Sleep Apnea Detection - Towards Wearables

Martin Kralik, Andrea Nemcova, Jiri Kozumplik, Eniko Vargova

Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering, Brno, Czechia

Abstract

Sleep apnea syndrome (SAS) is a common but serious sleep disorder affecting millions worldwide. It is marked by pauses in breathing over 10 seconds (apnea) or shallow breathing (hypopnea) during sleep, disrupting the sleep cycle and causing fatigue. Untreated SAS increases the risk of stroke, heart disease, high blood pressure, and mental health issues like depression and anxiety. Despite being treatable, many cases remain undiagnosed due to symptoms being misattributed or the inaccessibility of polysomnography (PSG), the diagnostic gold standard.

Recently, more accessible methods like electrocardiography or photoplethysmography (PPG) have shown promise, especially with wearable devices. Our work explores using single-channel PPG and deep learning methods (AlexNet, custom CNN, ZF-net) to classify SAS using data from the Multi-Ethnic Study of Atherosclerosis. Using all-night data from 120 patients with the highest apnea-hypopnea index, we trained on around 9,000 non-apneic and 9,000 SAS-positive intervals, achieving 87.7% accuracy and 85.1% sensitivity, showing the potential for early diagnosis without PSG.

1. Introduction

Sleep apnea syndrome (SAS) or sleep apnea in short is a sleep disorder, falling into category of sleep-disordered breathing. It is a condition characterized by decrease or disappearance of respiratory airflow during sleep, caused by partial or full airway obstruction or by lack of respiratory effort [1]. It is a disease that can be called common, yet it is unknown exactly how large of a population is affected by SAS. Some sources estimate prevalence of SAS in general population from 3% to 4% [2], while others estimate up to 17% of population is affected by moderate SAS [3]. Elderly population seems to be more prone to SAS, with other risk factors being obesity, gender, as well as smoking, alcohol consumption and usage of sedating meditation [4]. The results of untreated SAS does not only affect quality of daily life in the meaning of sleepiness, but can cause other problems, such as higher blood pressure or acute ischemic stroke [5].

SAS can be divided into three main categories: obstructive apnea (OA), caused by total blockage of airways, characterized by the presence of respiratory effort, central apnea (CA), caused by the lack of respiratory effort and therefore lack of airflow and hypopnea, caused by partial blockage of airways with respiratory effort present and airflow decreased by at least 10% comparing to normal. The severeness of disease is measured by apnea-hypopnea index (AHI), frequency of apneas and hypopneas per hour of sleep, with AHI = 10-15 being usually accepted as a threshold of diagnosed SAS [1].

The main problems of successful SAS diagnosis are two. First, the symptoms come during sleep and thus are unnoticed in many cases. Second problem is the need of all night polysomnography (PSG) to evaluate the severity and the cause of the disease. This can be problematic, as the sleep laboratory can be unavailable for many and the examination itself requires multiple diagnostic devices, being expensive and not to mention time consuming for both the patient and the expert, who evaluates the data. This inspired many authors in recent years, who are trying to find alternative approach for SAS detection, mainly focusing on the ECG signal. Wide spread of wearable devices however leads to another approach - utilizing photoplethysmographic (PPG) records from these devices to detect SAS, since respiratory signal can be estimated from both ECG and PPG [6].

While we do not think that wearable devices will replace PSG as a diagnostic tool for SAS or any sleep related disorder, they could serve as a warning tool or screening tool, giving the user ”raised finger” that examined sleep respiration might not be alright and that the consultation with a medical expert might be a solution. Modern devices come equipped with hardware powerful enough to evaluate 1D signals using artificial intelligence including deep networks. Measurement of a full night PPG and detection of SAS from the signal of such length using deep learning might be still problematic for battery life of a wearable device. Thus, our focus is on a per segment detection, since measurement of multiple one-minute segments per night might be doable and is in fact done by many devices to

evaluate sleep quality and heart rate changes. The goal of this paper is then to examine the possibility of utilization of deep convolutional neural networks (CNN) for binary SAS detection from one-minute segments of PPG signals without any prior filtration or feature extraction.

