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, EEG) activity, airflow, peripheral oxygen saturation (SpO2), chest and abdominal movements, sleeping position, and blood volume changes (photoplethysmography, 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 ground-truth, PSG-computed hypnogram and an F1-score of 0.631 with their Multi-Task model that predicted sleep stages based on ECG and respiratory effort (RE) only [11], highlighting the performance difference between sleep stages computed from the full PSG and those computed from only a subset. 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 during 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 watches 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. We show the performance on PPG only, which could be used in devices that can only measure PPG. Finally, we discuss the importance of correct sleep

stage identification and show that our model can be used without ground-truth, PSG-computed hypnograms.

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, cardiovascular 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 [18], which scored events based on the recommended criteria from the American Academy of Sleep Medicine (AASM) [19]: 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.

As SDB events manifest differently accross sleep stages, we used a modified version of the hypnogram prediction model from Bakker et al. [20], 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, showing moderate agreement.

[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].

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 folds, generated for cross validation as discussed later in this chapter. 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 ≥ 30). Table 2.2 shows their distribution. The number 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. [20], 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. We denoised the PPG signal 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. 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 convolutions to learn the 1Hz representation during training. The convolution stack consists of five *double convolution blocks* (DCB) which are composed of two 1D convolution layers with a kernel size of 3 or 5, 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.

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. [21]. 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. 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 in turn allows us or the user to analyse the prediction on an 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 [22]. Figure 2.1 shows the model architecture and the DCBs, that are also used for the preprocessing VAE.

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. [23], 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.

