Bielefeld University

Title
of
Thesis

Julian Hendrik Freiherr Bock von Wülfingen

Master Thesis

in Intelligent Systems

AG Machine Learning

Primary Supervisor: Michiel Straat
Secondary Supervisor: Pedro Fonseca
Date: XX.XX.2025

Contents

Abstract

Abstract text

Chapter 1

Introduction

A recent study estimated that over 900 million adults globally are affected by the common group of respiratory sleep disorders called Sleep-disordered breathing (SDB) [1], or in particular, Sleep Apnea, which clinical manifestations include sleepiness, fatigue, cardiovascular disease, and hypertension. SDB is even linked to higher cases of diabetes, stroke occurrences and increased morbidity [2, 3, 4]. Diagnosis of the disorder relies on detecting repeated respiratory events in which airflow is either reduced (hypopnea) or entirely paused (apnea) during sleep [2, 5]. These events can further be categorized into obstructive or central origin, depending on if the apnea happens due to a physical blockage of the upper airway or if caused by the brain failing to signal breathing resulting in missing breathing effort. In case the event shows features of both, it is classified as a mixed.Dividing the number of events by the total sleep time (TST) gives the Apnea-Hypopnea-Index (AHI), which indicates the severity of the disorder.

Gold standard for detecting SDB in Polysomnography (PSG) which captures physical and biological signals like heart (electrocardiogram, ECG) and brain (electroencephalogram, EGG) activity, airflow, peripheral oxygen saturation (SpO2), chest and abdominal movements, sleeping position, and blood volume changes (photoplethysmographie, PPG). This approach comes with a few downsides: Firstly, due to the vast amount of sensors and specialized equipment, setup and analysis of the full PSG is costly, requires human experts and might impact sleep quality. Secondly, looking only at a single night might have low diagnostic meaningfulness [6] and the analysis of multiple nights is needed. All this contributes to the fact that an estimated 93% of women and 82% of men with at least moderate SDB are undiagnosed [7].

In 2000, PhysioNet started interest in the topic of less complex apnea detection by holding a competition on their Apnea-ECG Dataset that only consists of labeled ECG recordings split into one-minute epochs. Although presented models reached high performances, later studies showed poor generalizability for these models and indicated that the dataset doesn't fully cover the broad spectrum of apneic events [8]. Therefore, in the last decades, a wide range of sleep disorder datasets and apnea detection architectures were published that focused on generalizability. For instance, Olsen et al. [9] used bidirectional GRUs on ECG data to achieve a sensitivity (Se) of 68.7%, a precision (Pr) of 69.1%, and an F1-score of 66.6% on their self-defined event-level metric and an AHI-correlation of $R^2 = 0.829$. Xie et al. [10] later validated Olsens model on the SOMNIA dataset and achieved an F1-score of 70.8% using the PSG-computed hypnogram and an F1-score of 0.631 with their Multi-Task model that computed sleep stages based on ECG and respiratory effort (RE) only [11]. Also using ECG and RE, Fonseca et al. [12] achieved intraclass correlation coefficient of 0.91 across different datasets.

Using the signal on which sleep apnea is mainly defined on, Airflow, also helps to increase performance greatly. Li et al. [13] achieved an F1-score of 85.7% on classifying one-minute segments of Airflow and ECG. Later, Yook et al. [14] used Airflow and SpO2 together to achieve an F1-score of 93% on classifying 10-second segments converted into scalograms. Downsides to this approach includes that the nasal cannula, a thin tube placed under the nostrils, might be uncomfortable during sleep and hard to set up properly.

One of the more simple signals to set up and record while sleep is PPG, which can be obtained through the use of a pulse oximeter that illuminates the skin to measure changes in light absorption. These devices come in a range of forms such as wrist-worn, like most modern smart-watch already have, or finger-worn, mounted typically on the index finger, which can also calculate SpO2. Lazazzera et al. [15] used PPG and SpO2 signals to achieve a Sensitivity of 76.9% and Specificity of 73.2%, although their dataset only consisted of 96 patients without any kind of co-morbidity.With the same input signals, Wu et al. [16] trained a trasnformer-based model on a dataset containing patients with co-morbidities and were able to validate their performance on PPG and SpO2 signals measured by a Smart Ring resulting in an F1-score of 64.9%.

