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]. The most common SDB disorder is Obstructive Sleep Apnea (OSA), which clinical manifestations include sleepiness, fatigue, cardiovascular disease, and hypertension. SDB in general is linked to higher cases of diabetes, stroke occurrences and increased morbidity [2, 3, 4].

The gold standard for assessing SDB is Polysomnography (PSG), which captures physical and biological signals like cardiac (electrocardiogram, ECG) and neurological (electroencephalogram, EEG; electrooculography, EOG; electromyography, EMG) activity, airflow, peripheral oxygen saturation (SpO2), thoracic and abdominal respiratory effort, sleeping position, and blood volume changes (photoplethysmography, PPG).

The 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]. The current guidelines for scoring of respiratory events [6] recommend scoring an apnea when there is a decrease of at least 90% in the airflow amplitude in respect to baseline, and a hypopnea when there is a decrease between 30% and 90% in airflow amplitude, associated with a cortical arousal (measured with EEG) or a decrease of $\geq 3\%$ in the level of SpO2 compared to the pre-event baseline. 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.

Besides respiratory events, the PSG is used to score sleep stages,distinguishing between periods of wakefulness (or wake), rapid eye movement (REM) sleep and non-REM sleep, which is further divided in N1, N2 and N3. By adding up the

time spent in each non-wake stage, the total sleep time (TST, measured in hours) is calculated. Dividing the number of apneas and hypopneas by the total sleep time (TST) gives the Apnea-Hypopnea-Index (AHI), which indicates the severity of SDB, and which combined with the clinical presentation is used for diagnosis.

Although PSG is the gold standard measure for assessing sleep and diagnosing SDB, it 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, limiting its use to one or two nights. Secondly, looking only at a single night might have low diagnostic meaningfulness [7] and hide within-subject variability in the assessment of the condition, which can only be elucidated by monitoring multiple nights. Polygraphic setups reduce the number of sensors to a subset of the full PSG required for adequately scoring respiratory events, recording only airflow, SpO2, and respiratory effort. These so called home sleep apnea tests (HSAT) are increasing in popularity due to their reduced complexity and cost, but they still remain relatively uncomfortable and are effectively limited to only a few nights of recording. All these factors contribute to an estimated 93% of women and 82% of men with at least moderate OSA that remain undiagnosed [8].

In 2000, a PhysioNet kick-started interest in the topic of surrogate assessment of sleep apnea with simpler sensors, by holding a competition on their Apnea-ECG Dataset that consisted only of labeled ECG recordings split into one-minute epochs. Although submitted models reached high performances, later studies showed that these exhibited poor generalizability, suggesting that the dataset doesn't fully cover the broad spectrum of apneic events and may not be representative of real-world, clinically meaningful cohorts [9]. Therefore, the last decades saw a wide range of studies, with a multitude of clinical datasets containing different sleep disorders and architectures for apnea detection that focused on reliability and generalizability. For instance, Olsen et al. [10] used the Sleep Heart Health Study (SHHS) [11] and the Multi-Ethnic Study of Atherosclerosis (MESA) [12] datasets to develope a neural network with bidirectional GRUs that used ECG inputs 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. [13] later validated Olsens model on the Sleep and Obstructive Sleep Apnoea Monitoring with Non-Invasive Applications (SOMNIA) [14] dataset achieving an F1-score of 70.8%. Both Olsen's and Xie's studies relied on the ground-truth sleep stages scored

from PSG to calculate the TST, required to obtain the AHI. In a follow-up study, Xie et al. [15] developed a multi-task model that in addition to SDB events, also predicted sleep and wake phases based on ECG and respiratory effort (RE) only, achieving an F1-score of 0.631. Xie's two studies highlighted the performance decrease when using surrogate signals to calculate sleep stages compared to using the full PSG. Also using ECG and RE, Fonseca et al. [16] achieved intraclass correlation coefficient of 0.91 across different datasets.

Notable about these studies is that they do not rely on the airflow signal acquired typically with PSG and HSAT, based on which apnea and hypopnea events are scored. Unsurprisingly, using airflow as input helps increase performance greatly. Li et al. [17] achieved an F1-score of 85.7% on classifying one-minute segments of Airflow and ECG. Later, Yook et al. [18] used Airflow and SpO2 together to achieve an F1-score of 93% on classifying 10-second segments converted into scalograms. The downsides to this approach are that the sensors used to obtain airflow, i.e. nasal cannulas, a thin tube placed under the nostrils, or thermistors, a thermocouple sensor placed on the upper lip, are uncomfortable during sleep and hard to set up properly.

One of the simpler signals to set up and record during sleep is PPG, which can be obtained with 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 as used in PSG and HSATs, mounted typically on the index finger, and which can also calculate SpO2. Lazazzera et al. [19] used PPG and SpO2 signals, achieving 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. [20] trained a transformer-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 with easy-to-use sensors, namely PPG and SpO2. We evaluate the performance using the combination of both sensors, and with PPG only, which can be used with devices that cannot accurately measure SpO2, such as wrist-worn wearables. Finally, we evaluate the impact of using sleep stages scored from the gold standard PSG versus using surrogate sleep stages predicted from PPG only.