2. Methods

2.1. Data

For our work, The Multi-Ethnic Study of Atherosclerosis (MESA) [7,8] database was used as a source of data for our simplified dataset. This database was chosen mainly due to the large amount of all-night recordings with included PPG signals, which is suitable for deep learning applications. All-night recordings from approximately 120 most apneic patients were split into one-minute intervals, with multiple rules applied.

At first, intervals were split into groups called "normal" (healthy, without apnea), "central apnea", "obstructive apnea" and "hypopnea" according to the expert annotations. Then, the one-minute segments were adjusted so that the apnea starts randomly during the first 30 seconds of the segment to ensure that all or most of the apnoic break is present in the segment but it does not always start immediately at the start of the segment. Then segments with more than one type of apnea present were removed. Even though we simply detect any form of apnoic presence in the segment, we wanted to future-proof the dataset for potential following studies, in case we would like to apply classification of SAS types on the same data. The last pruning was done by visually checking the data to remove any form of high noise or flat line segments, possibly caused by malfunction or disconnection of the PPG device.

Due to the expected difficulty of hypopnea detection, we created two datasets - Dataset A, which contains only CA and OA segments in the Apnea group and Dataset B, which also includes hypopnea segments. The ration of different kinds of apnea was selected based on the distibution of CA, OA and hypopnea segments in the selected 120 patients. Total number of segments in each created dataset is summarized in the Table 1.

Table 1. Summary of number of segments in Dataset A and Dataset B

	Dataset A	Dataset B
Central apnea	528	528
Obstructive apnea	7,110	4,350
Hypopnea	—	4,649
Normal (healthy)	7,638	9,527

2.2. Apnea detection

To simulate possible wearable use, little to no preprocessing approach was chosen. Only preprocessing applied to the PPG signal was down-sampling from 256 Hz to 64 Hz to decrease computational requirements. Each dataset was split randomly into training and testing datasets using 80:20 ratio and signals were normalized.

As for the SAS detection itself, multiple known CNN architectures were used, such as AlexNet [9], ZF-Net [10] and VGG-16 [11]. All of those were originally designed as 2D CNNs, we adapted the architecture to work with 1-D inputs. Additionally, one custom model, that was designed and tuned to be as simple as possible, while still giving reasonable results, was used. Simplified architecture visualisation of the simple custom CNN and the most complex CNN are seen in Figure 1. Output of CNNs is binary classification into two classes - Apnea or Normal.

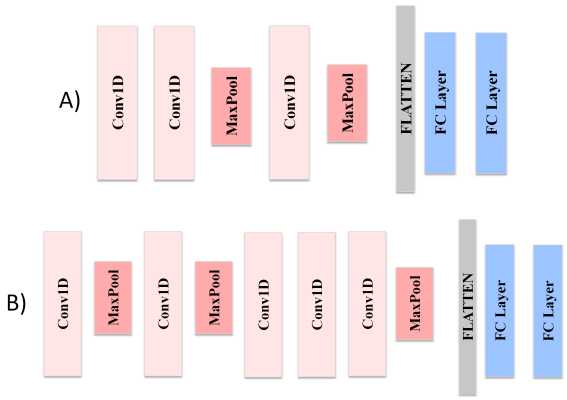

Figure 1. Examples of different CNN architectures: A) Custom CNN, B) ZF-Net

Multiple different hyperparameteres were applied during initial training phase. As a result, stochastic gradient descent (SGD) was selected as the best performing optimizer and binary cross-entropy (BCE) as a loss function. During the second half of the trainign we used smaller learning rate for fine tuning. Summary of final hyperparameters is in Table 2.