In this work, we present an event-level apnea detection model that relies solely on signals obtained by easy to use recording hardware, namely PPG and SpO2, and show the importance of correct sleep stage identification.

Chapter 2

Methods

2.1 Dataset

The data we used in this work came from the Multi-Ethnic Study of Atherosclerosis (MESA) [17], a large-scale sleep study aimed to investigate correlations between sleep quality, cardiovescular health, SDB, and other factors across different ethnic groups. Over 6,800 men and women from six different US communities were approached in the initial examination. For the final sleep exam ten years later, 288 participants were ineligible[1], roughly 2,700 were not contacted, and roughly 1,500 refused to participate. From the 2,261 participants undergoing the sleep exam, 2,060 had full-night PSG recordings, 2,156 had actigraphy data, and 2,240 completed a sleep questionnaire.

For the sleep event scoring, we used the automatic Somnolyzer system TODO: cite, which scored events based on the recommended criteria from the American Academy of Sleep Medicine (AASM) TODO: cite: apnea events were defined as a 90% or greater reduction in airflow for at least 10 seconds, while hypopnea events were defined as a 30% or greater reduction in airflow for at least 10 seconds, with either a $\geq 3\%$ oxygen desaturation or an associated arousal. TODO: obstructive vs central vs mixed?

As SDB events manifest differently accross sleep stages, we used a modified version of the hypnogram prediction model from Bakker et al. [18], that used only PPG signals, ensuring that our model works doesn't depend on signals outside of the finger-worn PPG sensor setup. Comparing the predicted hypnogram[2] with the Somnolyzer hypnogram, we achieved a Cohen's Kappa of 0.55,

[1] due to undergoing apnea treatment, living to far away, or other reasons

[2] Bakker's model combined N1 and N2 stages into one, resulting in four stages: Wake, N1/N2, N3, and REM. For calculating the Kappa, Somnolyzer scorings were adjusted to the same format

Fold	N	Age (years)	BMI (kg/m^2)	Sex (N male)	TST (h)
1	470	70 ± 9 **[55, 90]**	29 ± 5 **[19, 48]**	228 (48.5%)	6.2 ± 1.36 **[1.7, 10]**
2	470	70 ± 9 **[54, 90]**	29 ± 6 **[17, 56]**	208 (44.3%)	6.2 ± 1.36 **[1.6, 10]**
3	470	69 ± 9 **[55, 90]**	29 ± 5 **[16, 50]**	229 (48.7%)	6.2 ± 1.47 **[0.7, 10]**
4	470	69 ± 9 **[55, 90]**	28 ± 5 **[17, 50]**	210 (44.7%)	6.2 ± 1.32 **[0.9, 10]**
Full	1880	69 ± 9 **[54, 90]**	29 ± 6 **[16, 56]**	875 (46.5%)	6.2 ± 1.38 **[0.7, 10]**

Table 2.1: Demographic distribution and sleep times of the MESA dataset subset. Format for Age, BMI, and TST is mean ± std [min, max].

showing moderate agreement.

Filtering the MESA participants for those with PPG and SpO2 data, Somnolyzer scorings, and available predicted hypnograms, we ended up with a dataset size of 1,880 participants. Table 2.1 shows the demographic distribution and sleep times of our dataset subset together with the generated folds. To assess SDB severity, the AHI is often categorized into four classes. These so called severity classes are defined as follows: Normal (AHI $<$ 5), Mild (5 $\leq$ AHI $<$ 15), Moderate (15 $\leq$ AHI $<$ 30), and Severe (AHI $\geq$ 30). Table 2.2 shows their distribution. The amount of different apnea classes is shown in table 2.3.

2.2 Signals and Preprocessing

We used the PPG and SpO2 signals from the MESA dataset, which were recorded at 256Hz and 1Hz, respectively. A third input to the model is the hypnogram from Bakker et al. [18], which was predicted at $\frac{1}{30}$Hz and on PPG only, ensuring that the model still relies solely on data it can retrieve from the PPG sensor in the real world. The PPG signal has been denoised using a lowpass filter with a cutoff frequency of 5Hz.

To analyse the importance of correct sleep stage information, we also tested a version of the model that uses the ”ground-truth” Somnolyzer hypnogram instead of the predicted one.