Chapter 2

Methods

2.1 Dataset

The data we used in this work came from the Multi-Ethnic Study of Atherosclerosis (MESA) [12], 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.

To obtain ground-truth SDB events, we used the automatic Somnolyzer scoring system [21], which scored the respiratory events based on the recommended criteria from the American Academy of Sleep Medicine (AASM) [6]: 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.

To leverage the different expressions of respiratory events in different sleep stages, we explored the use of sleep stage classes as an additional input to our model. To do so, we used a previously developed sleep staging algorithm created by Bakker et al. [22], restricted to only use PPG signals as inputs, ensuring that our algorithm doesn't depend on signals outside of the finger-worn PPG sensor setup. We achieved a pooled Cohen's Kappa of 0.55 when measuring agreement between the PPG-predicted hypnogram[2] and the the ground-truth Somnolyzer

[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,

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].

PSG-derived hypnogram, showing moderate agreement. A detailed performance comparison can be found in TODO: Appendix XY.

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. [22], 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

N1/N2, N3, and REM. For calculating the Kappa, Somnolyzer scorings were adjusted to the same format

		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.

filter with a cutoff frequency of 5Hz.

TODO: data: events by real sleep stages, data: tst true vs pred correlation

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.

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. [23]. 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 [24]. 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. [25], computes relevance vectors for each input feature through their query (Q) and key (K) matrices. 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:

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

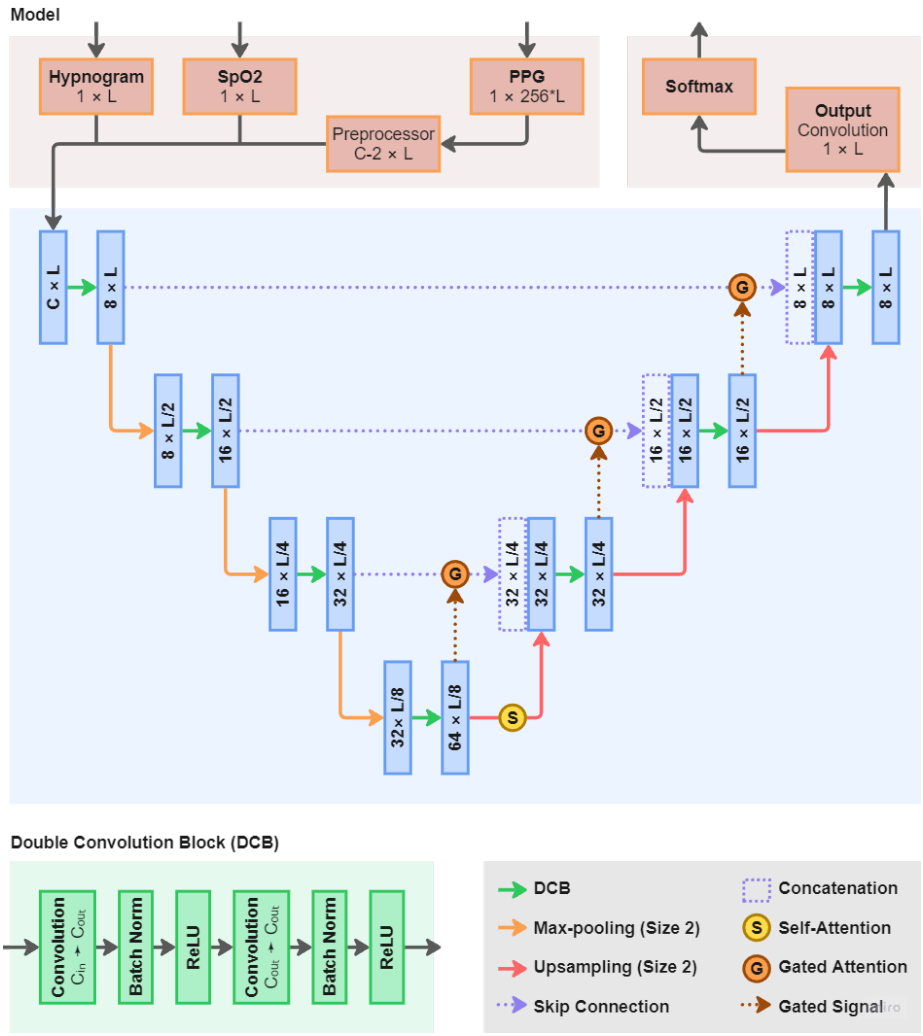


Figure 2.1: Architecture of the model and shape of the data flowing through the network. L is the sequence length, which is 30 minutes in our case. C is the number of input features (or channels for the convolutions), with the hypnogram and SpO2 signal having one channel each, while the number of PPG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

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. [26], 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