Table 2. List of training hyperparameters

Batch size	32
Epochs	10
Optimizer	SGD
Loss function	BCE
Learning rate	0.01 to 0.001

Table 3. Summary of results for both testing datasets

	Dataset A			Dataset B		
	Se (%)	Sp (%)	Acc (%)	Se (%)	Sp (%)	Acc (%)
Custom CNN	85.16	81.55	83.26	72.15	71.58	71.86
ZF-Net	85.12	**90.21**	**87.66**	**75.09**	**74.05**	**74.54**
Alex-Net	**86.97**	81.93	84.27	74.82	71.69	74.51
VGG-16	83.83	81.53	82.64	74.83	68.76	71.40

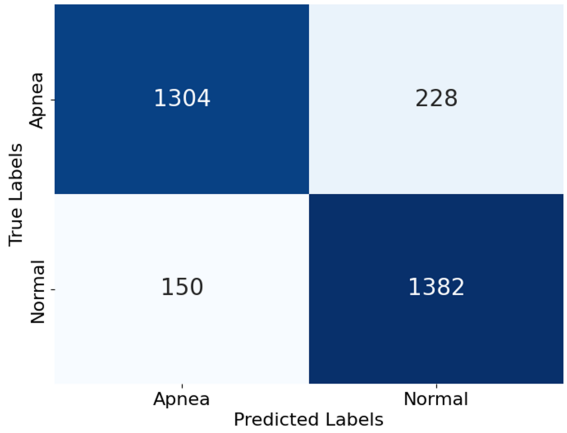

Figure 2. Confusion matrix of ZF-Net testing results on Dataset A

3. Results

Since both used datasets are balanced in the meaning of number of segments in both classes and the outputs of CNNs are binary, we used simple accuracy, sensitivity and specificity for evaluation of different CNN performance. Figure 2 shows confusion matrix of ZF-Net on Dataset A, as an example of best performing network on a dataset A. This shows high accuracy with balanced sensitivity and specificity over 85%. Detailed comparison of all used CNNs on both datasets can be seen in Table 3. This shows that all models give similar performance on respective datasets with ZF-Net giving the best results on both datasets.

What can be seen in Table 3 is that all models have decrease of all metrics by 10-15% on dataset B. This was expected, as nearly half of apnea segments in dataset B are hypopneas, which are much harder to detect even from respiratory signals and even experts do not always agree on hypopneas when annotating the PSG recordings. Downside of our approach is that with binary classification, we are unable to evaluate how much do hypopnea segments contribute to this accuracy decrease.

4. Discussion

When evaluating results, we can come into multiple interesting conclusions. At the first glance, we could say that PPG signals have potential for SAS detection. This could be a long run in the meaning of certified medical device with reliable results, however replacement of PSG as a tool for SAS diagnostics is not the aim of this article. What is shown is a potential for wearable devices to serve as a tool to increase awareness about SAS and to serve as a screening tool and preventative device, similarly how smart watches with ECG work in the area of atrial fibrillation detection [12].

Another conclusions can be done when comparing results on two datasets. As was said earlier, hypopneas can be harder to detect, since the range of respiration volume decrease can vary in each case. However, only 10-15% decrease of accuracy was observed for dataset B, even though hypopneas created almost half the the dataset apnea segments. While at the moment we cannot exactly evaluate the contribution of the hypopnea segments to that decrease, the overall results attacking 75% sensitivity for all models show potential for additional research.

Furthermore, modern wearable devices do not often come equipped with PPG sensor alone, but with SpO2 sensors as well. SpO2 signals are also usually measured during PSG, so they are not only part of MESA database, but almost all sleep databases publicly available. This opens another opportunity for researching this topic, as SpO2 decrease is one of the main features for classifying hypopneas.

When comparing performance between each model, we can see we achieved similar results in the meaning of sensitivity and accuracy with ZF-Net giving best results with 90% specificity, outperforming other models by far. It should be noted however that ZF-Net is the most complex network with almost 28 million trainable parameters, while our custom CNN is more simple and still giving similar sensitivity on dataset A and only 3% decrease of sensitivity on dataset B, this all with just 1 million trainable parameters. Further study could even decrease the complexity of this network while maintaining the performance, giving it the potential to be actually deployed on a wearable device.

When trying to compare our results with other authors, we discovered that there is no standard for evaluating alter-

native SAS detection algorithms. Overall when evaluating per recording detection, authors usually achieve better results, since they usually aim at the AHI exceeding certain threshold to call the patient positive with apnea. Our per segment approach on the other hand usually gives slightly worse numbers, nevertheless such approach might be better alternative, because all night recording of PPG could be problematic for wearables of today, however lets say multiple one-minute checks per hour could be probably done without battery concerns.

