POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Politecnico di Torino

ETH zürich

Master's Degree Thesis

Sleep Stages Classification in Sleep Disorder Patients: Integrating Wearable and Contactless Commercial Devices

Supervisors

Prof. Gabriella OLMO

Prof. Robert RIENER

M.D. Markus SCHMIDT

M.Sc. Oriella GNARRA

Candidate

Stefano GIODA

October 2023

Abstract

Sleep disorders decrease the quality of sleep for affected individuals, potentially leading to serious, negative health effects. Therefore, it is essential to promptly diagnose these disorders and subsequently monitor their progression. The diagnosis of sleep disorders involves the examination of sleep, categorized into distinct stages, with polysomnography (PSG) currently considered the gold standard for assessment. However, PSG has limitations: it is expensive, time-consuming, complicated to operate, obtrusive, and usually only performed on a single night. One possible solution to these limitations is to leverage commercially available wearable and contactless devices that are already capable of providing sleep stages classification. These devices are affordable, easy to use, comfortable, and suitable for multiple nights of use.

To investigate this alternative, this study analyzes the data collected from patients with sleep disorders to whom two wearable devices (Fitbit Inspire 2 and Empatica E4) and two contactless devices (Somnofy and Emfit) were added during PSG. The initial step consisted of an in-depth evaluation of the sleep stages automatically provided by Somnofy, Fitbit, and Emfit in comparison to the PSG for this particular type of patients. The devices demonstrated an overall accuracy of 67% for Somnofy, 64% for Fitbit, and 47% for Emfit. Statistically significant differences were found in all sleep measures, such as total sleep time and REM latency, with particular difficulty in detecting cases of very short sleep stages durations.

The dataset was then used to fine-tune some models using signals from Empatica to classify sleep stages. The results aligned with those of the other devices, performing better than Fitbit and worse than Somnofy.

Lastly, a novel approach to sleep stages classification was proposed: fusing sleep stages from multiple devices. A random forest was trained to classify the sleep stage of an epoch based on the sleep stages predicted by devices at that epoch. Sleep stages from Somnofy, Fitbit, Emfit, and the fine-tuned Empatica model were incorporated, and all possible combinations of two to four devices were tested. This method achieved the highest accuracy of 73% when fusing Somnofy, Fitbit, and Empatica. It was generally more accurate than the devices used alone, particularly when combining three or four devices.

In conclusion, this study demonstrates the potential of using commercially available devices for sleep stages classification. Encouraging results were achieved through the integration of multiple devices. Despite some limitations, these devices represent a promising path toward more comprehensive and accessible sleep monitoring for both healthy individuals and patients with sleep disorders.

Table of Contents

List of Tables

List of Figures

Acronyms

AP-M Analysis Period Manual

AP-A Analysis Period Automatic

BVP Blood Volume Pulse

EBE Epoch-By-Epoch

EDA Electrodermal Activity

EEG Electroencephalography

EMG Electromyography

EOG Electrooculography

GELU Gaussian Error Linear Unit

HR Heart Rate

HRV Heart Rate Variability

IBI Interbeat Interval

MCC Matthews Correlation Coefficient

NREM Non-Rapid Eye Movement

PPG Photoplethysmography

PSG Polysomnography

REM Rapid Eye Movement

WASO Wake After Sleep Onset

Chapter 1

Introduction

Sleep is crucial since it can impact several aspects of our lives, including health and cognitive performance. Numerous sleep disorders can interfere with sleep and, subsequently, detrimentally affect people's lives. Therefore, it is vital to monitor sleep in order to detect disorders promptly and treat them appropriately.

Sleep is analyzed by dividing it into different stages, typically defined according to the electrical activity of the brain as measured by electroencephalography (EEG) during polysomnography (PSG), which is the gold standard for sleep monitoring. However, PSG is an expensive, time-consuming, and obtrusive method that is typically only conducted for a single night.

As an alternative, various commercially available wearable and contactless devices with different sensors, such as accelerometers and photoplethysmography (PPG), are capable of estimating sleep stages. Nevertheless, the accuracy of these devices varies widely, and it is essential to verify and improve their performance compared to PSG, especially in patients with sleep disorders.

1.1 Sleep

Sleep is a fundamental physiological process that is essential for our health and well-being. During sleep, our bodies undergo a complex series of physiological and neurological changes that allow us to recover, and regenerate. Sleep plays a crucial role in many important aspects of our lives such as memory consolidation, learning, and cognitive functions, as well as in the regulation of mood, appetite, and immune function.

Sleep can be divided into several stages that were defined mainly based on the signals from EEG, which measures the electrical activity of the brain, but also electrooculography (EOG), detecting eye movements, and electromyography (EMG), measuring the electrical activity of muscles. There are two main types

of sleep stages: rapid eye movement (REM) sleep and non-rapid eye movement (NREM) sleep. According to the American Academy of Sleep Medicine guidelines, NREM sleep can be further divided into three stages: stage N1, N2, and N3. Stage N1 is the lightest stage of sleep, during which we may experience occasional muscle twitches and drifting thoughts. Stage N2 is a deeper stage of sleep characterized by slower brain waves and occasional bursts of rapid brain activity called sleep spindles. Stage N3 is the deepest stage of sleep, also known as slow-wave sleep, during which our brain activity slows down significantly and we experience minimal muscle tone or movement.

It has become common, especially with the emergence of commercial devices that can classify sleep stages, to group the N1 and N2 stages and call it light sleep, while the N3 state is called deep sleep. This division into 4 sleep stages (wake, REM, light sleep, and deep sleep) was used throughout the project.

1.1.1 Sleep disorders

Sleep disorders can affect the quality of sleep, leading to adverse health effects.

According to the International Classification of Sleep Disorders [1], sleep disorders can be divided into the following categories:

- Sleep-related breathing disorders: they manifest as deviations in respiratory patterns during sleep;
- Central disorders of hypersomnolence: they are characterized by excessive daytime drowsiness, unrelated to disrupted nighttime sleep or irregular circadian rhythms;
- Insomnia: it is characterized by ongoing struggles in starting, maintaining, or achieving satisfactory sleep, even when given sufficient opportunity and favorable conditions for rest;
- Parasomnias: it encompasses irregular sleep-related complex motions, actions, sentiments, perceptions, dreams, and activity of the autonomic nervous system;
- Circadian rhythm sleep-wake disorders: they arise due to changes in the circadian timing system, its synchronization mechanisms, or a mismatch between the internal circadian rhythm and the external surroundings;
- Sleep-related movement disorders: they are predominantly marked by relatively uncomplicated, commonly repetitive movements that disrupt sleep or its initiation;
- Other sleep disorders.

1.2 Polysomnography

Polysomnography (PSG) is a non-invasive diagnostic tool used to monitor and record various physiological parameters during sleep. It is considered the gold standard for sleep monitoring and involves the recording of brain waves, eye movements, muscle activity, heart rate, oxygen saturation, and respiratory effort.

During PSG, the patient is typically required to spend a night in a sleep laboratory, where they are monitored by a technician or sleep specialist. Electrodes and sensors are attached to the patient's scalp, face, chest, legs, and fingers, and the data are continuously recorded throughout the night. The recorded data are then analyzed by a trained sleep specialist, who can identify the different sleep stages based on the EEG, EOG, and EMG signals.

PSG is an important tool for diagnosing sleep disorders. However, it is uncomfortable to wear, expensive, time-consuming, and requires specialized equipment and expertise to perform it and analyze the results. Furthermore, the classification of sleep stages for certain epochs can vary among physicians when presented with the same data. The inter-rater agreement among different doctors has been estimated to have a Cohen's kappa of 63% [2], while the accuracy, or the percentage of times that different doctors classify the same sleep stage, has been reported to be 88% [3].

Therefore, there is a need for alternative sleep monitoring methods that are more practical and accessible, while still providing accurate and reliable information.

1.3 Commercial devices

Sleep monitoring devices have become increasingly popular for home-based monitoring of sleep quality and quantity. These devices can provide valuable information, such as sleep duration and sleep stages, which can help individuals optimize their sleep habits and improve their overall health and well-being. Commercial sleep monitoring devices typically use various sensors, such as accelerometers and PPG, to monitor different aspects of sleep. However, the accuracy and reliability of these devices in measuring sleep parameters vary widely, and there is a need to assess their performance against the gold standard for sleep monitoring (PSG), especially for patients with sleep disorders, before they can be used for medical purposes.

The current study evaluated four commercially available devices:

- Somnofy;
- Fitbit Inspire 2;
- Emfit;

- Empatica E4.

Somnofy and Emfit are contactless devices, while Fitbit and Empatica are wearable wristband devices. Somnofy, Emfit, and Fitbit devices automatically classify sleep stages.

Validation studies for the sleep stages provided by these devices exist in the literature, but the majority are only conducted on healthy individuals, such as the study on Somnofy [3] and the study on Fitbit [4]. Some validations have been performed on specific populations, such as paper [5], which tested Somnofy and Emfit, among two other devices, on a group of older individuals dealing with mild sleep disturbances, mainly sleep apnea. Another paper [6] examined Fitbit Charge 2, a different model from the one used in this study, in shift workers. Still, none of these studies involved individuals with diagnosed sleep disorders, as was done in the validation paper of Emfit [7], where 70% of patients experienced sleep disorders, specifically sleep-related breathing disorders and sleep-related movement disorders. However, there remains a gap in validation that includes individuals with all types of sleep disorders. This study aims to address this gap.

1.4 Study setup

Each patient underwent overnight polysomnography in the Sleep-Wake Epilepsy Center (SWEZ), Department of Neurology, at the Insel, University Hospital, in Bern (Switzerland). There, a team of qualified professionals ensured that the sensors were worn correctly and checked that everything was in order during the night.

Patients were also wearing the two wristbands (Fitbit Inspire 2 and Empatica E4), while the two contactless devices were positioned in the room: Emfit was placed under the mattress at chest level, and Somnofy was located above the bed at the foot, facing the headboard. Not every patient had all four commercial devices and some devices failed during the night.

The PSG data were then analyzed by sleep specialists to manually review each patient's sleep stages, supported by the RemLogic and SOMNOmedics PSG devices which provide automated scoring.

The patients in this study are affected by some sleep disorders with the exception of a small percentage of healthy patients.