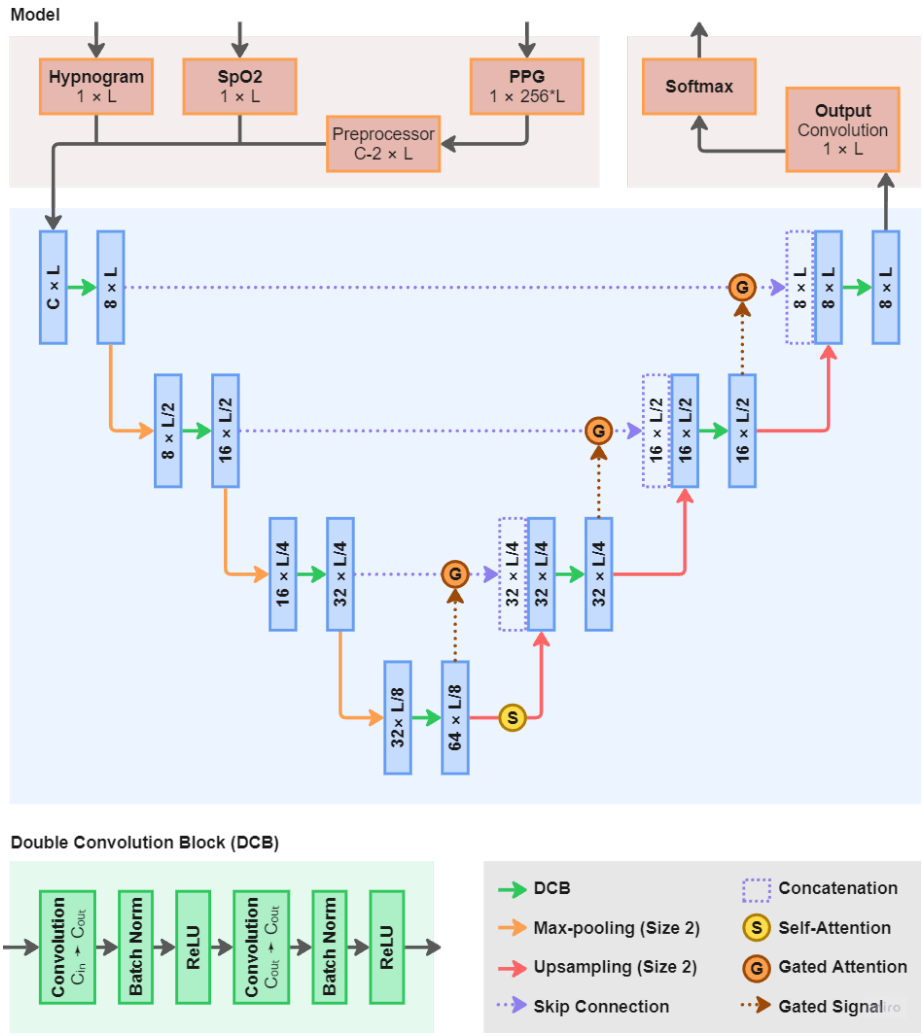

Figure 2.1: Architecture of the model and shape of the data flowing through the network. L is the sequence length, which is 30 minutes in our case. C is the number of input features (or channels for the convolutions), with the hypnogram and SpO2 signal having one channel each, while the number of PPG channels depend on the preprocessing technique used. The convolutions kernel sizes are 3 for the DCBs and 1 for the output convolution.

- **Attention gates**: Originally proposed for the task of medical pancreas image segmentation by Oktay et al. [24], 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 incoming skip connection features and the decoder features from the layer below.

2.4 Training and Evaluation

Cross-Validation

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 trains 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 evaluation results on the test sets 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. Figure 2.2 illustrates the cross-validation process. 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 and Setup

The training targets are one-dimensional vectors with the same length as the input sequence, where each second is labeled as either 0, indicating normal breathing, or 1, indicating an apnea event. The final layer of the model is a sigmoid activation function, that maps each second to an event probability value between 0 and 1. During training, we optimize the binary cross-entropy loss[3] (BCE) between the predicted probabilities and the true labels. The BCE loss is defined as:

$$\text{BCE} = \frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \tag{2.2}$$

[3]specifically, we use PyTorch's `nn.BCEWithLogitsLoss()`, which combines the sigmoid activation and BCE loss into one class, as it is more stable than doing these operation in sequence.

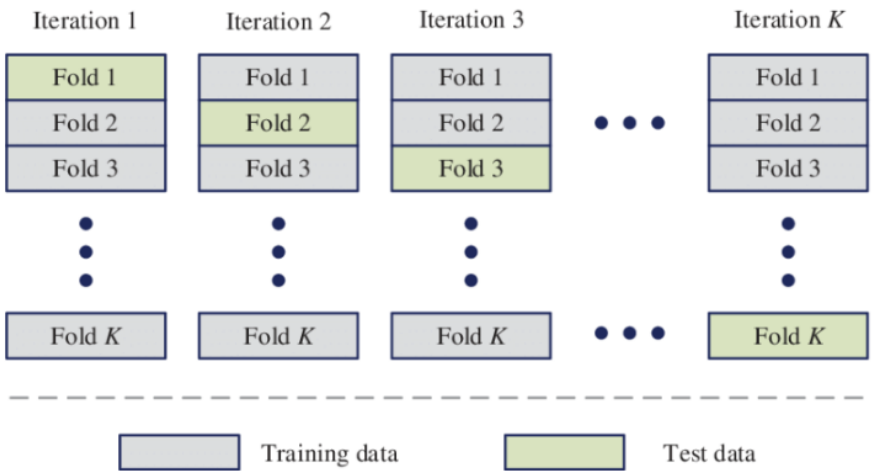

Figure 2.2: Structure of a k-fold cross-validation. Source: `https://www.researchgate.net/figure/K-fold-cross-validation-method_fig2_331209203`

where N is the number of samples (or seconds in our case), y_i is the true label (0 or 1, normal or abnormal breathing), and p_i is the predicted probability for sample i. The loss is then averaged over each batch and parsed to the optimizer, in our case the Adam optimizer with a learning rate of 0.001.

As for batching, we used randomly selected 32 30-minute segments from four different recordings each to mitigate batch overfitting on sleep patterns of a single participant.

During testing, we used a sliding window apprach, were each recording was split into 30-minute segments with a 2-minute overlap 2.3. Model predictions were then concatinated, disregarding each first and last minute to create the final prediction for the whole night. This approach allows us to predict recordings of arbitrary length, on which metrics like the AHI can be calculated.