		Severity Class			
Fold	AHI	normal	mild	moderate	severe
1	22.2 ± 18.3 **[0.4, 100]**	61	153	136	120
2	22.0 ± 18.3 **[0.3, 93]**	61	151	134	124
3	21.3 ± 17.1 **[0.4, 95]**	61	151	138	120
4	22.0 ± 18.3 **[0.4, 107]**	61	150	140	119
Full	21.9 ± 18.0 **[0.3, 107]**	244	605	548	483

Table 2.2: AHI and severity class distribution accross folds and full dataset subset.TODO: better use [25%,75%] interval instead of [min, max]? Format for the AHI is mean ± std [min, max].

Fold	obstructive apnea	central apnea	mixed apnea	hypopnea
1	15k (24%)	4k (7%)	1k (2%)	42k (67%)
2	16k (26%)	4k (6%)	1k (2%)	42k (66%)
3	15k (24%)	3k (6%)	1k (2%)	41k (67%)
4	17k (26%)	4k (6%)	1k (2%)	42k (66%)
Full	63k (25%)	16k (6%)	5k (2%)	167k (67%)

Table 2.3: Total number of apnea events per fold and in total. Important to note is the imbalance of the different apnea types, especially the underrepresentation of central and mixed apnea.

PPG Preprocessing

To deal with the high temporal resolution of the PPG signal, we tested three different preprocessing methods that would transform the 256Hz signal into a 1Hz signal with multiple dimensions:

- **Statistical**: On a 1Hz basis we extracted the mean, standard deviation, minimum, maximum, and mean peak interval of the PPG signal, resulting in a 5-dimensional representation of the PPG signal. Due to the nature of PPG showing the heartbeats at 1Hz, we used a sliding window of 5s around the 1Hz point to calculate the statistics.

- **Variational Autoencoder**: The Variational Autoencoder (VAE) is an unsupervised generative model that learns to encode the input data into a lower-dimensional latent space and then reconstruct it back to the original space. The VAE consists of an encoder and a decoder, where the encoder maps the input data to a distribution in the latent space, and the decoder samples from this distribution to reconstruct the input. Using the same sliding window approach as in the statistical method, we trained the VAE to reconstruct the middle 1s from the 5s input window. With that, the encoder learns to compress the input into a lower temporal dimension while preserving the relevant information. For training the main SDB detector model, this encoder is used to transform the 256Hz PPG signal into a 1Hz signal with 8 dimensions.

- **In-model Convolution Stack**: While the prior methods calculated the 1Hz representation of the PPG signal before training the model, we also tested a method that would use a stack of convolution to learn the 1Hz representation during training. The convolution stack consists of five *double conv blocks* TODO: is the name ok so?, which are composed of two 1D convolution layers with a kernel size of 3 or 5 TODO: write letters out or not?, each followed by a batch normalization layer and ReLU activation. Between these blocks are max pooling layers with a kernel size of 4 resulting in the downsampling of the signal to 1Hz, while bringing the number of channels up from 1 to 8. TODO: should I explain the whole model and every single line in more detail in the appendix?

Each preprocessing method brings the PPG signal down to 1Hz with multiple dimensions, which is then stacked together with the 1Hz SpO2 signal and the hypnogram that was upsampled to 1Hz. The input to the detection model is therefore a 1Hz signal with $2 + d$ dimensions, with d being the number of dimensions from the selected PPG preprocessing method(s).

2.3 Model Architecture

The core of the detection model is an adapted version of the U-Net architecture, originally proposed for 2D image segmentation by Ronneberger et al. [19]. The U-Net architecture improves an encoder-decoder structure by adding skip connections between the corresponding encoder and decoder layers, which allows the model to learn both low-level and high-level features TODO: stolen formulation. The adapted model uses 1D convolutions on the temporal dimension instead of 2D convolutions on the width and height of images. The output of the U-Net has the same resolution as the input, which allows the model to classify each second as either part of an event or of normal breathing. This inturn allows us or the user to analyse the prediction on a event level, instead of just the AHI level, which can be important, as studies showed links between apnea event duration and health that go beyond the AHI severity classifications [20]. TODO: create a model figure. how detailed should it be? show the complete u-net? TODO: channel sizes and temp resolution through the data flow