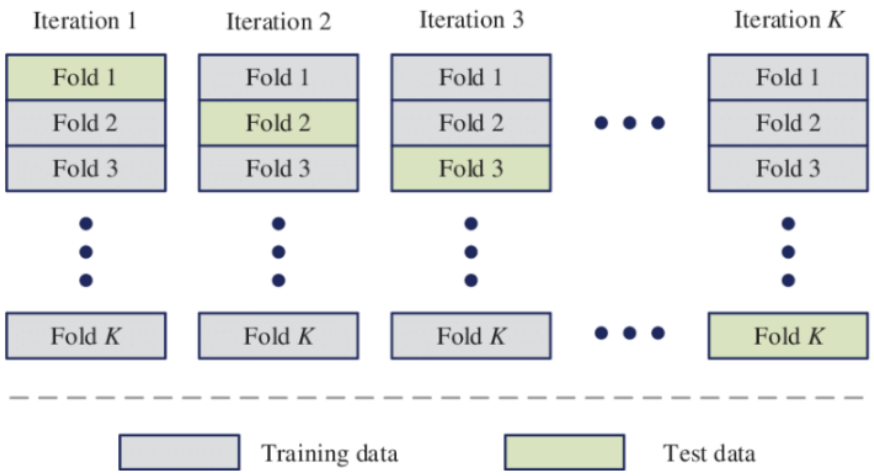


Figure 2.2: Structure of a k-fold cross-validation. Source: `https://www.researchgate.net/figure/K-fold-cross-validation-method_fig2_331209203`

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:

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

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 approach, were each recording was split into 30-minute segments with a 2-minute overlap 2.3. Model predictions were then concatenated, 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

[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.

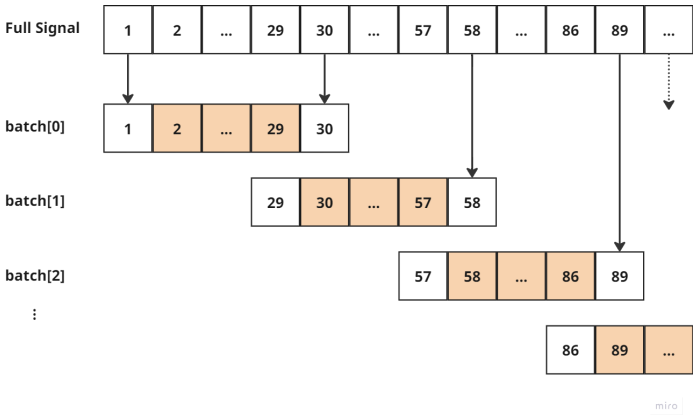


Figure 2.3: Dividing the full recording into 30-minute segments with a 2-minute overlap. The first and last minute of each segment are disregarded, so that the concatenation resembles the original recording for the final prediction.

with a time limit of 2 days for 30 epochs.

Evaluation Metrics

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

A. Event-level metrics

To assess the model performance 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. [10] 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. [13], and in which each event can only be used for scoring one time. Meaning that if a predicted event

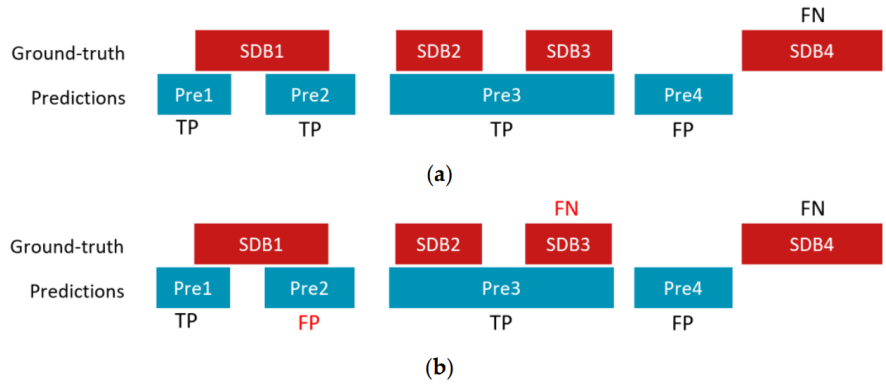


Figure 2.4: Example of the event scoring. (a) shows the version from Olsen et al. [10]. (b) illustrates the extra rules of the version we use: As every pair of TPs can only be scored once, Pre2 and SDB3 are counted as FP and FN respectively. Source: Figure taken from [13].

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:

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

As these metrics depend on the selection 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 calculate the **Intraclass Correlation Coefficient** (ICC), which is a measure of reliability between two or more raters, in our case the predicted and true AHI values. We use the ICC(2,1) version, which is a two-way random effects model for absolute agreement. It ranges from 0 to 1, where 0 indicates no agreement and 1 indicates perfect agreement. TODO: no formula

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), which is a measure of agreement between the predicted and true classes, taking into account the possibility of random agreement. Cohen's Kappa is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.6}$$

where p_o is the observed agreement and p_e is the expected agreement. Its value ranges from -1 to 1, where -1 indicates perfect disagreement, 0 indicates no agreement, and 1 indicates perfect agreement.

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:

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

$$LR- = \frac{1 - Sensitivity}{Specificity} \tag{2.8}$$

Chapter 3

Results

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. We tried two versions: The first got a 1-second input and had to reconstruct the exact second. The other one go a 2-second window around the second it should reconstruct. Figure 3.1 shows the example reconstructions of the two variations and Figure 3.2 plots the reconstruction losses over the epochs. As the 5s VAE had a lower loss, we used its encoder for the VAE preprocessing option.