Each fold has been trained on a single NVIDIA A40 GPU with 48GB VRAM with a time limit of 2 days for 30 epochs.

Evaluation Metrics

Several metrics are employed to measure the perfomance of the model, that can be divided into three categories:

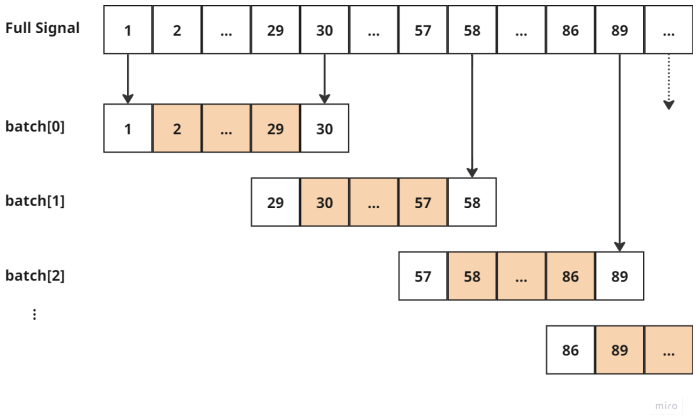

Figure 2.3: Dividing the full recording into 30-minute segments with a 2-minute overlap. The first and last minute of each segment are disregarded, so that the concatenation resembles the original recording for the final prediction.

A. Event-level metrics

To assess the model perfomance on an event level, regardless of the length of the night, we use the event-level metrics. They are calculated by extracting events from the predicted probabilities by thresholding the probabilities and counting each consecutive sequences of 1s as a single event. To mitigate outliers, we disregarded events shorter than 3 seconds and combined consecutive events that are less than 3 seconds apart into one event. We call this the *output correction*.

Olsen et al. [9] defined scoring rules on the event-level that are defined as follows: A predicted event, that overlaps with a true event, gets classified as a true positive (TP). If a predicted event has no overlapping true event, it gets counted as a false positive (FP). If a true event doesn't overlap with any predicted event, it gets counted as a false negative (FN). Note that there are no true negatives (TN) on the event-level. In this work, we use a more strict version of their rules, that were adjusted by Xie et al. [10], and in which each event can only be used for scoring one time. Meaning that if a predicted event overlaps with multiple true events, only one true event gets counted as TP, while the others are counted as FN. The same applies the other way around, where a true event can only be counted as TP once and other overlapping predicted events are counted as FP. A visual example can be seen Figure 2.4. From this, we can now compute the following metrics:

TODO: Add newpage if the table isn't above this As these metrics depend on the selection

Metric	Calculation	Meaning
Recall (Rec)	$\frac{TP}{TP+FN}$	What % of real events got detected?
Precision (Pr)	$\frac{TP}{TP+FP}$	What % of predicted events where real events?
F1-score	$2 * \frac{Pr*Re}{Pr+Re}$	Harmonic mean of Precision and Recall

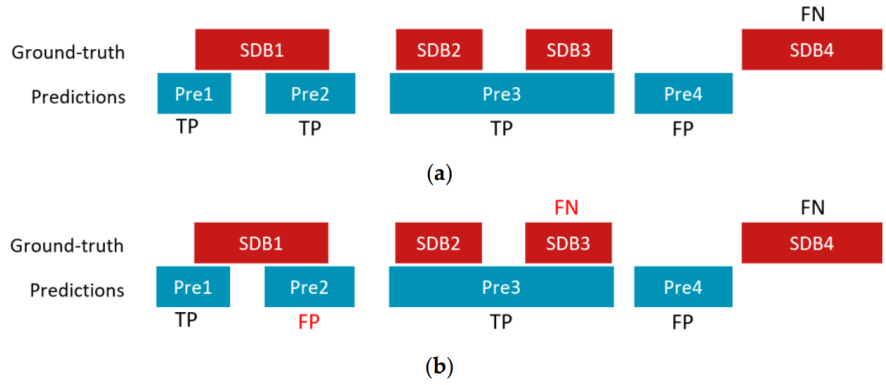

Figure 2.4: Example of the event scoring. (a) shows the version from Olsen et al. [9]. (b) illustrates the extra rules of the version we use: As every pair of TPs can only be scored once, Pre2 and SDB3 are counted as FP and FN respectively. Source: Figure taken from [10].

for a proper threshold, we TODO: we or we'll? show these metrics as a function of the threshold.

