the apnea-hypopnea index (AHI), which corresponds to the number of abnormal respiratory events per hour of sleep and evaluates the severity of the syndrome [4]. This human expertise, based on the analysis of all PSG signals, also aims to identify the type and origin of each event, in order to personalize the therapeutic solution.

This manual scoring of PSG recordings by expert, coupled with the cumbersome acquisition itself, makes the SAS diagnosing process complex, long and costly and is one of the explanations of the current underdiagnosis of this syndrome. SAS is therefore a real public health problem, including the challenge of early detection with minimally intrusive devices, and also the challenge of a diagnostic aid with an automatic scoring of recordings. Considering this last challenge, several approaches of machine learning or deep learning have been proposed these last years for automatic detection of SAS [5]–[9]. They are mainly focused on AHI estimation and only a few considers the automatic classification of events according to their types and origins [10]–[12]. Considering the technological challenge of reducing the complexity of PSG sensors, accelerometry has already been investigated for sleep monitoring [13]–[15]. We recently proposed a solution for night monitoring based on the use of 2 accelerometers placed on the subject's chest, from which thoracic and abdominal efforts are measured and lead to an estimation of the respiratory airflow. This estimation, called ADR for "Accelerometry-Derived Respiration" was first validated during sleep without abnormal events [16], and then, in terms of AHI estimation, for sleep-apnea screening [9].

The present study intends now to further exploit this double-accelerometry proposition and to evaluate its ability to differentiate the various types of respiratory events in SAS. For that purpose, several deep learning approaches are considered and evaluated on different combinations of signals including thoracic and/or abdominal respiratory movements from accelerometry, with or without ECG, and also nasal cannula airflow as reference.

II. Materials and Methods

We first describe our proposed methodology, starting with the data considered, followed by the description of the four different deep learning architectures used to classify our data samples and finishing by the training strategy.

A. Data

a) Signals of interest: The sleep study protocol included 28 volunteers and untreated SAS patients at the sleep laboratory of Grenoble University Hospital for an overnight recording. Subjects were equipped with classical polysomnography sensors and a dual accelerometry system. They provided written information consent and the study was approved by the relevant ethics committee (CHU Grenoble Alpes). From the recorded data and in this particular study, we were only interested in 4 signals: the abdominal and thoracic efforts reconstruction calculated from accelerometer data [16], the reference nasal airflow measured by nasal cannula and the cardiac activity measured by electrocardiogram, respectively noted {ABD, THO, CAN, ECG} and all acquired at 256 Hz.

Besides, all normal and abnormal respiratory events were manually annotated by a unique sleep expert, according to all PSG signals, and not only the 4 signals of interest. This expert analysis leads to 5 annotated classes of events, associated with their time locations: normal respiration, obstructive and central sleep apneas, obstructive and central sleep hypopneas, respectively noted {resp, apnobs, apncen, hypobs, hypcen}. An illustration of signals with two examples of abnormal events is given in Fig. 1. More details about the data acquisition, formatting, labelling and post-acquisition processing are given in previous studies [9], [16].

b) Preprocessing: A preprocessing step was applied to each signal of interest {ABD, THO, CAN, ECG} in order to make them suitable for deep learning processes. It consists of:

1. Segmenting the overnight recordings into 12-seconds samples with a 11-seconds overlap maximizing the number of samples (the shortest possible event being of 10 s [2]),
2. Resampling each sample at 128 Hz (being a good compromise between reducing the amount of data and still correctly assess ECG signals),
3. Normalizing each sample by z-score based on 5-minutes sliding window means μ and standard deviations σ (5 min being the reference time for sleep scorers): $z = \frac{x-\mu}{\sigma}$
4. Separating each class in separate datasets,
5. Shuffling each dataset,
6. Removing some events of over-represented classes to reduce data imbalance (indeed *e.g.* respiration events outnumber central apnea ones by a factor 400),
7. Splitting data into training/validation/testing sets with a stratified 80/10/10 distribution, such that each class is evenly distributed in each set.