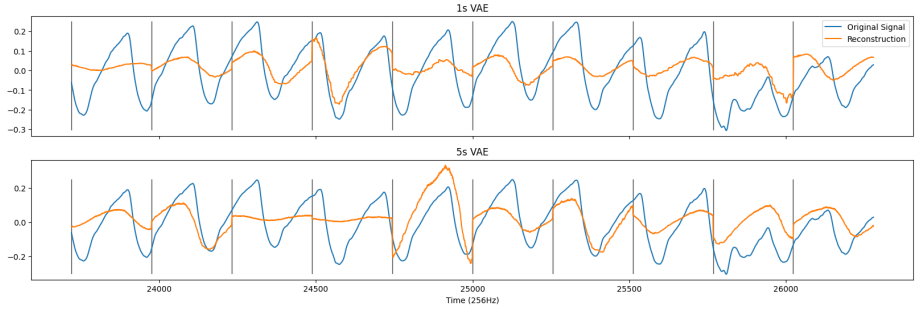


Figure 3.1: Example difference in reconstructions of the 1s and 5s VAE for the same signal. Each second has 256 values, which are reduced to only 8 values.

3.2 Preprocessing impact on performance

Figure 3.3 shows the recall, precision, and F1-score for the SDB detection model with the different preprocessing techniques. While neither the statistical nor

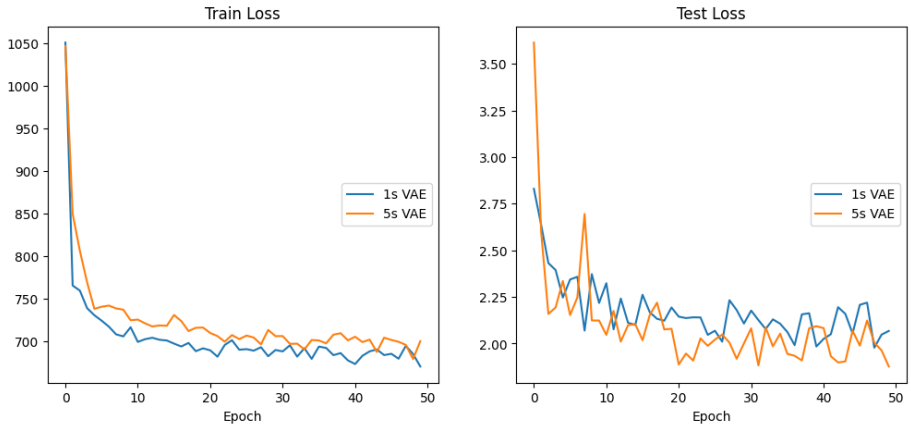


Figure 3.2: Train and test loss of both VAEs. Although the 1s VAE had a lower train loss, the 5s performed better on the test set, which could be a sign of better generalization.

Method	Training time	Testing time
In-model	145min	34min
Statistical	46min	12min
VAE	47min	12min
Stat. + VAE	50min	12min

Table 3.1: SDB detection model training and testing times in minutes. The in-model approach took roughly three times as long.

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. This can be seen in Table 3.1.

3.3 SDB Detection Model

Event-level performance

Figure 3.4 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

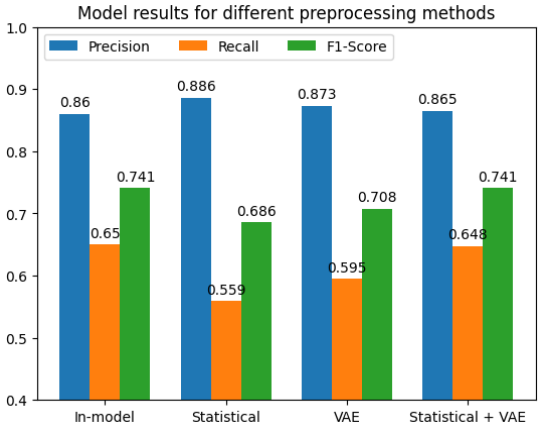


Figure 3.3: Precision, recall, and F1-score of the SDB detection model with different preprocessing techniques. Although the precision didn't change much, the recall and therefore the F1-score dropped significantly, when using the statistical or VAE preprocessing only. Important to note is that these results came from experiments with the ground-truth hypnogram, which is not the final model, as the final model uses the PPG-predicted hypnogram.

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 69.7% to 61.6%. The threshold for the best performance was determined to be 0.25.

Test and training losses together with the peak F1-score over the epochs are displayed in Figure 3.5. 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.

We show our final event-level metrics together with a comparison to other studies in Table 3.2.

With the threshold of 0.25, we can analyse the performance based on the event class and sleep stage. Table 3.3 shows the distribution of event classes in the dataset and the models detection rate. The dataset is highly imbalanced, with 2/3 of all events being hypopneas. Still, the model was best in detecting mixed apneas, with a near 90% detection rate, while hypopneas were only detected about 2/3 of the time.

In appendix TODO: ref we show and discuss the results per sleep stage and the length of the predicted events.