1.5 Objectives

The goals of this work can be summarized as:

- Evaluating of the sleep stages given by the commercial devices with respect to the PSG;
- Building a custom model that classifies sleep stages starting from the raw signals of the Empatica E4;
- Building a custom model that fuses the sleep stages of different devices.

The assessment of the device performance is important because, as mentioned above, validation studies with patients suffering from different sleep disorders are lacking, and at the same time it is fundamental before these devices can be used in sleep clinics. The simultaneous testing of multiple devices on the same patients is an added value, as it allows for a direct comparison of the results obtained.

The second objective is to fine-tune a model for this specific population. Empatica was chosen because it provides high-frequency data. It does not offer sleep stages classification, and only one study in the literature has investigated sleep stages classification using Empatica E4 [8]. In this work, the network from [9], developed for sleep stages classification with another wearable device, is adapted for use with Empatica E4, as it is a more flexible network than the one used in [8]. The results can then be compared with those obtained from the other devices in the first step.

Finally, a novel approach to sleep stages classification is introduced. This approach attempts to improve performance by using multiple devices simultaneously and fusing the sleep stages classified by each of them.

Chapter 2

Methods

2.1 Commercial devices

The devices included in the study are:

- Somnofy;
- Fitbit Inspire 2;
- Emfit;
- Empatica E4.

They are all consumer devices except the Empatica E4, which is a research device.

Within this clinical study, the aim of my thesis was to analyze the data from the given dataset and note that I was not involved in the selection of the devices.

2.1.1 Somnofy

Somnofy is a contactless device that uses radar sensors to detect movement, breathing, and heart rate. It also collects data about the environment, such as light intensity, noise level, and room temperature.

The raw sensors data are accessible at a frequency of 1 Hz, while the other derived measurements, such as sleep stages, have a resolution of 30 seconds.

The Somnofy device is connected directly to the power plug, so it does not need to be charged.

2.1.2 Fitbit Inspire 2

Fitbit Inspire 2 is a wrist-worn device developed by Fitbit Inc. It is a fitness tracker that can monitor the heart rate (HR) and the activity, but also the sleep, through an accelerometer and an optical sensor.

The device automatically detects sleeping periods and provides a breakdown of sleep stages. In our dataset, sleep stages are represented by the times at which a stage change occurs (e.g. from light sleep to deep sleep) and it is calculated at 30-second intervals so that the change can occur at the beginning or in the middle of each minute.

The producer claims up to 10 days of battery life.

2.1.3 Emfit

Emfit is a contactless device placed under the mattress that uses piezoelectric sensors to monitor movement and breathing patterns as well as heart rate using the ballistocardiograph signal.

From these measurements, it calculates other metrics such as heart rate variability (HRV) and sleep stages.

Heart rate, respiration rate, and movement activity are provided every 4 seconds, and sleep stages every 30 seconds.

It also connects directly to electricity.

2.1.4 Empatica E4

Empatica E4 is a wristband developed by Empatica Inc. and designed to collect several physiological data through different sensors:

- PPG sensor to measure blood volume pulse (BVP) at a sampling frequency of 64 Hz, from which the HRV and the interbeat interval (IBI) are derived;
- 3-axis accelerometer to detect activity in the range $[-2\,\mathrm{g}, 2\,\mathrm{g}]$ at a sampling frequency of 32 Hz;
- Infrared thermopile to measure the skin temperature with the following specifics:
  - sampling frequency: 4 Hz
  - range: skin temperature from $-40\,^{\circ}\mathrm{C}$ to $115\,^{\circ}\mathrm{C}$
  - sensitivity: $0.02\,^{\circ}\mathrm{C}$
  - accuracy: $\pm 0.2\,^{\circ}\mathrm{C}$ in the range $36\,^{\circ}\mathrm{C}$ to $39\,^{\circ}\mathrm{C}$

- Electrodermal activity (EDA) sensor to detect changes in sweat gland activity at a sampling frequency of 4 Hz.

The device only provides the data described above and has no built-in algorithm for predicting sleep stages.

The battery can last more than 32 hours.

2.2 Evaluation sleep stages of commercial devices

To comprehensively evaluate the performance of the sleep stages provided by Somnofy, Fitbit, and Emfit compared to the PSG sleep stages, several analyses were performed based on previous validation studies, namely:

- Analysis period: compare the duration of the estimated period of sleep;
- Sleep summary measures: compare metrics that summarize all-night sleep, such as the total duration of each sleep stage;
- Epoch-by-epoch (EBE) concordance: compare sleep stages at the same epoch;
- Distribution of sleep stages consecutive durations: compare the distributions of duration of successive sleep stages for each different sleep stage.

All devices utilize the following division of sleep stages: wake, REM, light sleep, and deep sleep. However, the PSG uses a different set of stages: wake, REM, N1, N2, and N3. In order to make the stages consistent, the N1 and N2 stages of the PSG were converted to light sleep, while the N3 stage was treated as deep sleep. Both the devices and PSG classify sleep stages in 30-second epochs.

2.2.1 Analysis period

The devices automatically detect when a person begins sleeping and classify the sleep stages until the end of sleeping is detected. This start and end detection determines the period during which the sleep stages are provided, referred to as the Analysis Period Automatic (AP-A) as it is automatically detected by the device.

During the PSG, the personnel manually recorded the time when the lights were switched off in the evening and the time when they were switched on in the morning. This period is named the Analysis Period Manual (AP-M) and is considered the ground truth for the analysis period.

In this assessment, the AP-A has been compared to the AP-M using a Bland-Altman plot [10].

2.2.2 Sleep summary measures

Some common measures are derived from the hypnogram and are important indicators to assess sleep, in particular, they are used to diagnose sleep disorders.

The measures analyzed were:

- Total sleep time: total amount of time spent in either REM, light, or deep sleep;
- Sleep onset latency: amount of time it takes to fall asleep after turning off the lights;
- REM latency: amount of time from the sleep onset to the first occurrence of REM sleep;
- Sleep efficiency: ratio between the total sleep time and the time spent in bed, measured as the total duration of the analysis period;
- Wake after sleep onset (WASO) duration: total amount of time spent awake after sleep onset;
- REM duration: total amount of time spent in a state of REM sleep;
- Light sleep duration: total amount of time spent in a state of light sleep;
- Deep sleep duration: total amount of time spent in a state of deep sleep.

All the measures were calculated on a single full night of a patient and were computed both for the manual analysis period (AP-M) and for the automatic analysis period (AP-A).

Bland-Altman plots were used to assess the degree of agreement between devices and PSG results.

In addition, t-tests were executed to determine if the differences were statistically significant: a paired t-test if the differences were normally distributed and a Wilcoxon signed-rank test if they were not. To verify normality, the Shapiro-Wilk test was used.

2.2.3 Epoch-by-epoch concordance

The EBE concordance analysis consists of comparing the sleep stage of each device with the PSG at every epoch. This is important in order to obtain a correct hypnogram that allows the assessment of sleep cycles and sleep fragmentation.

The epoch duration was 30 seconds both for the PSG and the devices. The analysis was performed in all the epochs where both the device and the PSG were present, which can be viewed as the intersection between the automatic and manual

analysis periods. Only the valid epochs were taken into account, while the epochs with artifacts in PSG or missing stage in the devices were discarded.

Different metrics can be computed by comparing the device's stage classification to the ground-truth classification of PSG and they can be calculated separately for each sleep stage or globally for all stages. By computing the metrics at a single sleep stage level, it becomes a binary classification where each epoch is either classified as the sleep stage considered (True) or classified as another sleep stage (False), which allows to define:

- True positives (TP): epochs where both the device and the PSG are the stage taken into account;
- True negatives (TN): epochs where both the device and the PSG are not the stage taken into account;
- False positives (FP): epochs where the device detects the considered stage, while the PSG does not;
- False negatives (FN): epochs where the PSG detects the considered stage, while the device does not.

The following metrics at single stage level were used:

- Sensitivity $= \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
- Specificity $= \dfrac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$
- Accuracy $= \dfrac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$
- Matthews correlation coefficient (MCC) $=$

$$\frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}}$$

The MCC was favored over Cohen's kappa, a similar correlation metric frequently used, due to its higher reliability, as Cohen's kappa exhibits some undesired behaviors [11].

The following global metrics were used:

- Balanced accuracy $=$

$$\frac{\mathrm{Sensitivity}_{\mathrm{wake}} + \mathrm{Sensitivity}_{\mathrm{REM}} + \mathrm{Sensitivity}_{\mathrm{light\ sleep}} + \mathrm{Sensitivity}_{\mathrm{deep\ sleep}}}{4}$$

  i.e. the balanced accuracy is the mean of the sensitivities for each sleep stage;

- Accuracy = $\frac{\text{total correct epochs}}{\text{total number of epochs}}$, where the correct epochs are all the epochs where the device and the PSG agree;
- Matthews correlation coefficient (MCC) =

$$\frac{s \times c - \sum_k^4 p_k \times t_k}{\sqrt{(s^2 - \sum_k^4 p_k^2) \times (s^2 - \sum_k^4 t_k^2)}}$$

  s = total number of epochs, c = total correct epochs, k = class from 1 to 4 (wake, REM, light, and deep sleep), p_k = total epochs classified as class k by the device, t_k = total epochs classified as class k by the PSG.

Sleep stages distribution was unbalanced, with most epochs being light sleep. For this reason, balanced accuracy should be preferred to accuracy, which is also reported only for the purpose of comparing results with those of other publications. The reported MCC formula represents the generalization in the multiclass scenario.

2.2.4 Distribution of sleep stages consecutive durations

This analysis, presented in the Fitbit Charge 2 validation study [6], compares the distribution of the consecutive durations of each sleep stage between the device and the PSG. A consecutive duration of a sleep stage was defined as the time interval (sum of adjacent epochs) in which the specific sleep stage remains uninterrupted, followed by a shift to another sleep stage. By identifying the consecutive durations for each patient, the total frequency for each possible duration of each sleep stage can be obtained and the distribution of consecutive durations for each sleep stage can be estimated.

The estimations of the distributions were plotted, allowing for a visual comparison between the devices and the PSG distributions. This comparison is essential for understanding the reliability of sleep fragmentation representation in the devices' hypnograms, pointing out whether they overestimate fragmentation, resulting in a distribution that is more concentrated in short durations, or whether they underestimate fragmentation, with a distribution concentrated on longer durations than PSG.