5. Conclusion

We trained and tested multiple CNN models to examine possibility of SAS detection from 1-minute segments of PPG signal, extracted from MESA database. We compared performance of selected models on data excluding and including hypopnea segments. The performance of all utilized models exceeded 83% accuracy for non-hypopnoic data and 71% accuracy for data with hypopneas included. Best results were achieved by ZF-Net model, with sensitivity 85.12%, specificity 90.21% and overall accuracy 87.66%. We were able to compare models with different complexity to show that even more simple CNNs can be used to detect SAS, making it possible to be implemented in wearable devices and to serve as a screening tool or first warning for millions worldwide.

Acknowledgments

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

References

[1] Flemons WW, Buysse D, Redline S, Oack A, Strohl K, Wheatley J, Young T, Douglas N, Levy P, Mcnicolas W, Fleetham J, White D, Schmidt-nowarra W, Carley D, Romaniuk J. Sleep-Related Breathing Disorders in Adults. Sleep New York NY 1999;22(5):667–689. ISSN 0161-8105.

[2] Acquavella J, Mehra R, Bron M, Suomi JM, Hess GP. Prevalence of Narcolepsy and Other Sleep Disorders and Frequency of Diagnostic Tests From 2013-2016 in Insured Patients Actively Seeking Care. Journal of Clinical Sleep Medicine 2020;16(8):1255–1263. ISSN 1550-9389.

[3] Senaratna CV, Perret JL, Lodge CJ, Lowe AJ, Campbell BE, Matheson MC, Hamilton GS, Dharmage SC. Prevalence of Obstructive Sleep Apnea in the General Population. Sleep Medicine Reviews 2017;34:70–81. ISSN 10870792.

[4] Neikrug AB, Ancoli-Israel S. Sleep Disorders in the Older Adult – A Mini-Review. Gerontology 2010-3-1;56(2):181–189. ISSN 0304-324X.

[5] Bassetti CL, Milanova M, Gugger M. Sleep-Disordered Breathing and Acute Ischemic Stroke: Diagnosis, Risk Factors, Treatment, Evolution, and Long-Term Clinical Outcome. Stroke 1970 2006;37(4):967–972. ISSN 0039-2499.

[6] Charlton PH, Birrenkott DA, Bonnici T, Pimentel MAF, Johnson AEW, Alastruey J, Tarassenko L, Watkinson PJ, Beale R, Clifton DA. Breathing Rate Estimation From the Electrocardiogram and Photoplethysmogram. IEEE Reviews in Biomedical Engineering 2018;11:2–20. ISSN 1937-3333.

[7] Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S. The National Sleep Research Resource. Journal of the American Medical Informatics Association 2018-10-01;25(10):1351–1358. ISSN 1067-5027.

[8] Chen X, Wang R, Zee P, Lutsey PL, Javaheri S, Alcántara C, Jackson CL, Williams MA, Redline S. Racial/Ethnic Differences in Sleep Disturbances. Sleep 2015;38(6):877–88. ISSN 0161-8105.

[9] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification With Deep Convolutional Neural Networks. Communications of the ACM 2017-05-24;60(6):84–90. ISSN 0001-0782. URL https://dl.acm.org/doi/10.1145/3065386.

[10] Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks, 2013. URL https://arxiv.org/abs/1311.2901.

[11] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015. URL https://arxiv.org/abs/1409.1556.

[12] Pepplinkhuizen S, Hoeksema WF, van der Stuijt W, van Steijn NJ, Winter MM, Wilde AA, Smeding L, Knops RE. Accuracy and Clinical Relevance of the Single-Lead Apple Watch Electrocardiogram to Identify Atrial Fibrillation. Cardiovascular Digital Health Journal 2022;3(6):S17–S22. ISSN 26666936.

Address for correspondence:

Martin Kralik

Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technicka 3082/12, 616 00 Brno, Czechia

kralikmartin@vut.cz