B. AHI-level metrics

Dividing the total number of apnea events by the TST (in hours) gives us the most common metric for SDB severity, the AHI. We compare the predicted AHI (AHI_{pred}) and the Somnolyzer AHI (AHI_{true}) using the following metrics:

- Plotting AHI_{pred} against AHI_{true} shows their correlation which can also be expressed in the **Root Mean Square Error** (RMSE) and the R^2 value. RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(AHI_{pred} - AHI_{true})^2} \tag{2.3}$$

where n is the number of samples (or participants AHIs in our case). R^2 is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{2.4}$$

where y_i is the i-th predicted AHI, $\bar{y}_i$ is the mean of the predicted AHIs, and f_i is the i-th true AHI. The R^2 value ranges from 0 to 1, where 0 indicates no correlation and 1 indicates a perfect correlation.

- The **Bland-Altman plot** is another way to visualize agreement by plotting the difference between the two AHI values against their mean. This allows us to see the bias, defined as the mean difference, and limits of agreement, defined as the bias $\pm$ 1.96 times the standard deviation of the differences. The limits of agreement are the range in which 95% of the differences between the two AHI values are expected to fall.

- The **Spearman's rank correlation coefficient** (ρ) ignores the actual values of the AHIs and only looks at their ranks. This is useful for measuring the strength of the monotonic relationship between the two AHI values, regardless of their actual values. It is computed as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2.5}$$

  where d_i is the difference between the ranks of the two AHI values for each participant, and n is the number of participants. The coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

- Finally, we also calculated the **Intraclass Correlation Coefficient** (ICC) TODO: what is it? which one do we use? how is it computed maybe?

C. Severity-class-level metrics

The last set of metrics we used is the severity-class-level metrics, which are calculated on the AHI severity classes. As mentioned, the boundaries for the severity classes are defined as follows: Normal (AHI < 5), Mild ($5 \leq$ AHI < 15), Moderate ($15 \leq$ AHI < 30), and Severe (AHI ≥ 30). As small errors in the AHI around these hard thresholds can lead to a wrong classification, we used near-boundary double-labeling (NBL), which allows us to assign two classes for

AHIs that fall in the range of about 2.5 around the boundaries. Exact values can be found in TODO: appendix.

Using these four classes we can plot the confusion matrix and compute model **Accuray** (Acc), defined as the number of correctly classified patients divided by the total number of patients, and the **Cohen's Kappa** (κ) TODO: what is it? how is it computed?.

Finally, this matrix can be binarized to assess the discrimination ability between normal to abnormal (mild-severe), mild to moderate, and moderate to severe SDB. TODO: should I create a table explaining all (Acc, Sen, Spe, PPV, NPV, LR+, LR-) metrics again or is LR enough? Metrics on this binarized view include the **Likelihood ratios** (LR), which give insight in how much a test result changes the odds of having the disease, or in our case, the specific severity classes. These likelihoods can be computed as positive and negative likelihood ratios (LR+ and LR-), which are defined as:

$$LR+ = \frac{Sensitivity}{1 - Specificity} \tag{2.6}$$

$$LR- = \frac{1 - Sensitivity}{Specificity} \tag{2.7}$$

Chapter 3

Results

- TODO: should I show the loss and reconstruction of the VAE?
- TODO: should I show an example of the models output
- TODO: should I talk about the denoising techniques? should I show an example of them? Or just the one I used (lowpass)? Results are without cross validation
- TODO: don't show the kfold balancing analysis, right? table was enough
- TODO: should I show the correctify size analysis
- TODO: Can I show the no hypno model results (no cross validation)
- TODO: Should Ishow the results of the runs with no improvements: Multiscale CNN (what about SE), Reduce LR