2.3 Empatica sleep stages

As mentioned before, Empatica E4 is able to collect several types of data, in particular, the data exported from it consisted of the following files:

- `ACC.csv`: accelerometer data for each of the three axes at a frequency of 32 Hz. Each value is in the range [-128, 128] which can be converted to the gravity of

Earth (g) by dividing the value by 64, obtaining the original range of [−2 g, 2 g];

- `BVP.csv`: blood volume pulse data derived from the PPG sensor at a frequency of 64 Hz;
- `EDA.csv`: electrodermal activity data expressed in Microsiemens at a frequency of 4 Hz;
- `HR.csv`: average heart rate values computed in spans of 10 seconds every second (1 Hz frequency) derived from BVP;
- `IBI.csv`: interbeat intervals, since each line contains the number of seconds from the previous beat, there is no predefined frequency. The incorrect peaks are removed automatically by the Empatica algorithm and they are reported in the file;
- `TEMP.csv`: skin temperature data in degrees Celsius at a frequency of 4 Hz.

Two different models were tested:

- Random forest classifier using features extracted from accelerometer, IBI and EDA signals;
- An encoder-decoder fully convolutional neural network.

2.3.1 Features extraction and random forest

The first approach to developing an algorithm for classifying sleep stages from raw Empatica data aimed to be very simple in order to have a baseline for more complex models. It consisted of extracting features from the signals and using them as input to a random forest classifier.

The features were extracted using FLIRT [12] which stands for Feature generation tooLkIt for weaRable daTa: it is a Python library that can extract features from accelerometer, IBI and EDA signals from wearable devices, and it supports Empatica E4.

The features are divided between:

- IBI: 52 features regarding heart rate and heart rate variability, it consists of statistical features, as well as features in time and frequency domain;
- EDA: 44 features extracted from time and frequency domains;
- Accelerometer: 22 features for each axis (x, y, z) plus 22 general features for a total of 88 features from time and frequency domains.

The detailed list of all the features can be found in [12].

Some features have been discarded either because they contained infinite values (EDA and accelerometer discarded features) or the features values were invalid (i.e. NaN values) for all rows (IBI features). The features discarded are:

- *hrv_lf_hf_ratio*: IBI feature for the ratio of low frequency to high frequency;
- *hrv_hfnu*: IBI feature for the high frequency power in normalized units;
- *hrv_lfnu*: IBI feature for the low frequency power in normalized units;
- *phasic_entropy*: EDA feature for the entropy of the phasic component of EDA, also called skin conductance response;
- *tonic_entropy*: EDA feature for the entropy of the tonic component of EDA;
- *x_entropy*: accelerometer feature for the entropy of the x-axis data;
- *y_entropy*: accelerometer feature for the entropy of the y-axis data;
- *z_entropy*: accelerometer feature for the entropy of the z-axis data.

By removing these features, the final number of features was 49 for IBI, 42 for EDA, and 85 for the accelerometer, which gives a total of 176 features.

The signals were divided into windows and the features were computed for each window separately. The FLIRT library allows to set both the window length and the step size to move the window. In line with the principle of having a simple model in this approach, the window length was set to 30 seconds which is the duration of the epoch, and also the step size was set to 30 seconds, in order to have non-overlapping windows for each epoch in which the features are based on data only of that epoch. Another simplification adopted was to discard all epochs in which there was any invalid feature value (i.e. NaN value).

The model was validated through a leave-one-group-out cross-validation approach, with each group consisting of a different patient. At each iteration, one patient was utilized as the test group while the remaining patients were used for model training until every patient had been tested.

2.3.2 Fully convolutional neural network

The previous approach is very dependent on the quality of the extracted features, which can be a major limitation.

To overcome this limitation, the model in the second approach received the data as input after minimal preprocessing. This is possible using deep learning networks that can learn to extract features and make predictions.

The architecture chosen is based on U-Sleep [13], a fully convolutional network based on U-Net [14], [15], which is a state-of-the-art algorithm for classifying sleep stages from EEG and EOG signals of PSG.

This model was selected because it had been previously tested in another wearable device [9], which made slight modifications from [13] and leveraged PPG and accelerometer signals, both of which are also present in Empatica. Furthermore, this model was preferred over the one previously tested on the Empatica device in [8] due to its utilization of state-of-the-art techniques as well as its flexibility to accommodate any type and number of input signals. The main contributions of this step include adapting the network presented in [9] to work with Empatica signals, experimenting with various signal combinations, identifying the optimal parameters, and conducting a comprehensive model evaluation similar to that performed on the other devices.

The data were divided into consecutive non-overlapping windows, each containing a fixed number of epochs, to ensure uniform input sizes. The number of epochs in a window is an arbitrary parameter, and experiments were conducted with different lengths to see which was most effective. The segment size is expressed in seconds as the multiplication of the number of epochs per window and the number of seconds per epoch, which is 30.

The network is shown in Figure 2.1 and can be divided into four parts:

1. Conformation module: it serves to prepare the input for the network by reshaping and zero padding it;

2. Encoder module: it extracts features from the data by compressing it to a lower temporal dimension;

3. Decoder module: it brings the compressed feature representation of the input back to the initial temporal dimension;

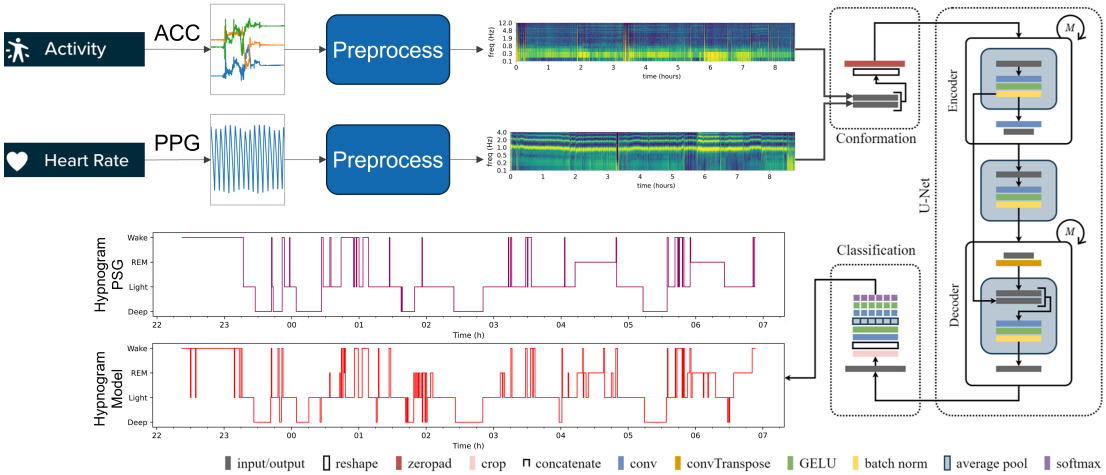
4. Segment classifier module: it segments and classifies the decoded vector into the single epochs.

The conformation module has been introduced in [9], while the other modules were already present in U-Sleep [13].

Each module is described in more detail below.

Conformation. The conformation handles concatenation, reshaping, and zero padding of the input. The network is able to receive multiple signals as input, and each signal is composed of three dimensions: the temporal dimension, which is the product of the segment size and the frequency of the signal; the spatial dimension, representing the number of data points for each time instant; and the channel dimension, since a signal can be composed of several channels. The first

Figure 2.1: Convolutional neural network for sleep staging classification with Empatica



Representation of the network, adapted from [9]. It shows an example of how 1024 epochs (equivalent to 8 hours and 32 minutes) are classified based on the activity and heart rate of the Empatica, but other combinations have also been tested. The raw activity signal comprises three separate signals from a 3-axis accelerometer, while the raw heart rate signal is the PPG, also called BVP. After minimal preprocessing, the raw signals can be directly fed into the model, or the spectrograms can be extracted and fed into the model. Both options have been tested, and this example presents the use of spectrograms. The two spectrograms are fused into a single array by the conformation module, which is then processed by M encoders followed by M decoders. The output is then divided and converted into 30-second epochs, which are classified into the 4 classes: wake, REM, light, and deep sleep. The hypnogram in red shows the classification of the model by taking at each epoch the class with the highest softmax output, while the hypnogram in purple shows the corresponding classification of the PSG which is the ground truth.
M = number of encoder and decoder blocks, GELU = Gaussian Error Linear Unit activation function [16], conv = convolution, convTranspose = transposed convolutional, batch norm = batch normalization [17], ACC = accelerometry, PPG = photoplethysmography.

operation of this module is to combine all input signals and reshape them into a 3-dimensional vector along the channel axis, which requires that all signals have identical temporal and spatial dimensions. The last step is to ensure that the first two dimensions are a power of 2 by adding zeros to them until the nearest power of 2 is reached, if necessary.

Encoder. The encoder module consists of a stack of M units. Each unit comprises a 2D convolution, a Gaussian Error Linear Unit (GELU) activation function [16], a batch normalization [17], and another 2D convolution, which reduces the temporal and spatial dimensions by half. The output of the last encoder unit undergoes a 2D convolution, a GELU activation function, and a batch normalization before

being fed to the decoder module.

Decoder. The decoder module, like the encoder, comprises a stack of M units. The unit has the peculiarity of receiving two inputs: one from the decoder at the previous level and one from the batch normalization output of the encoder at the same level of this decoder. The input from the previous decoder undergoes a transposed convolution to upsample it to the same dimension as the encoder output, allowing the next operation, which is the concatenation of the two. The concatenated array is then passed through a 2D convolution, a GELU activation function, and a batch normalization.

Segment classifier. In order to classify one sleep stage for each 30-second epoch, the array with the initial temporal size needs to be downsized. The steps are: removal of the zeros added during padding in the conformation modules, reshaping into a 2D array, a 1D convolution with GELU activation, average pooling in the temporal dimension to reduce it to the desired number of epochs, a 1D convolution with GELU activation, and a dense convolution with a softmax activation function. The final softmax function gives the probability of each stage for each epoch, and the stage with the maximum probability is the one classified by the model.

The architecture is very flexible and can be configured to accept any type of signal as input. Both direct input of the raw signals and input of the spectrogram of the signals have been tested.