In order to have reproducible results and compare models, shuffling and splitting were set to be rigorously the same across experiments. This was done by fixing the random seed, responsible for initializing the random number generator.

B. Deep Learning Architectures

a) Multilayer perceptron (MLP): Firstly, we considered an MLP [17] as it is the most classical type of neural network, often used as a standard baseline for its simplicity. Our MLP model is composed of an input flatten layer, followed by 3 hidden dense layers of 32 units with relu activations and batch normalizations, a dropout layer of rate 0.2 and an output dense layer of 5 units with a softmax activation.

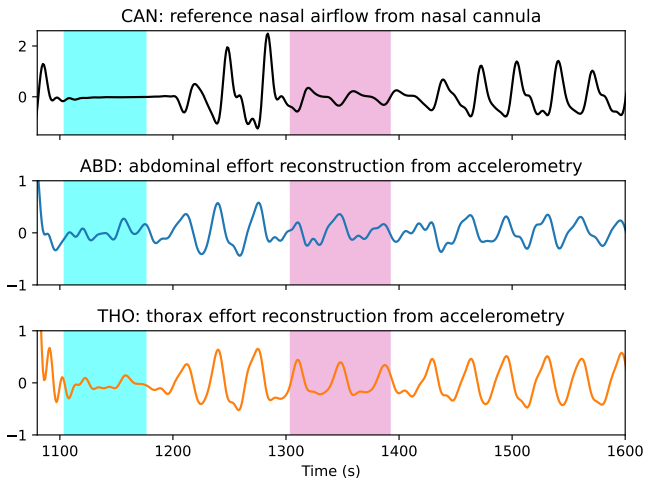

Fig. 1. An illustration of our signals of interest with two abnormal events highlighted, one central apnea (cyan) and one obstructive hypopnea (pink), according to the expert annotations.

b) Convolutional neural network (CNN): Secondly, we considered a CNN [18] in its one-dimensional form. CNNs usually involve the spatial study of two-dimensional images via convolutions but they are also relevant for time series study as time can be seen as a spatial dimension. Our CNN model is composed of 3 1D-convolution layers of 32 filters of size 3 with relu activations and batch normalizations, followed by a flatten layer, a dense layer of size 32 with relu activation and a batch normalization, a dropout layer of rate 0.2 and an output dense layer of 5 units with a softmax activation.

c) Long short-term memory (LSTM): Thirdly, we considered a LSTM network [19], a famous recurrent neural network (RNN) designed to capture temporal dependencies in time series by using an internal memory. Our LSTM model is composed of a LSTM layer of 256 units, followed by an output dense layer of 5 units with a softmax activation.

d) Gated recurrent unit (GRU): Fourthly, we considered a GRU network [20], another RNN simpler than the LSTM network with less memory consumption and also a bit faster. Our GRU model is composed of a GRU layer of 256 units, followed by a dropout layer of rate 0.2 and an output dense layer of 5 units with a softmax activation.

C. Training Strategy

Being in a multi-class classification, we used the categorical cross-entropy (CCE) as loss function. CCE is defined as:

$$\text{CCE}(y, \hat{y}) = -\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) \tag{1}$$

where N is the number of classes (5 in our case), y_i is the true value (0 or 1) indicating whether class i is the correct classification for the observation y, and $\hat{y}_i$ is the probability

value, predicted by a model, of the observation y being in class i.

For training, the Adam algorithm was selected as the optimizer and a batch size of 128 was used. All networks were trained with a learning rate of 10^{-3} on 150 epochs until convergence. The implementation was done in Keras and massive parallelization was exploited via GPU computation.

III. Results and Discussion

In this section, we are first interested in the computational performances of the deep learning models and their comparison through classification metrics. Then, with the exploitation of the best model according to our application context, we compare several configurations of input signals, later called (input) channels.