3.1 PPG Preprocessing through the VAE

One of the preprocessing techniques used to "downsample" the PPG signal was the Variational Autoencoder (VAE), whose encoder could be used to transform each 256Hz second into a 1Hz value of 8 dimensions. Figure TODO: ref show the example reconstructions of the four model variations we tried. *linVAE* and *convVAE* refere to the type of the layers in the model, being either linear or convolutional respectively. *1s* and *5s* refer to the input length. While the *1s* model would reconstruct the inputted second, the *5s* model would reconstruct only the middle one second of the 5 second window around it. Figure TODO: ref plots the reconstruction loss over the epochs and shows that the convolutional model with the 5-second window performed best and was therefore used in the next experiments. TODO: linVAE 5s was obviously stuck in a local minima and

could not converge during training, should I still show it? Also the linVae 1s has only been trained for 10 epochs instead of 50. Should I run it again?

3.2 Preprocessing impact on performance

Figure TODO: ref shows the recall, precision, and F1-score TODO: do I have the exact training time? for the SDB detection model with the different preprocessing techniques. While neither the statistical nor the VAE preprocessing approach reached the same performance as the in-model approach, using both statistical and VAE preprocessing together did reach a similar performance. As both these values would only be needed to be calculated once before the training and not during each epoch, which the in-model approach did, training time got reduced significantly by a factor of 3.

3.3 SDB Detection Model TODO: put this section before preprocessing?

Event-level performance

Figure TODO: ref shows the recall, precision, and F1-score over each threshold for the main SDB detection model, which uses the PPG-predicted hypnogram, the PPG itself with the in-model technique, and the SpO2. The Figure also shows a version of the model without the SpO2 signal, which means it relies solely on the PPG data. As can be seen, omitting the SpO2 signal has a significant impact on the performance, as the peak F1-score drops from TODO: exact values.

Test and training losses together with the peak F1-score over the epochs are displayed in Figure TODO: ref. While the version without SpO2 seems to train slightly more stable, learning convergences much slower than the one with SpO2, which reaches the area of the final peak F1-score in the first few epochs.

The threshold for the best performance was determined to be TODO: exact value and the final event-level metrics are shown in Table TODO: ref.

TODO:

- Evaluation per Event Classes
- Evaluation per Sleep Stage

- Event lengths

AHI-level performance

Figure TODO: ref shows the scatter plots for the predicted and true AHI values of both versions of the model TODO: address the bias towards predicting lower AHIs. To assess agreement, Figure TODO: ref displays the corresponding Bland-Altman plots. We found a TODO: highlight the interesting metrics, like level of agreement, ICC, Spearman's rank, ... All AHI-level metrics can be found in Table TODO: ref.

Severity-class-level performance

Figure TODO: ref shows the confusion matrices for the predicted severity classes using the hard thresholds and the NBL version. Although a strong focus on the true prediction diagonal can be seen, the bias towards predicting lower severity classes is also visible like in the AHI-level results.

We shows the models discrimination ability in Table TODO: ref. TODO: which values to highlight?

3.4 Importance of correct Sleep Stages

TODO: figure .. shows metrics of predHypno model compared to GTHypno and maybe NoHypno. Performance decreases significantly

3.5 Correcting model output

After applying the threshold for the prediction, a correctification step was applied. This step removed events shorter than a specified number of seconds (called the correctification size) and merged events that were closer than the correctification size. Figure TODO: ref, F1-heatmap and prec-recall plots by corr size displays the impact of the correctification size and shows that settings this value too low allows more prediction errors to pass through, while setting it too high removes many true positives.

Chapter 4

Discussion

In this work, we presented a, automatic, data-driven SDB detection model based on an Attention U-Net with state-of-the-art performance, that can tackle the problem of the huge number of undiagnosed sleep apnea cases, due to its selection of uncomplicated and inobtrusive input sensors. We achieved a peak F1-score of 76.6% in event detection and an AHI prediction correlation of ???, which could be even further improved through the use of linear of MLP-based corrections. We showed great diagnostic results with positive and negative likelihood ratios of ≥ 16 and ≤ 0.28 respectively with very few participants being wrongly classified more than one severity class apart. All metrics were based on our strict event scoring that is more transparent than minute-to-minute, segment classification.