Attention mechanisms

Our model can leverage three types of attention:

- **Self-Attention in the bottleneck**: The self-attention mechanism, originally proposed by Vaswani et al. [21], computes relevance vectors for each input feature through their query (Q) and key (K) matricies. By multiplicating this vector with the value matrix (V), the model learns long-range dependencies throughout the sequence, making it possible to focus on the important parts of the input data. The self-attention mechanism is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

  where d_k is the dimension of the key matrix and the softmax function normalizes the attention scores, ensuring that they sum to 1. Using Self-Attention can increase model complexity greatly due to their quadratic complexity, which is why we apply it in the bottleneck, where the temporal resolution is at its lowest.

- **Attention gates**: Originally proposed for the task of medical pancreas image segmentation by Oktay et al. [22], attention gates are employed at the skip connections of the U-Net and help the model highlight important regions while suppressing irrelevant ones. They work by learning a gate that refines the skip connection (encoder) features before concatenation. This gate is computed from the same incomming skip connection features and the decoder features from the layer below. TODO: add the formula?
- **Squeeze-and-Excitation (SE) blocks**: The SE block, proposed by Hu et al. [23], is a lightweight attention mechanism that computes channel-wise feature importance. It works in two parts: First, the squeeze operation creates feature maps accross the spatial dimensions using global average pooling, resulting in a 1D vector for each channel. Then, in the excitation part, two fully connected linear layers with ReLU activation transform this vector and multiply it with the input, effectively weighting the channels by importance. TODO: add the formula?

TODO: talk about complexity impact? quadratic in time, quadratic in channels

2.4 Training and Evaluation

To ensure statistical validity, we used a fixed seed of 42 and a 4-fold cross-validation approach balanced for AHI severity class. In k-fold cross-validation, the dataset is split into k equal parts (called folds). One then selects one fold as the test set and train the model on the remaining k-1 folds. This process is repeated k times, each time with a different fold as the test set, and the results are averaged to obtain a more reliable estimate of the model's performance. This approach helps to mitigate the risk of overfitting and provides a more robust evaluation of the model's generalization ability. As seen in Tables 2.1, 2.2, and 2.3, the folds are not only balanced for AHI severity class, but also show a good distribution of demographic data.

- Training Parameters (Optimizer, LR, BS, ...) and Setup (Machines, ...)
- Sigmoid and threshold

- Seed and Cross-Validation (mainly balanced for AHI severity, but with seed 42 we got a good distribution fo demographic data in the folds)
- Train on 30min (?) segments. For Testing: Concat 30min Windows with Overlap for full night result.
- Correct results (like Olsen, 10sec minimum event and distance between events)
- Event-based metrics (Se, Pr, F1) and when to count TP, TN, FP, FN
- AHI-based metric (Linear Correlation, Severity Classes, Near-Boundary Double-Classification)

Chapter 3

Results

- Baseline Model vs PPG Preprocessing vs Attention Model Results
- PlethPre reduced training time of detection model by 3x
- Significance of SpO2 and the Hypnogram (No SpO2/Hypnogram, Only PPG Baseline Model)
- AHI correlations and Severity class results

Chapter 4

Discussion

- Discussion and Implications (Is this way applicable in the real world)
- Limitations
- Further work

Chapter 5

Conclusion

- Summerization of paper
- Significance of work
- Outlook

Bibliography

[1] Adam V Benjafield, Najib T Ayas, Peter R Eastwood, Raphael Heinzer, Mary SM Ip, Mary J Morrell, Carlos M Nunez, Sanjay R Patel, Thomas Penzel, Jean-Louis Pépin, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet respiratory medicine*, 7(8):687–698, 2019.

[2] JA Dempsey, SC Veasey, BJ Morgan, and Cp O'DONNELL. Dempsey ja, veasey sc, morgan bj, o'donnell cp. pathophysiology of sleep apnea. physiol rev 90: 47-112, 2010. *Physiological reviews*, 90(2):797–798, 2010.

[3] Susheel P Patil, Hartmut Schneider, Alan R Schwartz, and Philip L Smith. Adult obstructive sleep apnea: pathophysiology and diagnosis. *Chest*, 132(1):325–337, 2007.