Empatica signals were trimmed between lights off and lights on times in the original dataset. PPG and accelerometer data were preprocessed according to [9] and consisted of a z-score normalization (zero mean and unit standard deviation) for PPG and a median normalization to 0 for accelerometer. An adaptive interquartile range normalization on a 300-second sliding window was then applied to both signals, followed by clipping of outliers above 20 times the interquartile range. EDA, temperature, and HR signals were not included in the study [9]. After conducting experiments, it was found that performance did not improve with normalization for these signals compared to passing them directly without preprocessing. Therefore, no normalization was performed. When combining multiple signals for raw data input, they were resampled to 32 Hz, as done in [9], by downsampling (PPG) or upsampling (EDA, temperature, HR) to have the same temporal dimension.

2.4 Fusion of devices sleep stages

In this step, a novel approach to sleep stages classification is introduced: fusing together the sleep stages provided by different devices.

A random forest model was used to fuse the sleep stages because the data in input is very simple: there is only one sleep stage per device. The model is built to classify one epoch at a time, taking as input the sleep stage given by each device for that specific epoch. In addition to Somnofy, Fitbit, and Emfit, the sleep stages classification of the Empatica model is also used to allow the merging of all devices in the study. The Empatica model that obtains the best performance among all those previously proposed is used.

Performances were evaluated using leave-one-group-out cross-validation. This means that during each iteration, one patient is used for testing while the others are utilized for training, and the process is repeated until all patients have been tested.

Chapter 3

Results

3.1 Study population

The starting dataset consisted of 160 one-night PSG recordings of different patients as a result of the study conducted in the Sleep-Wake Epilepsy Center (SWEZ), Department of Neurology, at the Insel, University Hospital, in Bern (Switzerland) with Dr. Markus Schmidt M.D. Ph.D. as principal investigator and Prof. Tobias Nef as sponsor. The commercial devices (Somnofy, Fitbit, Emfit, and Empatica) were added to the usual PSG, resulting in the presence of data from these devices in the dataset. The number of patients with each device varied: 144 for Somnofy, 58 for Fitbit, 103 for Emfit, and 149 for Empatica, noting that not all patients had all devices. The so marked difference in the Fitbit is due to the fact that this device was introduced after the study had already begun. All data were subject to error checking, as described in the following section, which resulted in the exclusion of some patients for certain analyses. The demographic characteristics of the patients with valid data are then described.

3.1.1 Error checking

The sleep stages from PSG, Somnofy, Fitbit, Emfit, and Empatica raw data were checked for errors prior to analysis. Three patients were excluded from all subsequent analyses because they exhibited issues with their PSG data including sleep stages reported during the daytime rather than nighttime, a considerable discrepancy between the specified lights on time of 3 p.m. and the actual end of the PSG stages at 7 a.m., and sleep stages being reported twice for each epoch with conflicting epochs.

Somnofy is the only device that classifies some epochs as missing during the night, and for some patients, this classification represents a good part of the night, so 10 patients with more than 30% missing epochs were excluded. The device also

sometimes failed to detect sleep and classified most epochs as wake, so 5 patients with more than 90% of epochs classified as wake were discarded.

No errors were found in Fitbit stages, and only one patient was excluded from the Emfit analysis due to a classification lasting only 2 hours instead of the full night.

Regarding Empatica, 13 patients were removed because data were only available for less than 5 hours instead of a full night. In addition, 6 patients not among those removed had a malfunctioning temperature sensor that provided no data, so these patients were not excluded from all analyses because the other sensors worked, but when the temperature was used, these patients were discarded.

3.1.2 Demographics

The demographics of the entire study population after removing patients with invalid data and those whose data were correctly collected for each device are shown in Table 3.1.

Table 3.1: Study population characteristics

	All	Somnofy	Fitbit	Emfit	Empatica
Valid patients: n	157	128	56	101	133
Females: n (%)	86 (54.78)	74 (57.81)	29 (51.79)	56 (55.45)	73 (54.89)
Age (years): mean ± SD [range]	44.94 ± 16.20 [18.2, 84.6]	44.30 ± 16.40 [18.2, 84.6]	41.39 ± 16.02 [18.6, 79.1]	43.29 ± 15.59 [18.2, 75.3]	45.32 ± 16.23 [18.2, 84.6]
Diagnosis: %					
Breathing disorders	56.05	53.91	51.78	55.45	57.14
Hypersomnolence	17.84	17.97	25.00	19.80	15.79
Insomnia	5.09	5.47	1.79	4.95	6.02
Parasomnias	3.82	3.12	1.79	3.96	4.51
Circadian disorders	1.27	1.56	0	0.99	1.50
Movement disorders	1.91	2.35	1.79	1.98	2.26
Other disorders	3.19	3.12	3.57	0.99	3.01
Healthy controls	4.46	5.47	8.93	4.95	4.51
Missing diagnosis	6.37	7.03	5.35	6.93	5.26

SD = standard deviation

The device with the most valid patients is Empatica (133), followed by Somnofy (128), Emfit (101), and Fitbit (56).

In general, the proportion of women is close to half (54.78%), and the average age is about 45 years, ranging from a low of 18 years to a high of 84 years. All patients suffer from a sleep disorder except for a small percentage (about 5 percent) of healthy patients, and the majority of them suffer from sleep-related breathing disorders. There is also a portion of patients whose diagnosis is missing, typically

these are special cases in which it is not easy to diagnose the type of disorder and for which there was a need for a more in-depth examination.

27 patients have valid data for all four commercial devices. Some results are reported using only these patients so that the results of different devices on the same patients can be directly compared. The characteristics of this group are shown in Table 3.2.

Table 3.2: Characteristics of patients with valid data from all commercial devices

	Patients with all devices
Valid patients: n	27
Females: n (%)	14 (51.85)
Age (years): mean ± SD	42.24 ± 16.15
[range]	[19.0, 72.8]
Diagnosis: %	
Breathing disorders	51.86
Hypersomnolence	22.23
Parasomnias	3.70
Insomnia	3.70
Circadian disorders	0
Movement disorders	3.70
Other disorders	0
Healthy controls	11.11
Missing diagnosis	3.70

SD = standard deviation

3.2 Evaluation sleep stages of commercial devices

Results are reported separately for each analysis performed, which were:

- Analysis period;
- Sleep summary measures;
- Epoch-by-epoch concordance;
- Distribution of sleep stages consecutive durations.

3.2.1 Analysis period

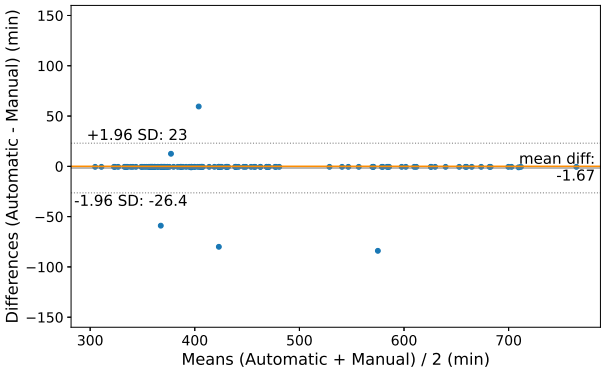
The Bland-Altman plots in Figure 3.1 provide a comprehensive visualization of the comparison between the durations of automatic analysis periods derived from the devices and the manual analysis periods obtained from PSG. These plots serve to illustrate the level of agreement between the two sets of durations. The y-axis of the plots represents the difference between the duration of the automatic period and the manual period, while the x-axis depicts the average of the two durations. The orange line, denoting the null difference, indicates cases in which the durations are perfectly in agreement. Points above the orange line indicate instances where the device has overestimated the duration of the sleeping period, while points below the line indicate an underestimation of the sleeping period by the device. Notably, the plots also include the average difference and the 95% limits of agreement.

It is important to note that within the original dataset provided by the study, Somnofy and Emfit data were already trimmed to include only the data within the lights off and lights on times, i.e. the manual period. Consequently, the automatic period for these devices never exceeds the manual period, except for a few instances where the lights off and on times were later corrected. Therefore, for these devices, the information we can get from the plots is only about the possible underestimation of sleep periods in cases where the automatic period either starts after the lights off time or ends before the lights on time.

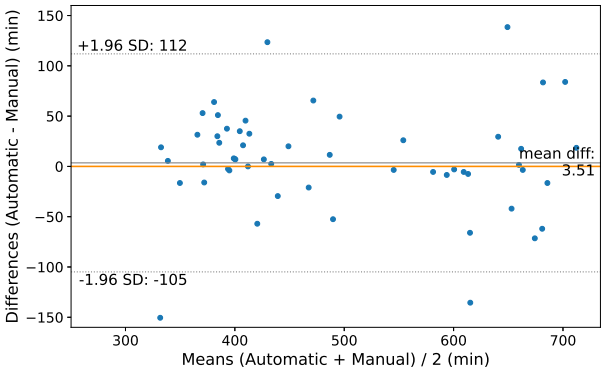
Given this context, Figure 3.1a demonstrates that Somnofy can recognize sleep throughout the entire period without any systematic underestimation. Differences are mostly concentrated around zero, with a few exceptions. This is further supported by the mean difference of less than 2 minutes and narrow limits of agreement. In contrast, Figure 3.1c indicates a clear tendency towards underestimation for Emfit. This trend is confirmed by a substantial mean difference of more than 20 minutes and wider limits of agreement, implying greater variability in underestimation.

From Figure 3.1b illustrating the Fitbit differences, it is evident that they are widely scattered both above and below the null difference line, as reflected in the limits of agreement, which are at more than 1 and a half hours. However, there appears to be negligible systematic bias as the mean difference results in an overestimation of 3 and a half minutes.