A. Parameters and Training Time Evaluation

We evaluated the time duration of the training phase on a PC with an Intel Xeon(R) Silver 4214R CPU performing at 2.40GHz x 48 that exploited GPU computations on an Nvidia Quadro RTX 6000. In addition, we also compared the number of parameters per model with different numbers of input channels (more details about the performance results of each configuration are given in III-C).

Tab. I summarises the results. As we can see, changing the number of input channels has a big influence on the number of parameters in the MLP model but it does not significantly influence the other models. Indeed, an MLP architecture flattens all the input channels and then connects each point to each neuron of the first layer, multiplying therefore the number of weights of the first layer. Whereas the three other architectures treat multiple input channels in parallel, which does not really impact the size of the network. About CNN, the high number of parameters is due at 99% to the dense layer preceded by the flatten layer. To reduce it, pooling layers or less convolution filters could be used, but in our case it decreases a bit the performances without being significantly faster.

Looking at the training times, we can see that the MLP and CNN models are much faster than the LSTM and GRU models. This is a normal behaviour since processing MLP and CNN models are well adapted to GPU computations, while it is not the case for RNN models like LSTM and GRU as they need to deal with temporal memory, which is hardly parallelizable. However, about 70 min to train a model is still an acceptable time as it is done only once. Moreover, it is worth noting that a forward pass to classify the test set was less than a minute for all models.

B. Model Comparison

During the test phase, for a given input, a probability is given for each class at the output of the models. The highest probability is chosen to be the predicted class. To compare the performances of our models, we considered classification metrics such as the balanced accuracy for its ability to deal with imbalanced data and the f1-score as it is the traditional way to measure classification's accuracy. They are defined according to numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as:

$$\text{balanced-accuracy} = \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right) \Big/ 2 \quad (2)$$

$$\text{f1-score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (3)$$

Tab. II gives the scores of balanced accuracy and f1-score of the four models of classification into 5 classes. Here only one configuration of inputs with the three channels ABD, THO and ECG was chosen to compare the models.

In our case study and in terms of classification performances, we have the following hierarchy of models: GRU>LSTM>CNN>MLP. With higher performances, the RNN models better capture the temporal features of our channels allowing a good classification. Even if CNN is capable of considering spatial information as temporal information, it is not sufficient and less adapted than RNN here. Outperforming the other models, GRU will therefore be chosen for deeper analyses in the next section.

C. Quantitative Evaluation

Now, we want to study which input channels are needed to well classify the samples into the right class over the 5 {resp, apnobs, apncen, hypobs, hypcen}. In this purpose, we tried different configurations of input channels such as a 1-channel input with ABD or THO samples, a 2-channels input with a combination of ABD-THO or CAN-ECG samples, and a 3-channels input with ABD-THO-ECG samples.

Fig. 2 shows the confusion matrices of these configurations of input channels. A confusion matrix allows the visualization of the performance of an algorithm with a classification problem. It represents, here in percentage, the ratio of correct predictions (called "predicted label" on the figures) compared to the reality (called "true label") for each of the 5 classes.

TABLE I
Number of parameters and training time for the 4 classification models according to the number of input channels

Model	1-channel	2-channels	3-channels	Training time*
MLP	58 K	113 K	169 K	5 min
CNN	1574 K	1574 K	1574 K	11 min
LSTM	265 K	267 K	268 K	71 min
GRU	200 K	201 K	202 K	67 min

*Training times are almost the same whatever the number of channels.

TABLE II
Metric scores of the 4 models of classification in 5 classes with 3 input channels (ABD, THO, ECG)

Model	balanced-accuracy	f1-score
MLP	0.527	0.508
CNN	0.594	0.583
LSTM	0.827	0.819
GRU	**0.920**	**0.916**

*Best results are highlighted in bold.