Our model demonstrated higher detection rates for apnea events compared to hypopnea events. This result, while counterintuitive given the event distribution in the dataset, is expected because hypopneas are harder to detect due to their lower impact and not always being associated with desaturation events.

One goal of our work was to use only PPG and SpO2, as the finger-worn sensor recording these signals, is easy to set up and unobtrusive during sleep. However, an even less obtrusive option are smart watches or smart rings, that already today can record PPG at good quality. To our knowlegde, SpO2 cannot be realiably be recorded using these devices, especially not the subtle drops in saturation on which some classification is based on. We showed that omitting SpO2 data from the training signals decreased performance, but not as much as expected. The model was still able to detect SDB events with a peak F1-score of TODO: X% and predict AHI with a correlation of TODO: X. This means that our model is still usefull on these unobtrusive technologies that only measure PPG, and can be used over many nights. This greatly helps with diagnositic meaningfulness, screening, analyzing night-to-night variability, and evaluating

effectiveness of SDB treatment methods. Also, TODO: appendix XY showed that training without SpO2 data was, while plateauing way slower, more stable than than training with it, which is likely due to the fact that SpO2 in itself is less stable and prone to artifacts.

Another important factor of our models performance is the use of sleep stage labels. The model performed best when using ground truth sleep stages from Somnolyzer and worst without any sleep stage information. The PPG-derived sleep stages from [20] greatly improved results over using no sleep stage information but were still not as good as using the ground truth. Further improvements in predicting the hypnogram from only PPG signals could lead to better results in detecting SDB, that are still relying solely on data from PPG sensors.

We also looked into preprocessing the PPG signal using statistical analysis and a VAE. While achieving the same level of perfomance as using the in-model approach, training time of the detection model decreased significantly. Inference time will not be affected, as the benefit comes only from not needing to recompute the preprocessed signal for each training run, but this approach could help with rapid prototyping and hyperparameter tuning.

Finally, correcting the models output by filtering out events shorter than 3 seconds and merging events less than 3 seconds apart into one, we saw an increase in event-level results, while general AHI-level performance decreased slightly. Looking at the positive and negative diagnostic performance, we see that correcting the output improves results on high AHI participants while slightly lowering results on low AHI participants. This explains lower general AHI-level performance, as the dataset is slightly biased towards low AHIs.

An important limitation of our work is the lack of validation on other datasets. While the MESA dataset we used is large and greatly balanced in some regards, like AHI, BMI, smoking habits, or co-morbidities, other factors like age are not balanced. Recordings have also been made in a clinical setting and with the same hardware. Even further, first-night effects haven't been adressed. Validation our work on other datasets, is crucial to show generalizability and usefulness of our model in the real world.

Future work could tackle the bias and errors in AHI prediction. While a linear correctioin could help correcting the bias, other studies have shown that using demographic data to refine the AHI through a small MLP can increase correlation greatly.

As the use of smart watches or smart rings maximizes or goal of inobtrusive

SDB detection even further, future work could also look into using other sensors that are already available on these devices. One example is the accelerometer, which records movements during sleep, or breathing sounds, that are mainly used for detecting snoring. Both signals are indicators of SDB events and might improve our results even further.

Chapter 5

Conclusion

In this work we aimed to create an easy to set up and unobtrusive way to diagnose SDB, an illness that, while having significant health implications, is often undiagnosed. We have shown that using only signals from a finger-worn PPG sensor, we can detect SDB with high accuracy. While our method is not perfect and performance of the gold standard PSG is still not reached, we believe that our method is a step in the right direction, helping with screening and pre-diagnosis of SDB.

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] Peter Anderer, Marco Ross, Andreas Cerny, and Edmund Shaw. Automated scoring of sleep and associated events. In *Advances in the Diagnosis and Treatment of Sleep Apnea: Filling the Gap Between Physicians and Engineers*, pages 107–130. Springer, 2022.

[19] Matthew M Troester, Stuart F Quan, Richard B Berry, American Academy of Sleep Medicine, et al. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine, 2023.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.