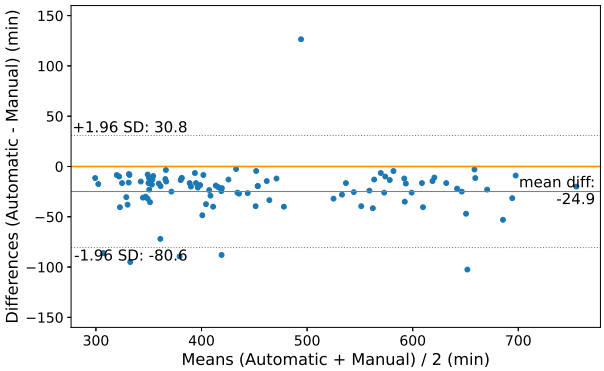
Figure 3.1: Bland-Altman plots analysis period of each device



(a) Somnofy



(b) Fitbit



(c) Emfit

Differences between the durations of the Automatic analysis periods and the durations of the Manual analysis periods are plotted for each patient. The orange center line shows the zero difference. The black solid line shows the mean of the differences, while the two dashed lines indicate the upper and lower limits of agreement at 95%. Reported values are in minutes.
min = minutes, SD = standard deviation

3.2.2 Sleep summary measures

The sleep measures results for each device are reported in separate tables (Somnofy: Table 3.3, Fitbit: Table 3.4, and Emfit: Table 3.5) as each device has different patients. For both the automatic and manual analysis periods, mean values are presented, along with the resulting P value from the t-test between the device and the PSG (paired t-test if the differences follow a normal distribution, otherwise Wilcoxon signed-rank test). Additionally, Bland-Altman plots are included for the individual measures during the manual period, as results with the automatic period are very similar and the data for Somnofy and Emfit were already cut for the manual period. The tables also show the differences that do not follow a normal distribution (Shapiro-Wilk test), which are the total sleep time, sleep onset latency, REM latency, sleep efficiency, and, WASO duration, which should be taken into account when reading the Bland-Altman plot of these measurements. Nonetheless, the Bland-Altman plots remain a useful visual tool for assessing agreement and bias between the two measurement methods.

The Somnofy results in Table 3.3 show that the sleep onset latency, REM latency (automatic period), light sleep duration, and deep duration had no statistically significant differences compared to PSG. Significant differences were found for the other parameters. The Bland-Altman plots (Figure 3.2) revealed remarkable proximity to zero in most cases, suggesting a degree of agreement between the automatic analysis of Somnofy and PSG data. However, the substantial dispersion in the differences, which in some instances extended towards notably larger values, is reflected in the wide agreement limits.

In Table 3.4 significant differences are reported across all sleep metrics when comparing Fitbit to PSG, except only the REM latency and sleep onset latency (automatic period). Moreover, the Bland-Altman plots (Figure 3.3) reveal differences in biases, with some metrics showing considerable biases, such as a notable underestimation of 33 minutes in WASO duration. In contrast, other metrics demonstrate biases that are extremely close to zero, with a difference of less than one minute in the case of REM latency. Nonetheless, it is essential to note a persistent trend observed across all metrics: a clear dispersion in the differences, as reflected in the broad limits of agreement. Furthermore, a pattern seems to emerge when considering sleep onset latency and WASO duration, suggesting that as the values of sleep onset latency and WASO duration increase, the underestimation also increases.

Looking at Table 3.5 with the Emfit results, all metrics exhibited statistically significant differences. Additionally, the Bland-Altman plots (Figure 3.4) indicate the presence of considerable biases in most cases and substantial scatter among data points, which implies a wide dispersion of the differences between Emfit and PSG. A clear trend exists in the metrics of sleep onset latency, which subsequently

affects REM latency and WASO duration. This trend is a consequence of the partial or complete lack of classification of wake epochs by Emfit. As Emfit fails to detect any episodes of wakefulness, sleep onset latency is mostly estimated as zero, assuming a linear trend with the value given by PSG, and the same happens with WASO duration.

3.2.3 Epoch-by-epoch concordance

The results of epoch-by-epoch concordance are first presented individually for each device (Somnofy: Table 3.6, Fitbit: Table 3.7, Emfit: Table 3.8) using all patients with valid data, followed by Table 3.9 with the results of each device on the 27 patients with valid data for all devices, enabling a direct comparison.

The results from Somnofy (Table 3.6) indicate comparable sensitivity across sleep stages and very good specificity for all stages except light sleep. Light sleep is the most common stage in PSG and is frequently classified by devices, leading to the highest number of misclassified epochs. This is reflected in the accuracy as well, with accuracies nearing 90%, whereas it is 70% for light sleep. MCC demonstrates a moderate correlation for both individual sleep stages and overall. Global balanced accuracy and accuracy are equal (67%) in this case, indicating consistent performance across sleep stages.

In Table 3.7, Fitbit exhibits considerably lower sensitivity for wake than for the other stages, indicating that Fitbit has difficulty detecting wake, and again lower specificity for light sleep, as well as accuracy. MCC scores are moderate for both individual sleep stages and the aggregate. Overall accuracy is 64%, with balanced accuracy lower at 60%, largely because Fitbit struggles to classify wake epochs.

Emfit (Table 3.8) has a sensitivity of less than 50% for REM and deep sleep and is particularly low for the wake stage, which Emfit also struggles to detect, confirming what was previously mentioned, that in many patients no awakenings are classified while they are present. Light sleep again has lower specificity and accuracy than the other sleep stages. It is important to note that MCC shows a weak correlation in all cases. Overall accuracy is quite low (47%) as well as balanced accuracy (40%), again lower mainly due to the wake stage.

For a direct comparison on the same data, Table 3.9 shows the performance of the devices in patients with valid data from all devices. Overall, Somnofy outperforms the other devices with consistently superior accuracy, MCC, and balanced accuracy metrics. Fitbit is ranked second, with the highest sensitivity for light sleep and the best specificity for deep sleep. Emfit performs considerably worse than the other two, with the only higher metric being wake specificity, however, this is due to the fact that this stage is hardly ever classified by the device.

3.2.4 Distribution of sleep stages consecutive durations

The distributions of consecutive sleep stages durations are shown in Figure 3.5 for the 27 patients with valid data from all devices because the distributions are very similar with all patients and the results can then be compared directly. For all sleep stages, the PSG's distribution is concentrated on short durations with the peak occurring during consecutive durations of 1-2 epochs (30 seconds-1 minute). This characteristic is particularly pronounced in WASO, whereas in REM there is a longer tail towards longer durations.

Somnofy is the device with the closest distributions to PSG, particularly in WASO and light sleep. In REM, the peak is observed to occur during the duration of three consecutive epochs, with no occurrences for shorter durations. For these patients, Somnofy is unable to identify durations lasting one to two epochs in REM. Additionally, the distribution of deep sleep shows a less pronounced peak on shorter epochs and a longer tail than PSG.

The Fitbit distributions deviate substantially from those of PSG, with all durations concentrated on longer time periods, as in the case of WASO, where the peak is at the 5-minute duration. In the distributions of REM and deep sleep, a discontinuity is observed with no durations less than 4 and a half minutes for REM and 3 and a half minutes for deep sleep. This nonbiological discontinuity, i.e. solely due to a device limitation, since biologically there can be shorter durations as shown in the PSG, is the same as that found in [6] on another Fitbit model, indicating that the algorithm used in these devices is similar and suffers from this limitation.

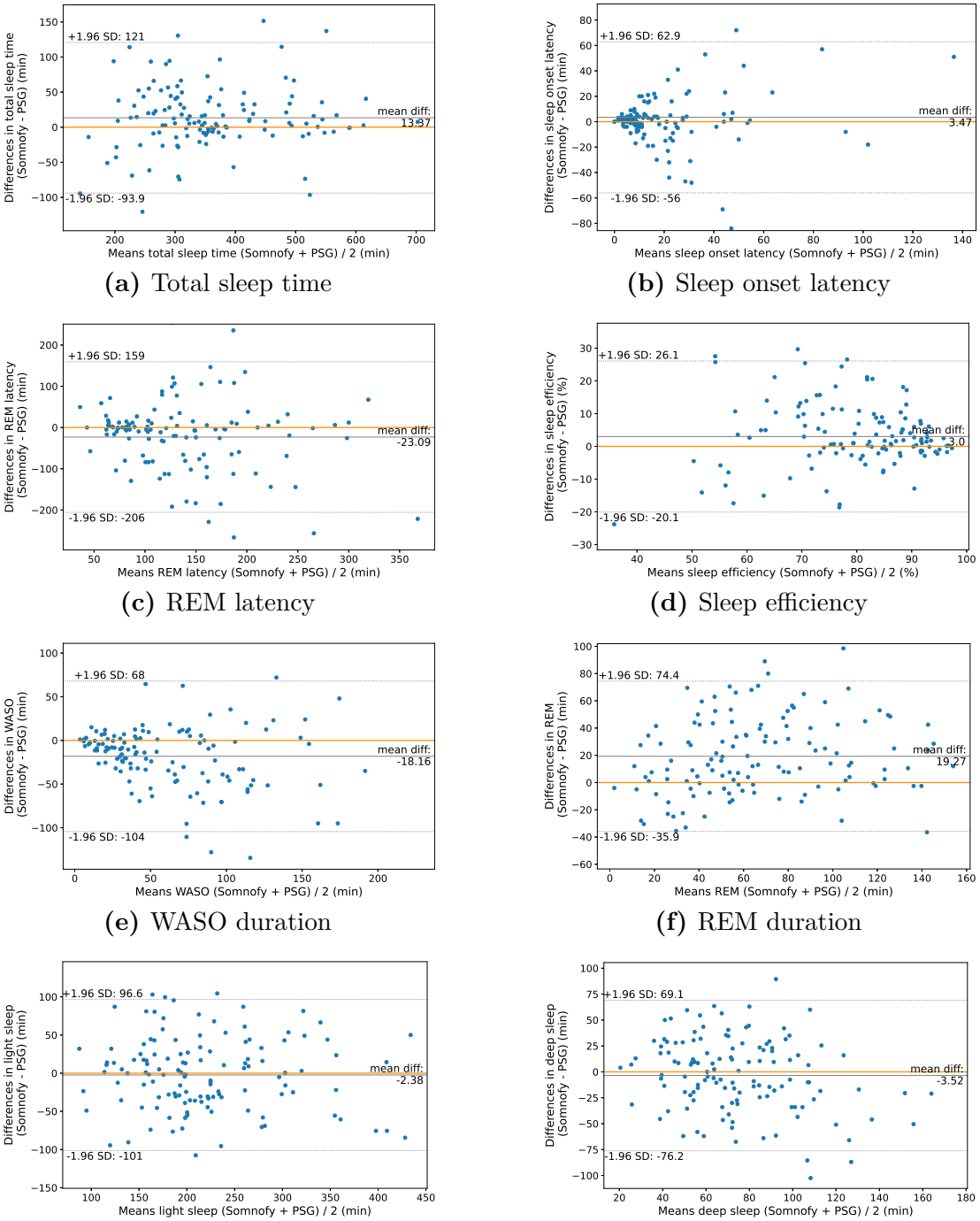
In Emfit distributions, these discontinuities are even more pronounced. No durations shorter than 4 and a half minutes exist in WASO, and in deep sleep, few durations are shorter than 9 minutes, with none shorter than 2 and a half minutes. Similarly, in light sleep, durations shorter than 8 consecutive minutes are very rare. Notably, the distributions' peaks are significantly shifted towards longer time intervals, with 13 minutes for deep sleep and 20 minutes for light sleep.