[4] Terry Young, Paul E Peppard, and Daniel J Gottlieb. Epidemiology of obstructive sleep apnea: a population health perspective. *American journal of respiratory and critical care medicine*, 165(9):1217–1239, 2002.

[5] GA Gould, KF Whyte, GB Rhind, MAA Airlie, JR Catterall, CM Shapiro, and NJ Douglas. The sleep hypopnea syndrome. *American Review of Respiratory Disease*, 2012.

[6] Michel Toussaint, Remy Luthringer, Nicolas Schaltenbrand, Gabriella Carelli, Eric Lainey, Anne Jacqmin, Alain Muzet, and Jean-Paul Macher. First-night effect in normal subjects and psychiatric inpatients. *Sleep*, 18(6):463–469, 1995.

[7] Terry Young, Linda Evans, Laurel Finn, Mari Palta, et al. Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. *Sleep*, 20(9):705–706, 1997.

[8] Gabriele B Papini, Pedro Fonseca, Jenny Margarito, Merel M van Gilst, Sebastiaan Overeem, Jan WM Bergmans, and Rik Vullings. On the generalizability of ecg-based obstructive sleep apnea monitoring: merits and lim-

itations of the apnea-ecg database. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6022–6025. IEEE, 2018.

[9] Mads Olsen, Emmanuel Mignot, Poul Jorgen Jennum, and Helge Bjarup Dissing Sorensen. Robust, ecg-based detection of sleep-disordered breathing in large population-based cohorts. *Sleep*, 43(5):zsz276, 2020.

[10] Jiali Xie, Pedro Fonseca, Johannes P van Dijk, Xi Long, and Sebastiaan Overeem. The use of respiratory effort improves an ecg-based deep learning algorithm to assess sleep-disordered breathing. *Diagnostics*, 13(13):2146, 2023.

[11] Jiali Xie, Pedro Fonseca, Johannes van Dijk, Sebastiaan Overeem, and Xi Long. A multi-task learning model using rr intervals and respiratory effort to assess sleep disordered breathing. *BioMedical Engineering OnLine*, 23(1):45, 2024.

[12] Pedro Fonseca, Marco Ross, Andreas Cerny, Peter Anderer, Fons Schipper, Angela Grassi, Merel van Gilst, and Sebastiaan Overeem. Estimating the severity of obstructive sleep apnea using ecg, respiratory effort and neural networks. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[13] Fan Li, Yan Xu, Junjun Chen, Ping Lu, Bin Zhang, and Fengyu Cong. A deep learning model developed for sleep apnea detection: A multi-center study. *Biomedical Signal Processing and Control*, 85:104689, 2023.

[14] Soonhyun Yook, Dongyeop Kim, Chaitanya Gupte, Eun Yeon Joo, and Hosung Kim. Deep learning of sleep apnea-hypopnea events for accurate classification of obstructive sleep apnea and determination of clinical severity. *Sleep Medicine*, 114:211–219, 2024.

[15] Remo Lazazzera, Margot Deviaene, Carolina Varon, Bertien Buyse, Dries Testelmans, Pablo Laguna, Eduardo Gil, and Guy Carrault. Detection and classification of sleep apnea and hypopnea using ppg and spo _2 signals. *IEEE Transactions on Biomedical Engineering*, 68(5):1496–1506, 2020.

[16] Zetong Wu, Hao Wu, Kaiqun Fang, Keith Siu-Fung Sze, and Qianjin Feng. A transformer-based deep learning model for sleep apnea detection and application on ringconn smart ring. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2024.

[17] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38(6):877–888, 2015.

[18] Jessie P Bakker, Marco Ross, Ray Vasko, Andreas Cerny, Pedro Fonseca, Jeff Jasko, Edmund Shaw, David P White, and Peter Anderer. Estimating sleep stages using cardiorespiratory signals: validation of a novel algorithm across a wide range of sleep-disordered breathing severity. *Journal of Clinical Sleep Medicine*, 17(7):1343–1354, 2021.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[20] Matthew P Butler, Jeffery T Emch, Michael Rueschman, Scott A Sands, Steven A Shea, Andrew Wellman, and Susan Redline. Apnea–hypopnea event duration predicts mortality in men and women in the sleep heart health study. *American journal of respiratory and critical care medicine*, 199(7):903–912, 2019.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[22] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[23] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.