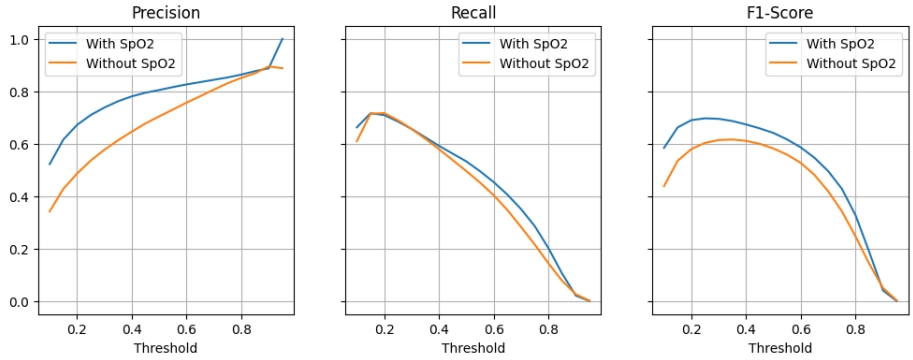


Figure 3.4: Comparison of the event-level metrics of the SDB detection model with and without SpO2. The model with SpO2 reached a peak F1-score of 69.7% at a threshold of 0.25, while the one without SpO2 only reached a peak F1-score of 61.6% at a threshold of 0.35.

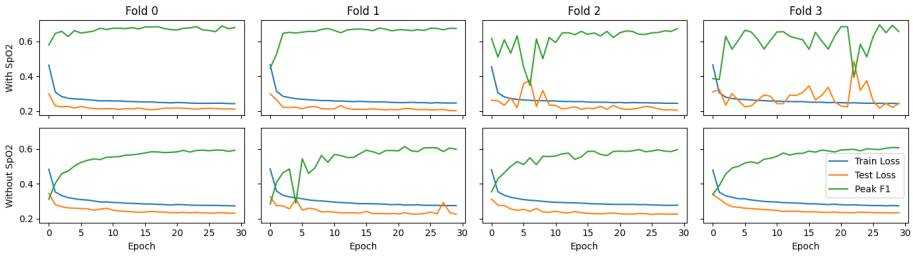


Figure 3.5: Losses and peak F1-score by epoch for every fold.

Model	Signals	Prec.	Rec.	F1
[10]*	ECG	73.4%	70.9%	72.1%
[13]*	ECG, RE	56.5%	77.4%	70.8%
[15]	ECG, RE	63.3%	63.0%	63.1%
[17]**	Airflow, EEG	87.3%	83.7%	85.7%
[18]**	Airflow, SpO2	93.0%	91.0%	93.0%
[19]**	PPG, SpO2	-	76.9%	-
Ours	PPG, SpO2	70.94%	68.46%	69.68%

Table 3.2: Result comparison between other work and our SDB detection model. Models that use the airflow signal achieve the best results. *Studies that use the ground-truth, PSG-computed sleep stages. **Studies that classify 60- or 10-second long epochs instead of events, as our event scoring metric does.

	Obstructive Apnea	Mixed Apnea	Central Apnea	Hypopnea
Total (N)	61161	4811	15240	162536
% of all	25.1%	2.0%	6.3%	66.7%
Found (N)	43243	4299	12486	111826
% found	70.7%	89.4%	81.9%	68.8%

Table 3.3: Distribution of event classes in the full dataset and how many of the different classes were detected by our model. Although the dataset is greatly inbalanced to hypopneas (2/3 of all) and against mixed apneas (only 2%), the detection rates greater for apneas than for hypopneas.

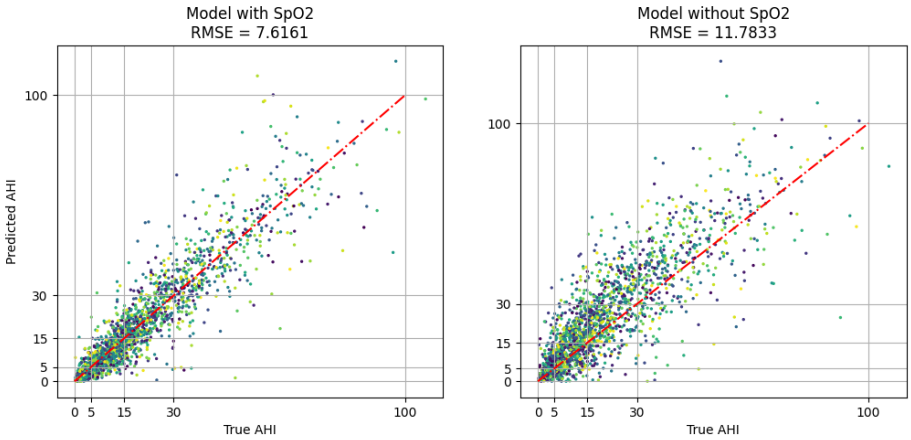


Figure 3.6: Preedicted AHI plotted against the ground-truth AHI. The left plot shows the model with SpO2 and PPG. The right one shows the result of using only PPG as input. The red line is the identity line. The grid shows the differen AHI severity class boundaries.

AHI-level performance

Figure 3.6 shows the scatter plots for the predicted and true AHI values of both versions of the model. To assess agreement, Figure 3.7 displays the corresponding Bland-Altman plots. Both plots show a bias towards predicting higher AHIs for the model without SpO2, while the one with PPG and SpO2 shows lower deviation and near to no bias. We also got a lower RMSE of 7.6 instead of 11.8 when using SpO2. For the model with SpO2 we achieved a Spearman rank correlation of 0.917 and an intra-class correlation of 0.91. All AHI-level metrics and a comparison to other work can be found in Table 3.4.