Table 3.3: Sleep measures Somnofy

Sleep measure		PSG	Somnofy-A	Somnofy-M
TST (min)	Mean ± SD	354.95 ± 112.81	368.61 ± 114.73	368.32 ± 114.80
	95% CI mean	[335.40, 374.49]	[348.73, 388.48]	[348.43, 388.20]
	P value t-test	-	<.001*	<.001*
SOL (min)	Mean ± SD	19.01 ± 21.82	20.01 ± 26.89	22.48 ± 33.55
	95% CI mean	[15.23, 22.79]	[15.35, 24.67]	[16.66, 28.29]
	P value t-test	-	.587*	.202*
REML (min)	Mean ± SD	151.44 ± 90.07	128.46 ± 72.67	126.27 ± 71.76
	95% CI mean	[135.84, 167.05]	[115.87, 141.05]	[113.84, 138.70]
	P value t-test	-	.052*	.019*
SE (%)	Mean ± SD	77.86 ± 13.45	81.19 ± 13.67	80.86 ± 13.82
	95% CI mean	[75.53, 80.19]	[78.82, 83.55]	[78.47, 83.26]
	P value t-test	-	<.001*	<.001*
$WASO_d$ (min)	Mean ± SD	72.99 ± 55.26	58.97 ± 47.39	54.82 ± 45.35
	95% CI mean	[63.42, 82.56]	[50.76, 67.18]	[46.97, 62.68]
	P value t-test	-	<.001*	<.001*
REM_d (min)	Mean ± SD	57.50 ± 36.25	76.77 ± 40.53	76.77 ± 40.54
	95% CI mean	[51.22, 63.78]	[69.75, 83.79]	[69.75, 83.79]
	P value t-test	-	<.001	<.001
$Light_d$ (min)	Mean ± SD	220.23 ± 79.27	218.14 ± 75.42	217.86 ± 75.42
	95% CI mean	[206.50, 233.96]	[205.07, 231.20]	[204.79, 230.92]
	P value t-test	-	.642	.597
$Deep_d$ (min)	Mean ± SD	77.21 ± 39.31	73.70 ± 28.89	73.69 ± 28.89
	95% CI mean	[70.41, 84.02]	[68.69, 78.70]	[68.69, 78.70]
	P value t-test	-	.287	.286

Underlined P values represent statistically significant differences.
Somnofy-A = Somnofy measures during automatic analysis period, Somnofy-M = Somnofy measures during manual analysis period, TST = total sleep time, SOL = sleep onset latency, REML = REM latency, SE = sleep efficiency, $WASO_d$ = WASO duration, REM_d = REM duration, $Light_d$ = light sleep duration, $Deep_d$ = deep sleep duration, min = minutes, SD = standard deviation, CI = confidence interval, * = differences not normally distributed

Figure 3.2: Bland-Altman plots for Somnofy sleep measures



(a) Total sleep time

(b) Sleep onset latency

(c) REM latency

(d) Sleep efficiency

(e) WASO duration

(f) REM duration

(g) Light sleep duration

(h) Deep sleep duration

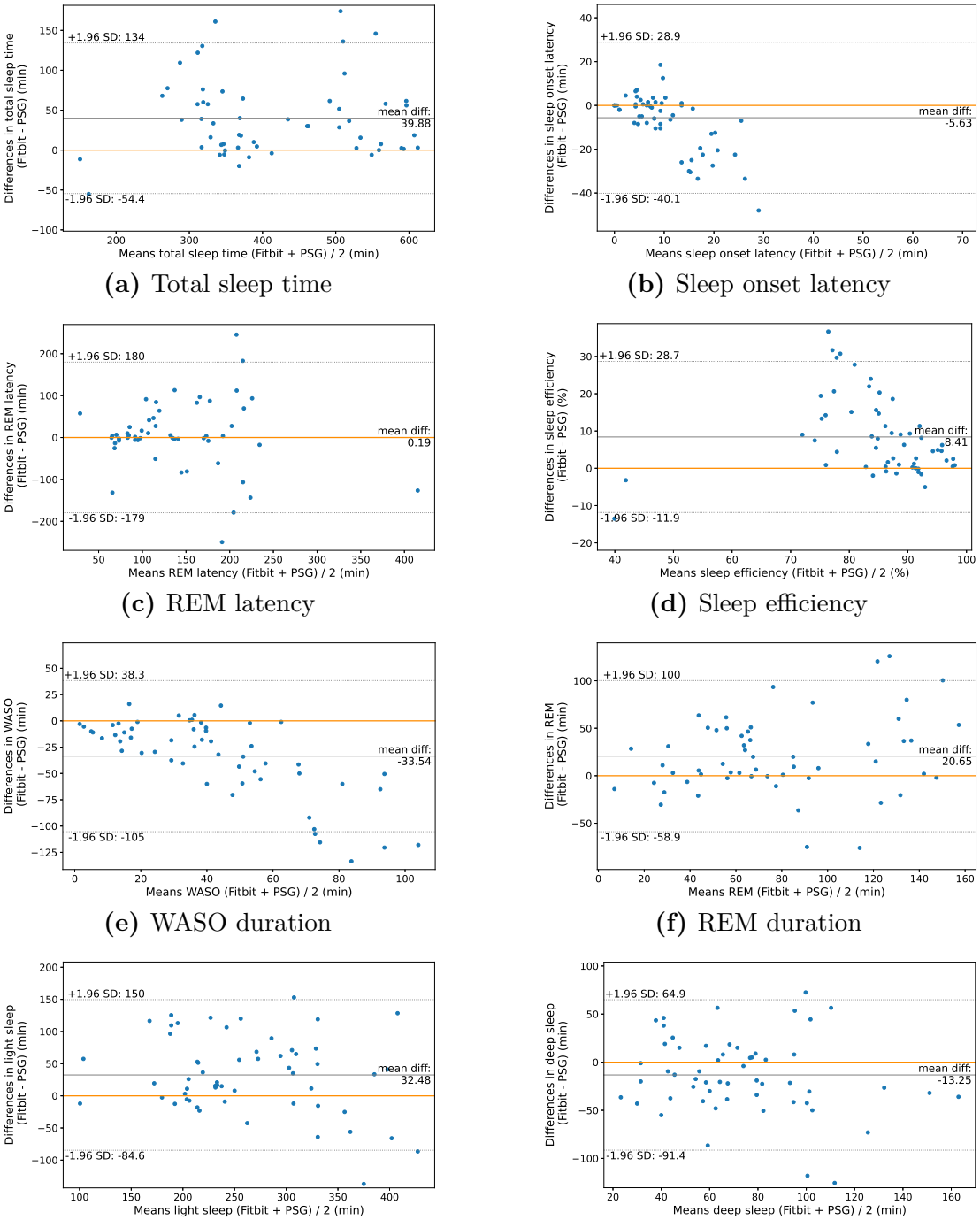
Differences between the sleep measures of Somnofy and PSG during the manual analysis period. min = minutes, SD = standard deviation

Table 3.4: Sleep measures Fitbit

Sleep measure		PSG	Fitbit-A	Fitbit-M
TST (min)	Mean ± SD	393.90 ± 118.19	444.84 ± 122.09	433.79 ± 120.17
	95% CI mean	[362.95, 424.86]	[412.86, 476.82]	[402.31, 465.26]
	P value t-test	-	<.001*	<.001*
SOL (min)	Mean ± SD	14.05 ± 12.64	13.41 ± 16.76	8.42 ± 14.95
	95% CI mean	[10.74, 17.36]	[9.02, 17.80]	[4.50, 12.34]
	P value t-test	-	.687*	<.001*
REML (min)	Mean ± SD	140.95 ± 84.27	148.68 ± 81.84	141.76 ± 75.34
	95% CI mean	[118.88, 163.03]	[127.25, 170.12]	[122.03, 161.50]
	P value t-test	-	.209*	.496*
SE (%)	Mean ± SD	80.43 ± 12.51	89.92 ± 6.02	88.84 ± 11.63
	95% CI mean	[77.15, 83.71]	[88.34, 91.49]	[85.79, 91.89]
	P value t-test	-	<.001*	<.001*
$WASO_d$ (min)	Mean ± SD	58.84 ± 41.79	29.71 ± 18.86	25.29 ± 17.43
	95% CI mean	[47.89, 69.78]	[20.73, 29.86]	[24.77, 34.66]
	P value t-test	-	<.001*	<.001*
REM_d (min)	Mean ± SD	67.62 ± 39.46	89.92 ± 47.66	88.28 ± 48.23
	95% CI mean	[57.29, 77.96]	[77.44, 102.40]	[75.64, 100.91]
	P value t-test	-	<.001	<.001
$Light_d$ (min)	Mean ± SD	246.73 ± 86.61	287.85 ± 76.16	279.21 ± 74.42
	95% CI mean	[224.05, 269.42]	[267.90, 307.80]	[259.72, 298.70]
	P value t-test	-	<.001	<.001
$Deep_d$ (min)	Mean ± SD	79.54 ± 39.29	67.07 ± 32.84	66.29 ± 32.93
	95% CI mean	[69.25, 89.83]	[58.47, 75.67]	[57.67, 74.92]
	P value t-test	-	.025	.017

Underlined P values represent statistically significant differences.
Fitbit-A = Fitbit measures during automatic analysis period, Fitbit-M = Fitbit measures during manual analysis period, TST = total sleep time, SOL = sleep onset latency, REML = REM latency, SE = sleep efficiency, $WASO_d$ = WASO duration, REM_d = REM duration, $Light_d$ = light sleep duration, $Deep_d$ = deep sleep duration, min = minutes, SD = standard deviation, CI = confidence interval, * = differences not normally distributed