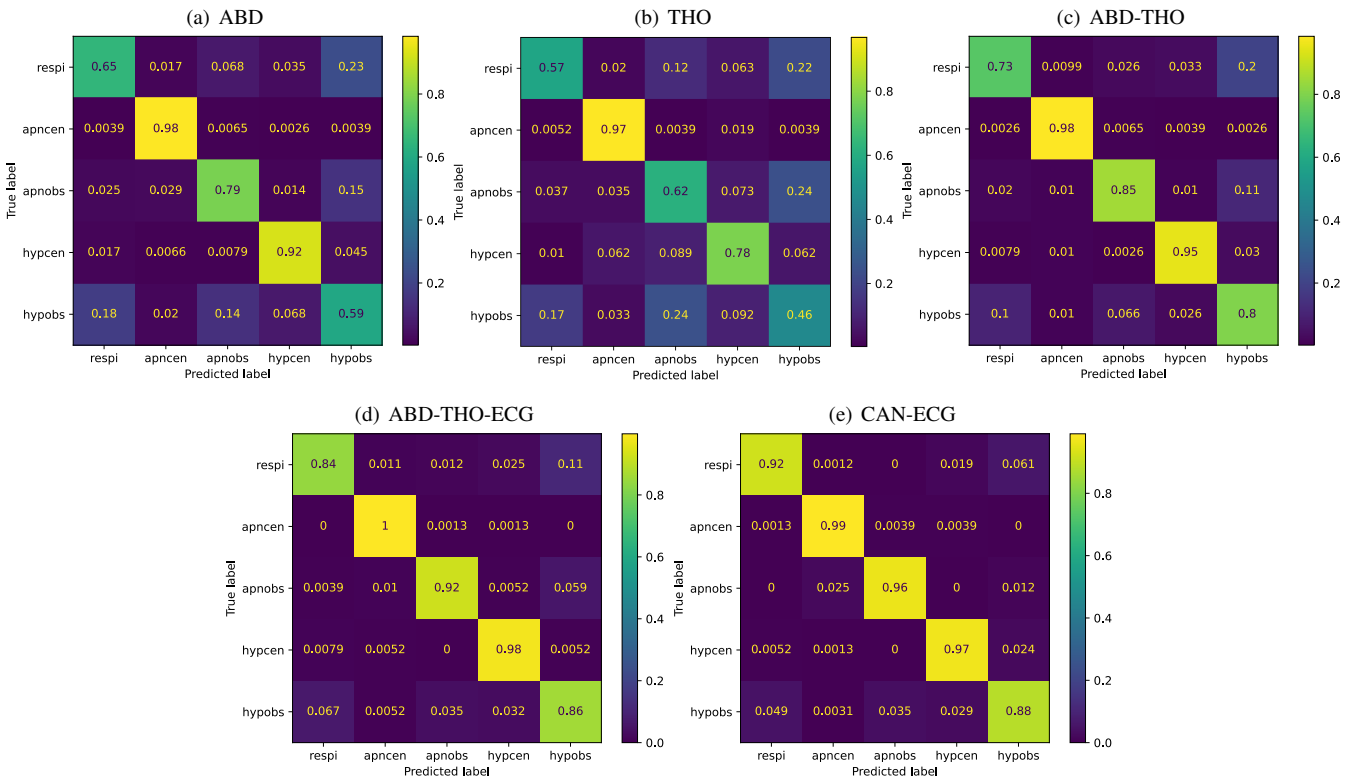

Fig. 2. Confusion matrices to classify into the 5 classes {resp, apnobs, apncen, hypobs, hypcen} with different configurations of input channels: (a) only abdominal ABD signal, (b) only thoracic THO signal, (c) ABD-THO signals, (d) ABD-THO-ECG signals, (e) reference nasal airflow CAN and ECG signals. *On the figures, the word "label" is similar to "class".

By analysing the confusion matrices, several results can be discussed.

Firstly and globally by looking at all confusion matrices from Fig. 2 (a) to (e), the events from a central origin are very well classified whatever the type and number of inputs. This behaviour may be explained by the specific signature of signals, caused by the disappearance of thoracic and abdominal movements during central events.