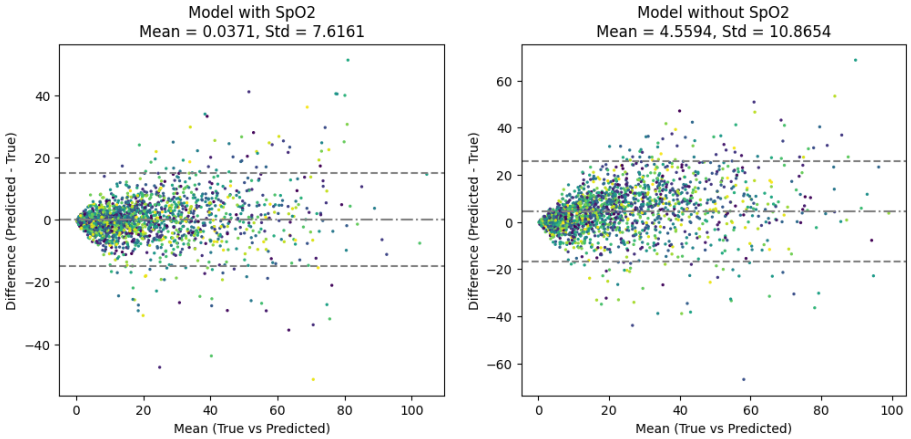


Figure 3.7: Bland-Altman plots for the true and predicted AHI values. The left plot shows the model with SpO2 and PPG. The right one shows the result of using only PPG as input. The grey line is the mean difference and the grey, dashed lines are levels of agreement, computed as 1.96 times the standard deviation of the differences.

Model	ρ	ICC, 95%CI	RMSE	Bias	SD error
PPG + SpO2	**0.917**	**0.91** [0.90, 0.92]	**7.62**	**0.04**	**7.62**
PPG only	0.842	0.81 [0.73, 0.86]	11.8	4.56	10.8
[16], MESA	0.87	0.88 [0.86, 0.90]	9.67	-0.58	9.66
[16], All datasets	0.89	0.91 [0.89, 0.92]	8.88	-0.85	8.84

Table 3.4: AHI-level metrics for our work with and without SpO2 compared to results from Fonseca et al. Both Spearman's ρ and the ICC are statistically significant with $p < 0.0001$.

Severity-class-level performance

Figure 3.8 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 in both models, the bias towards predicting higher severity classes for the PPG only model is still visible.

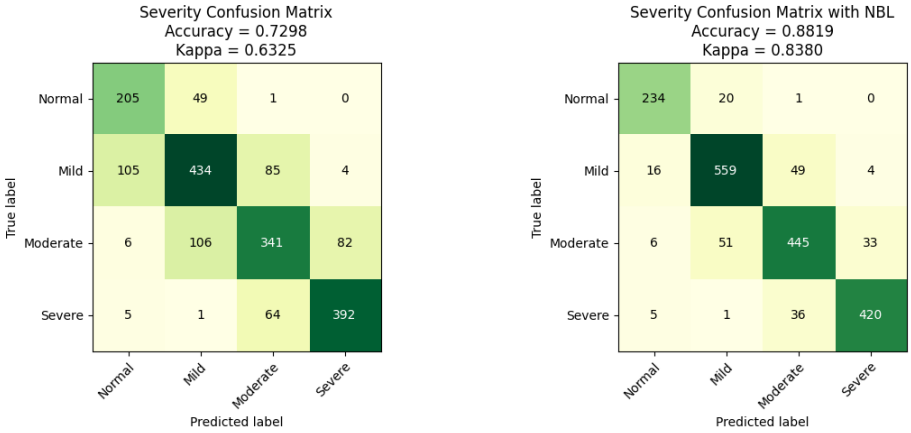
We show the models discrimination ability in Table TODO: ref, where we show binarized confusion matric results for no SDB vs SDB, mild vs moderate SDB, and moderate vs severe SDB. As before, the model with access to SpO2 performed better and NBL increased results. Most PPG + SpO2 model metrics exceed 90% while the PPG only model struggles especially with specificity and NPV. Likelihood ratios are also shown, achieving values of ≥ 4 and ≤ 0.2 for positive and negative likelihood ratios respectively in most cases, showing good diagnostic performance.