Figure 3.3: Bland-Altman plots for Fitbit sleep measures



(a) Total sleep time

(b) Sleep onset latency

(c) REM latency

(d) Sleep efficiency

(e) WASO duration

(f) REM duration

(g) Light sleep duration

(h) Deep sleep duration

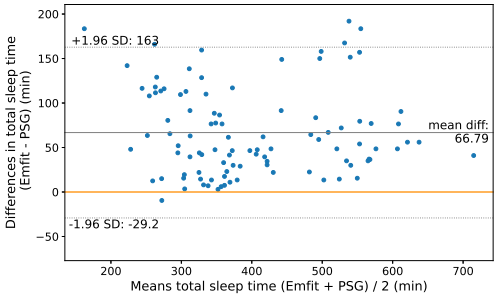
Differences between the sleep measures of Fitbit and PSG during the manual analysis period. min = minutes, SD = standard deviation

Table 3.5: Sleep measures Emfit

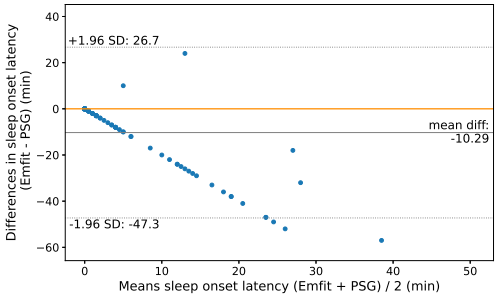
Sleep measure		PSG	Emfit-A	Emfit-M
TST (min)	Mean ± SD	367.50 ± 120.09	435.78 ± 118.73	434.29 ± 118.66
	95% CI mean	[344.08, 390.92]	[412.62, 458.93]	[411.15, 457.43]
	P value t-test	-	<.001*	<.001*
SOL (min)	Mean ± SD	11.03 ± 19.19	0.75 ± 3.55	0.74 ± 3.52
	95% CI mean	[7.29, 14.77]	[0.06, 1.44]	[0.06, 1.43]
	P value t-test	-	<.001*	<.001*
REML (min)	Mean ± SD	140.63 ± 86.99	47.57 ± 42.74	47.57 ± 42.74
	95% CI mean	[123.66, 157.59]	[39.23, 55.90]	[39.23, 55.90]
	P value t-test	-	<.001*	<.001*
SE (%)	Mean ± SD	77.09 ± 13.09	97.33 ± 3.49	91.75 ± 6.12
	95% CI mean	[74.54, 79.65]	[96.65, 98.01]	[90.56, 92.95]
	P value t-test	-	<.001*	<.001*
$WASO_d$ (min)	Mean ± SD	63.33 ± 42.47	10.78 ± 12.81	10.97 ± 12.55
	95% CI mean	[55.04, 71.61]	[8.28, 13.27]	[8.52, 13.41]
	P value t-test	-	<.001*	<.001*
REM_d (min)	Mean ± SD	62.40 ± 38.78	101.88 ± 34.79	101.70 ± 34.70
	95% CI mean	[54.83, 69.96]	[95.10, 108.67]	[94.94, 108.47]
	P value t-test	-	<.001	<.001
$Light_d$ (min)	Mean ± SD	229.82 ± 82.70	250.59 ± 71.03	249.64 ± 70.93
	95% CI mean	[213.69, 245.95]	[236.74, 264.45]	[235.81, 263.48]
	P value t-test	-	<.001	<.001
$Deep_d$ (min)	Mean ± SD	75.28 ± 37.05	83.30 ± 29.09	82.95 ± 29.19
	95% CI mean	[68.06, 82.51]	[77.63, 88.97]	[77.25, 88.64]
	P value t-test	-	.028	.035

Underlined P values represent statistically significant differences.
Emfit-A = Emfit measures during automatic analysis period, Emfit-M = Emfit measures during manual analysis period, TST = total sleep time, SOL = sleep onset latency, REML = REM latency, SE = sleep efficiency, $WASO_d$ = WASO duration, REM_d = REM duration, $Light_d$ = light sleep duration, $Deep_d$ = deep sleep duration, min = minutes, SD = standard deviation, CI = confidence interval, * = differences not normally distributed

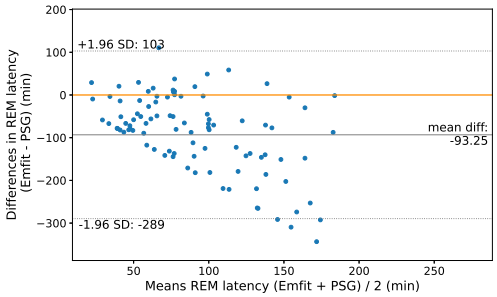
Figure 3.4: Bland-Altman plots for Emfit sleep measures



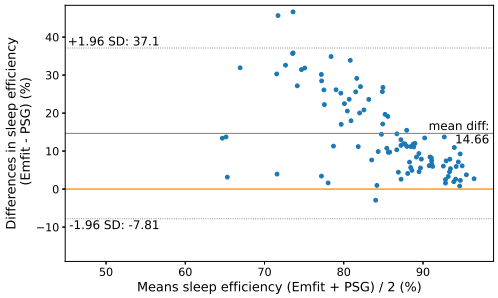
(a) Total sleep time



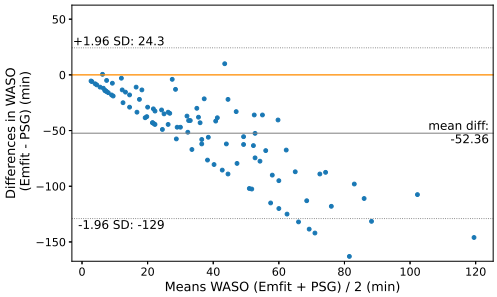
(b) Sleep onset latency



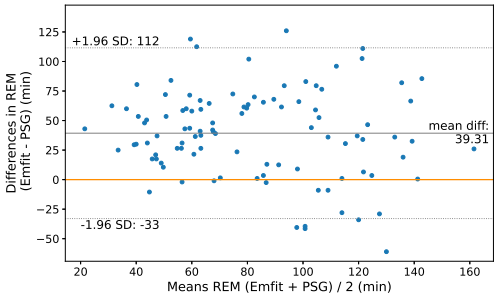
(c) REM latency



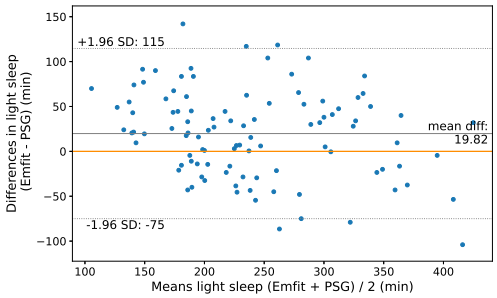
(d) Sleep efficiency



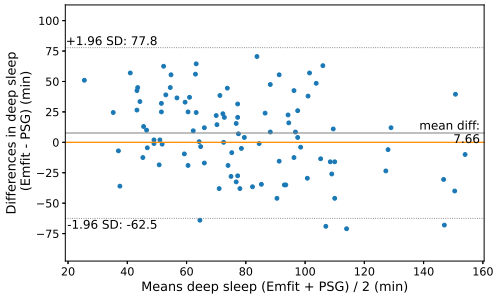
(e) WASO duration



(f) REM duration



(g) Light sleep duration



(h) Deep sleep duration

Differences between the sleep measures of Emfit and PSG during the manual analysis period. min = minutes, SD = standard deviation

Table 3.6: Epoch-by-epoch concordance Somnofy

Sleep stage	Sensitivity	Specificity	Accuracy	Matthews CC
Wake	0.60 ± 0.23 [0.56, 0.64]	0.94 ± 0.10 [0.92, 0.95]	0.87 ± 0.10 [0.85, 0.89]	0.57 ± 0.18 [0.54, 0.60]
REM	0.71 ± 0.29 [0.66, 0.76]	0.91 ± 0.07 [0.90, 0.92]	0.89 ± 0.07 [0.88, 0.90]	0.53 ± 0.29 [0.48, 0.58]
Light sleep	0.69 ± 0.14 [0.67, 0.72]	0.71 ± 0.13 [0.69, 0.74]	0.70 ± 0.10 [0.69, 0.72]	0.40 ± 0.20 [0.37, 0.44]
Deep sleep	0.67 ± 0.23 [0.63, 0.71]	0.93 ± 0.06 [0.92, 0.94]	0.88 ± 0.06 [0.87, 0.89]	0.59 ± 0.19 [0.55, 0.62]
Global	Balanced accuracy		0.67 ± 0.13 [0.64, 0.69]	
	Accuracy		0.67 ± 0.12 [0.65, 0.69]	
	Matthews CC		0.51 ± 0.17 [0.48, 0.54]	

Values presented as mean ± standard deviation followed by 95% confidence interval.
CC = correlation coefficient

Table 3.7: Epoch-by-epoch concordance Fitbit

Sleep stage	Sensitivity	Specificity	Accuracy	Matthews CC
Wake	0.39 ± 0.20 [0.33, 0.44]	0.97 ± 0.05 [0.96, 0.98]	0.87 ± 0.09 [0.85, 0.89]	0.45 ± 0.15 [0.41, 0.49]
REM	0.67 ± 0.28 [0.60, 0.74]	0.90 ± 0.06 [0.89, 0.92]	0.87 ± 0.06 [0.86, 0.89]	0.50 ± 0.25 [0.43, 0.56]
Light sleep	0.77 ± 0.11 [0.74, 0.80]	0.59 ± 0.16 [0.55, 0.63]	0.68 ± 0.10 [0.65, 0.70]	0.36 ± 0.19 [0.32, 0.41]
Deep sleep	0.56 ± 0.25 [0.50, 0.63]	0.95 ± 0.04 [0.93, 0.96]	0.87 ± 0.07 [0.85, 0.89]	0.52 ± 0.22 [0.46, 0.58]
Global	Balanced accuracy		0.60 ± 0.12 [0.56, 0.63]	
	Accuracy		0.64 ± 0.11 [0.61, 0.67]	
	Matthews CC		0.44 ± 0.16 [0.40, 0.49]	

Values presented as mean ± standard deviation followed by 95% confidence interval.
CC = correlation coefficient