Secondly, comparing 1-channel ABD (Fig. 2 (a)) or THO (Fig. 2 (b)) versus 2-channels ABD-THO (Fig. 2 (c)), we can see the importance of using two accelerometers instead of only one. Indeed, using two accelerometers improves a lot the classification of respiration and obstructive events, especially the hypopnea ones. That can be explained by their complementarity and their phase opposition found in this kind of events. With only one channel, such a phase opposition can not be observed, which depreciated the classification performances.

Thirdly, when adding ECG as an input channel (Fig. 2 (d)), we can notice that it improves the classification results for all types of events compared to Fig. 2 (c). In fact, ECG brings information about the cardiac activity such as the heart rate variability which is known to be linked to the sleep apnea syndrome. With our observation, we can therefore confirm that cardiac information plays an important role in SAS helping to differentiate events, and that we should use ECG to classify them.

Finally, considering ABD-THO-ECG the set of channels to make our classification, we need to investigate if this dual accelerometry approach combined with ECG provides performance of events' classification in line with those of a reliable reference. A reference classification is therefore considered with the airflow measured by nasal cannula CAN (as it is currently the respiratory reference in PSG analysis) and the ECG for cardiac activity. The confusion matrix of this reference CAN-ECG is given by Fig. 2 (e). Thus, our choice to consider the channels ABD-THO-ECG as inputs for events' classification seems convincing since the performances are getting very close to the reference CAN-ECG. However, we can still see some confusions between respiration and obstructive hypopnea events. An obstructive hypopnea event is nothing but a shallow breathing, making it similar to a respiration event. This similarity can lead to confusions during classification. Nevertheless, such confusions happen only around 9% of the time for these concerned events.

IV. Conclusion

Our study aimed at evaluating the ability of a dual accelerometry system to correctly classify sleep apnea events. We compared the performance of four famous deep learning architectures MLP, CNN, LSTM and GRU in different configurations of input channels, including accelerometric signals,

electrocardiography and reference respiratory airflow. GRU was found to be the most efficient model of all and the configuration with ABD-THO-ECG signals as input channels had classification performances very close to the reference classification, including airflow from a nasal cannula and ECG. Thus, we can conclude that our proposition of using accelerometry for abnormal respiratory events' detection is promising, provided to consider both thoracic and abdominal accelerometers, combined with ECG. It allows thus, with a few simple sensors, an automatic identification of events, very close to the expert annotations established from PSG multi-sensors.

In future clinical use, our proposed method might be coupled with the AHI estimation for a SAS diagnosis of new subjects. In this context, the fusion of the classified events will have to be considered in order to reconstruct the events' location on an overnight recording. Moreover, it could be interesting to investigate a leave-one-out cross-validation strategy instead of a stratified one.

References

[1] A. Benjafield, N. Ayas, P. Eastwood, R. Heinzer, M. Ip, M. Morrell, C. Nunez, S. Patel, T. Penzel, J.-L. Pépin, P. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," *The Lancet Respiratory Medicine*, vol. 7, Jul. 2019. doi:10.1016/S2213-2600(19)30198-5

[2] M. R. Mannarino, F. Di Filippo, and M. Pirro, "Obstructive sleep apnea syndrome," *European Journal of Internal Medicine*, vol. 23, no. 7, pp. 586–593, Oct. 2012. doi:10.1016/j.ejim.2012.05.013

[3] A. de Groote, J. Groswasser, H. Bersini, P. Mathys, and A. Kahn, "Detection of obstructive apnea events in sleeping infants from thoracoabdominal movements," *Journal of Sleep Research*, vol. 11, no. 2, pp. 161–168, Jun. 2002. doi:10.1046/j.1365-2869.2002.00291.x

[4] M. Billard and Y. Dauvilliers, *Les troubles du sommeil*. Elsevier Masson, 2005.