3.4 Importance of correct Sleep Stages

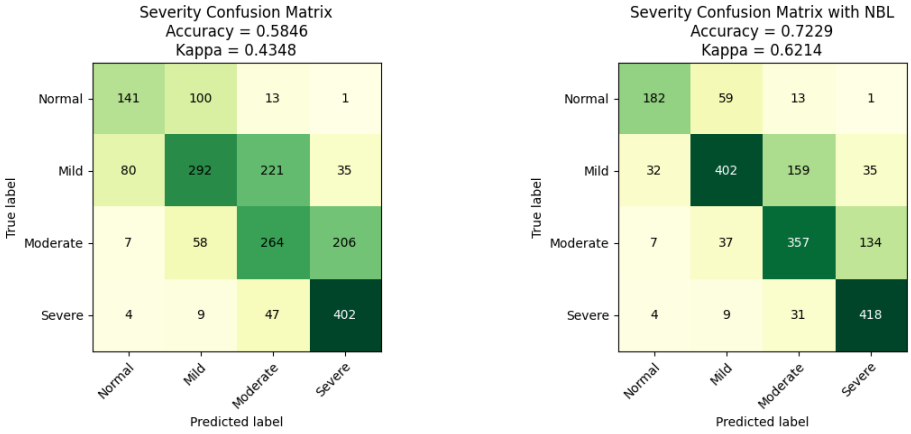
To assess the importance of the correct sleep stage prediction, we trained a model with the ground-truth hypnogram, the PPG-generated hypnogram, and finally without any sleep stage information, letting the model only rely on PPG and SpO2. Figure 3.9 shows the event-level metrics for each of these experiments. With the exception of the recall for lower thresholds, the models performance reduces consistently, the less certain it is on sleep stages. Peak F1-score for the model without a hypnogram was only 56.9%, a 13% drop from the version with the PPG-predicted hypnogram and a 20% drop from the model that has access to the ground-truth sleep stages, which got up to 76.1%.

3.5 Correcting model output

After applying the threshold for the prediction, a correction step was applied. This step removed events shorter than a specified number of seconds (called the correction size) and merged events that were closer than the correction size. Figure 3.10 displays the impact of the correction 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. A correction size of 3 seconds seems to be the best option.



(a) Model with SpO2 and PPG



(b) Model with PPG only

Figure 3.8: Confusion matrices for the predicted and true severity classes with and without NBL and for both models. (a) shows the model with SpO2 and PPG, while (b) shows the model with PPG only.

Min. Sev.	N ≥ thr (% ≥ thr)	Acc.	Sens.	Spec.	PPV	NPV	LR+	LR-
			PPG + SpO2					
Mild	1625 (86%)	0.912	0.929	0.804	0.968	0.639	4.736	0.089
Moderate	997 (53%)	0.889	0.882	0.898	0.907	0.870	8.650	0.132
Severe	462 (25%)	0.917	0.848	0.939	0.820	0.950	13.990	0.161
			PPG + SpO2 (NBL)					
Mild	1625 (86%)	**0.974**	**0.983**	0.918	**0.987**	0.897	11.941	**0.018**
Moderate	997 (53%)	0.938	0.937	0.939	0.945	0.929	15.319	0.067
Severe	462 (25%)	0.958	0.909	**0.974**	0.919	**0.970**	**34.840**	0.093
			PPG only					
Mild	1625 (86%)	0.891	0.944	0.553	0.931	0.608	2.112	0.101
Moderate	997 (53%)	0.815	0.922	0.694	0.773	0.887	3.015	0.113
Severe	462 (25%)	0.839	0.870	0.829	0.624	0.951	5.099	0.157
			PPG only (NBL)					
Mild	1625 (86%)	0.938	0.974	0.714	0.956	0.809	3.401	0.037
Moderate	997 (53%)	0.859	0.943	0.764	0.819	0.922	4.002	0.075
Severe	462 (25%)	0.886	0.905	0.880	0.711	0.966	7.547	0.108

Table 3.5: Diagnostic performance of our models. The best values all come from the PPG + SpO2 model with NBL.

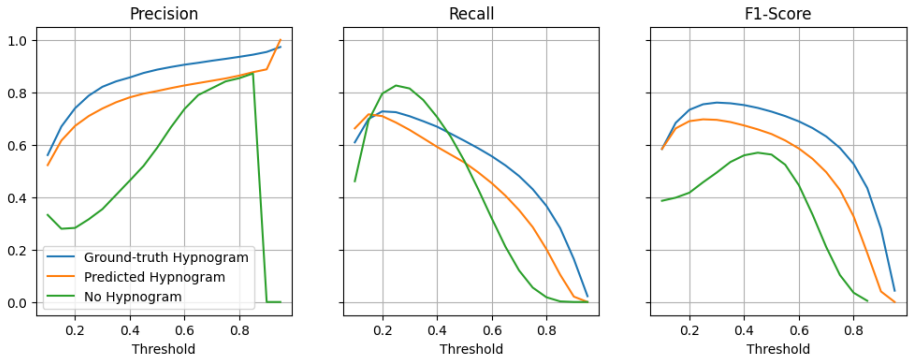


Figure 3.9: Precision, recall, and F1-score for the SDB detection model with different sleep stage information. The more the models sleep stage information gets to the ground-truth hypnogram, the better the performance.

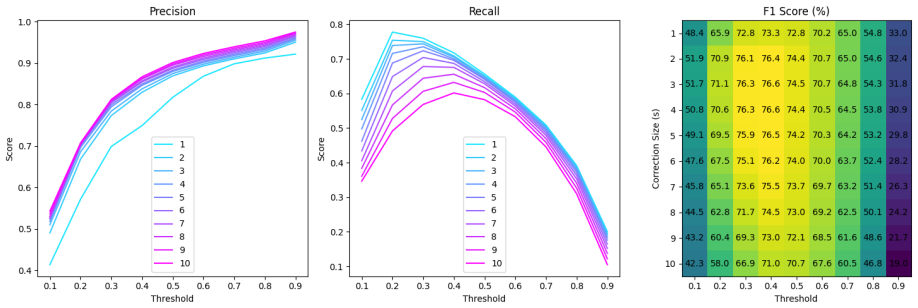


Figure 3.10: Event-level metrics for our SDB detection model for different correction sizes over the thresholds. As the precision grows with bigger correction sizes, the recall decreases. While most of both changes are somewhat evenly, there is a big difference in no correction (size of 1) to a small correction (of size 2) for the precision. Important to note is that, as with the preprocessing experiments, these results came from tests with the ground-truth hypnogram.