[5] C. Varon, A. Caicedo, D. Testelmans, B. Buyse, and S. Van Huffel, "A novel algorithm for the automatic detection of sleep apnea from single-lead ecg," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 9, pp. 2269–2278, 2015. doi:10.1109/TBME.2015.2422378

[6] K. Li, W. Pan, Y. Li, Q. Jiang, and G. Liu, "A method to detect sleep apnea based on deep neural network and hidden markov model using single-lead ECG signal," *Neurocomputing*, vol. 294, pp. 94–101, Jun. 2018. doi:10.1016/j.neucom.2018.03.011

[7] S. S. Mostafa, F. Mendonça, F. Morgado-Dias, and A. Ravelo-García, "Spo2 based sleep apnea detection using deep learning," in *2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)*, 2017, pp. 000 091–000 096. doi:10.1109/INES.2017.8118534

[8] S. Nikkonen, I. O. Afara, T. Leppänen, and J. Töyräs, "Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea," *Scientific Reports*, vol. 9, no. 1, Sep. 2019. doi:10.1038/s41598-019-49330-7

[9] A. Bricout, J. Fontecave-Jallon, J.-L. Pépin, and P.-Y. Guméry, "Accelerometry-derived respiratory index estimating apnea-hypopnea index for sleep apnea screening," *Computer Methods and Programs in Biomedicine*, vol. 207, p. 106209, Aug. 2021. doi:10.1016/j.cmpb.2021.106209

[10] M. Alimardani and G. de Moor, "Automatic classification of sleep apnea type and severity using EEG signals," in *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 2021, pp. 121–128. doi:10.5220/0010288301210128

[11] H. Yue, Y. Lin, Y. Wu, Y. Wang, Y. Li, X. Guo, Y. Huang, W. Wen, G. Zhao, X. Pang, and W. Lei, "Deep learning for diagnosis and classification of obstructive sleep apnea: A nasal airflow-based multi-resolution residual network," *Nature and Science of Sleep*, vol. 13, pp. 361–373, Mar. 2021. doi:10.2147/NSS.S297856

[12] A. Chatterjee and N. D. Jana, "Classification of sleep apnea event type using imbalanced labelled EEG signal," in *2022 IEEE Region 10 Symposium (TENSYMP)*, IEEE. IEEE, Jul. 2022, pp. 1–6. doi:10.1109/tensymp54529.2022.9864566

[13] C. L. Bucklin, M. Das, and S. L. Luo, "An inexpensive accelerometer-based sleep-apnea screening technique," in *Proceedings of the IEEE 2010 National Aerospace and Electronics Conference*, 2010, pp. 396–399. doi:10.1109/NAECON.2010.5712984

[14] K. T. Sweeney, E. Mitchell, J. Gaughran, T. Kane, R. Costello, S. Coyle, N. E. O'Connor, and D. Diamond, "Identification of sleep apnea events using discrete wavelet transform of respiration, ecg and accelerometer signals," in *2013 IEEE International Conference on Body Sensor Networks*, 2013, pp. 1–6. doi:10.1109/BSN.2013.6575488

[15] N. Selvaraj and R. Narasimhan, "Automated prediction of the apnea-hypopnea index using a wireless patch sensor," *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, vol. 2014, pp. 1897–900, Aug. 2014. doi:10.1109/EMBC.2014.6943981

[16] A. Bricout, J. Fontecave-Jallon, D. Colas, G. Gerard, J.-L. Pépin, and P.-Y. Guméry, "Adaptive accelerometry derived respiration: comparison with respiratory inductance plethysmography during sleep," in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Jul. 2019, pp. 6714–6717. doi:10.1109/embc.2019.8856561

[17] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989. doi:10.1016/0893-6080(89)90020-8

[18] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989. doi:10.1162/neco.1989.1.4.541

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. doi:10.1162/neco.1997.9.8.1735

[20] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014. doi:10.48550/arxiv.1412.3555