Chapter 4

Discussion

In this work, we presented an automatic SDB detection model based on an Attention U-Net using as input only PPG and optionally SpO2 signals. We achieved a peak F1-score of 69.7% in event detection and an AHI prediction correlation of $\rho = 0.917$. We showed great diagnostic results with positive and negative likelihood ratios of ≥ 11.0 and ≤ 0.1 respectively with very few participants ($\leq 1\%$) being wrongly classified more than one severity class apart. Event detection metrics were based on a strict event scoring that is more transparent than minute-to-minute, segment classification, as used in other studies.

Our model demonstrated higher detection rates for apnea events compared to hypopnea events. The much larger prevalence of hypopneas than any other event type in our dataset would suggest, that the classifier had more examples to learn from and could yield higher performance for these. However, this result is expected from a physiological perspective because hypopneas, which are associated with partial instead of full obstructions, have a much lower expression in cardiorespiratory changes in the PPG signal, and are not always associated with desaturation events, and thus have no expresion in the SpO2 signal.

One goal of our work was to use only PPG and SpO2, as the finger-worn sensor used to record these signals, is easy to set up and relatively unobtrusive during sleep. However, even less obtrusive options are smart watches or smart rings, that already today can record PPG with good quality. To the best of our knowledge, SpO2 cannot yet be realiably recorded using these devices, especially not the subtle drops in saturation of as little as 3% based on which respiratory events are scored. We showed that omitting SpO2 data from the training signals decreased performance, but not by a drastic amount. The model was still able to detect SDB events with a peak F1-score of 61.6% and predict AHI with a correlation of $\rho = 0.842$. This means that our model is still useful using these even less obtrusive technologies that only measure PPG, and can be used over many

nights. This may enable long-term home monitoring of SDB, enable accurate screening, help evaluate night-to-night variability, and evaluating effectiveness of SDB treatment methods.

Another important determinant of performance for our model is the source of sleep stage labels. The model performed best when using sleep stages scored with Somnolyzer, and worst without any sleep stage information. The PPG-derived sleep stages from the algorithm described by Bakker et al. [22] greatly improved results over using no sleep stage information but were still not as good as using the ground truth. As described in TODO: appendix, the algorithm can reach higher performance when besides cardiac information, also respiratory information is available. Further improvements in predicting the hypnogram from only PPG signals could lead to better results in detecting SDB, that still rely solely on data from PPG sensors. Although surrogate measures for predicting the hypnogram will likely never reach the same quality of the full PSG-derived sleep stages, our results indicate that even imperfect sleep stage information is well-suited for this task.

We also evaluated several approaches for preprocessing the PPG signal, namely using statistical analysis and a VAE. While achieving the same level of performance 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, we corrected the model output by filtering out events shorter than 3 seconds and merging events less than 3 seconds apart into one, which yielded better results than not correcting the output at all. Analyzing the precision and recall seperately, we found that increasing this correction size beyond 3 seconds, results in great decreases in recall, while the precision just improved slightly, resulting in a worse overall F1-score. The opposite happened when using correction sizes smaller than 3, where the recall improvements could not compensate for the big drop in precision. This big drop is likely due to single outliars for one or two seconds, which we can get rid of by selecting a correction size of 3 seconds.

4.1 Limitations and Future Work

An important limitation of our work is the lack of validation on other, external datasets. While the MESA dataset used in this work is large and greatly balanced in some cohort statistics, like AHI, BMI, smoking habits, or presence of 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 addressed. 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, especially with the PPG only model. 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.

Also, we 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. As the F1-score and losses did not seem to reach their peak in the 30 epochs we trained the model for, further work could look into training for longer, or using other preprocessing techniques for the SpO2 signal.

As the use of smart watches or smart rings maximizes or goal of unobtrusive 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 can be easily monitored with an associated smartphone, an which can be used for detecting snoring. Both signals are indicators of SDB and might improve detection performance 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. TODO: 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. TODO: summery: importance of correct sleep stage

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] 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.

[7] 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.

[8] 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.

[9] 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 limitations 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.

[10] 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.

[11] Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

[12] 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.

[13] 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.

[14] Merel M van Gilst, Johannes P van Dijk, Roy Krijn, Bertram Hoondert, Pedro Fonseca, Ruud JG van Sloun, Bruno Arsenali, Nele Vandenbussche, Sigrid Pillen, Henning Maass, et al. Protocol of the somnia project: an observational study to create a neurophysiological database for advanced clinical sleep monitoring. *BMJ open*, 9(11):e030996, 2019.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

Appendix A

Appendix

Todos:

- should I show an example of the models output
- 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
- don't show the kfold balancing analysis, right? table was enough
- Evaluation per Sleep Stage, and Event lengths