Table 3.8: Epoch-by-epoch concordance Emfit

Sleep stage	Sensitivity	Specificity	Accuracy	Matthews CC
Wake	0.11 ± 0.13 [0.08, 0.13]	0.99 ± 0.02 [0.99, 0.99]	0.83 ± 0.12 [0.81, 0.85]	0.17 ± 0.19 [0.14, 0.21]
REM	0.45 ± 0.25 [0.40, 0.50]	0.81 ± 0.07 [0.79, 0.82]	0.76 ± 0.07 [0.74, 0.77]	0.20 ± 0.21 [0.15, 0.24]
Light sleep	0.63 ± 0.07 [0.62, 0.64]	0.52 ± 0.08 [0.50, 0.53]	0.57 ± 0.06 [0.56, 0.59]	0.14 ± 0.13 [0.12, 0.17]
Deep sleep	0.43 ± 0.18 [0.39, 0.46]	0.86 ± 0.05 [0.85, 0.87]	0.78 ± 0.06 [0.77, 0.79]	0.26 ± 0.17 [0.23, 0.29]
Global	Balanced accuracy		0.40 ± 0.09 [0.38, 0.42]	
	Accuracy		0.47 ± 0.09 [0.45, 0.49]	
	Matthews CC		0.19 ± 0.11 [0.16, 0.21]	

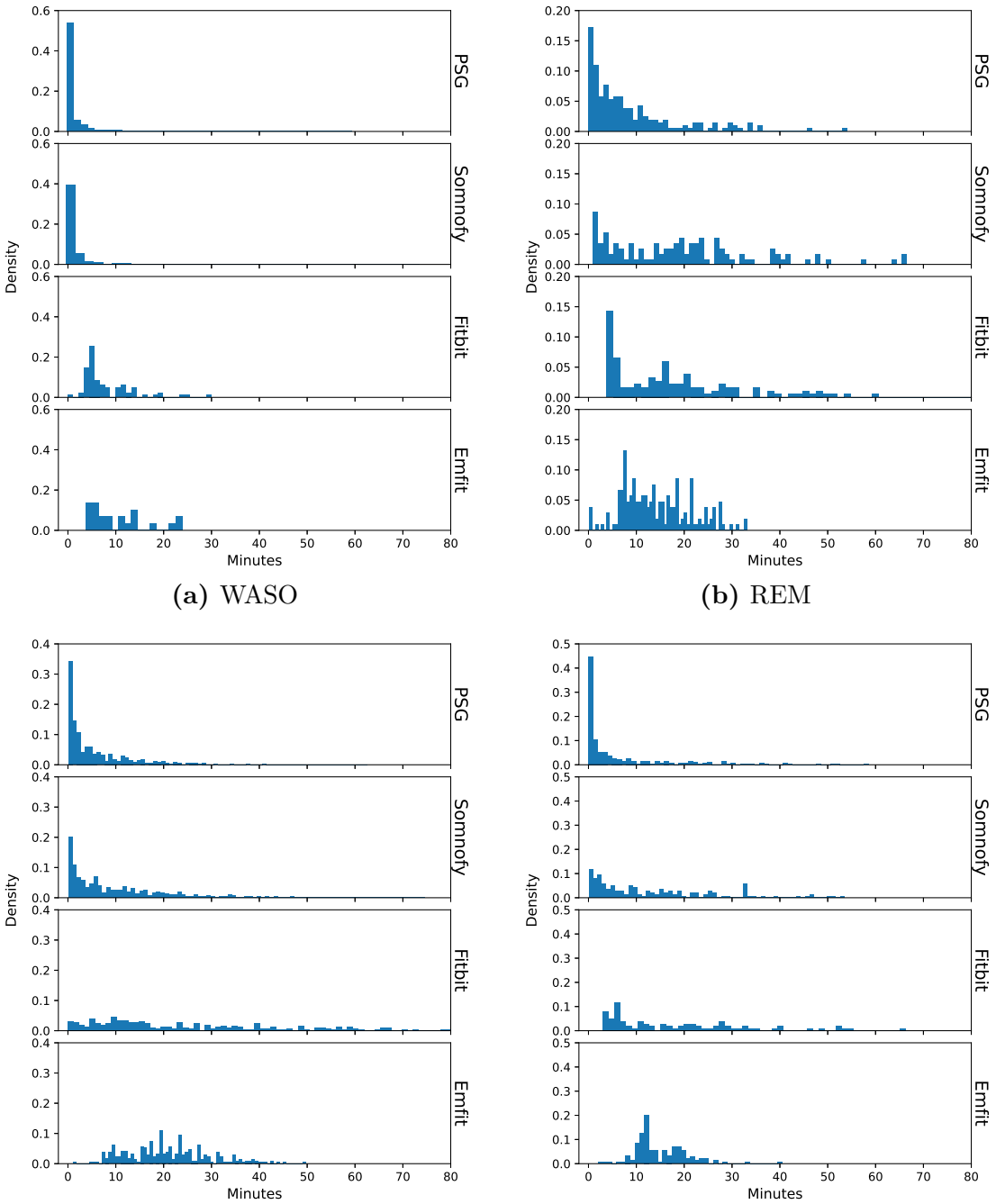
Values presented as mean ± standard deviation followed by 95% confidence interval.
CC = correlation coefficient

Table 3.9: Epoch-by-epoch concordance comparison on common patients

Sleep stage	Device	Sensitivity	Specificity	Accuracy	Matthews CC
Wake	Somnofy	**0.64** ± 0.21 [0.56, 0.72]	0.94 ± 0.07 [0.91, 0.97]	**0.89** ± 0.07 [0.86, 0.92]	**0.61** ± 0.15 [0.55, 0.67]
	Fitbit	0.38 ± 0.17 [0.31, 0.44]	0.98 ± 0.03 [0.97, 0.99]	0.88 ± 0.08 [0.85, 0.91]	0.46 ± 0.13 [0.41, 0.51]
	Emfit	0.09 ± 0.11 [0.05, 0.14]	**0.99** ± 0.03 [0.98, 1.00]	0.85 ± 0.10 [0.81, 0.89]	0.16 ± 0.18 [0.09, 0.23]
REM	Somnofy	**0.71** ± 0.29 [0.59, 0.82]	**0.93** ± 0.05 [0.91, 0.95]	**0.90** ± 0.05 [0.88, 0.92]	**0.57** ± 0.28 [0.46, 0.68]
	Fitbit	0.68 ± 0.29 [0.57, 0.79]	0.89 ± 0.07 [0.86, 0.91]	0.86 ± 0.07 [0.83, 0.89]	0.48 ± 0.26 [0.37, 0.58]
	Emfit	0.48 ± 0.22 [0.39, 0.57]	0.82 ± 0.07 [0.79, 0.85]	0.77 ± 0.06 [0.75, 0.80]	0.24 ± 0.20 [0.16, 0.32]
Light sleep	Somnofy	0.73 ± 0.09 [0.69, 0.77]	**0.72** ± 0.14 [0.67, 0.78]	**0.72** ± 0.08 [0.69, 0.76]	**0.45** ± 0.17 [0.38, 0.51]
	Fitbit	**0.77** ± 0.10 [0.73, 0.80]	0.60 ± 0.16 [0.53, 0.66]	0.68 ± 0.09 [0.65, 0.72]	0.36 ± 0.16 [0.30, 0.43]
	Emfit	0.64 ± 0.08 [0.61, 0.67]	0.51 ± 0.08 [0.48, 0.54]	0.58 ± 0.06 [0.55, 0.60]	0.15 ± 0.13 [0.09, 0.20]
Deep sleep	Somnofy	**0.65** ± 0.22 [0.57, 0.74]	0.94 ± 0.05 [0.92, 0.96]	**0.89** ± 0.06 [0.87, 0.91]	**0.59** ± 0.22 [0.50, 0.67]
	Fitbit	0.52 ± 0.24 [0.43, 0.61]	**0.95** ± 0.05 [0.93, 0.97]	0.87 ± 0.06 [0.85, 0.90]	0.50 ± 0.24 [0.41, 0.59]
	Emfit	0.40 ± 0.15 [0.34, 0.45]	0.85 ± 0.05 [0.83, 0.87]	0.77 ± 0.05 [0.75, 0.79]	0.23 ± 0.14 [0.18, 0.28]
Global	Balanced Accuracy	Somnofy	**0.68** ± 0.12 [0.63, 0.73]		
		Fitbit	0.59 ± 0.12 [0.54, 0.63]		
		Emfit	0.40 ± 0.08 [0.37, 0.43]		
	Accuracy	Somnofy	**0.70** ± 0.09 [0.67, 0.74]		
		Fitbit	0.64 ± 0.11 [0.60, 0.69]		
		Emfit	0.48 ± 0.09 [0.45, 0.52]		
	Matthews CC	Somnofy	**0.54** ± 0.15 [0.49, 0.60]		
		Fitbit	0.44 ± 0.16 [0.37, 0.50]		
		Emfit	0.19 ± 0.12 [0.15, 0.24]		

Values presented as mean ± standard deviation followed by 95% confidence interval.
CC = correlation coefficient

Figure 3.5: Distributions of sleep stages consecutive durations



(a) WASO

(b) REM

(c) Light sleep

(d) Deep sleep

Comparison of the estimated distributions of consecutive durations for each sleep stage between the PSG and the devices. The distributions are calculated for the common patients. Each plot is normalized so that the area under the histogram integrates to 1.

3.3 Empatica sleep stages

3.3.1 Features extraction and random forest

Table 3.10 displays the results of the random forest classification model using features extracted from the Empatica data. The table shows the epoch-by-epoch concordance metrics of the patients with valid data for all devices so that they can be directly compared with those of the other devices (Table 3.9), also the results with all Empatica patients are very similar. The results were obtained through leave-one-group-out cross-validation, where all patients with valid data from Empatica were utilized, but only the folds containing the test patient as one of the common patients were taken into account.

The model has very poor performance, in particular, it is unable to detect REM and deep sleep epochs (low sensitivity and high specificity), and most epochs are classified as light sleep (high sensitivity and low specificity). MCC ranges from weak correlation (wake and light sleep) to no correlation (REM and deep sleep). Overall metrics are also low, and the model generally performs worse than any of the other devices, despite being trained on this specific population.

Table 3.10: Sleep staging classification performance based on features extracted from Empatica on common patients

Sleep stage	Sensitivity	Specificity	Accuracy	Matthews CC
Wake	0.33 ± 0.15 [0.27, 0.39]	0.95 ± 0.04 [0.94, 0.97]	0.83 ± 0.08 [0.80, 0.86]	0.34 ± 0.13 [0.29, 0.39]
REM	0.00 ± 0.01 [0.00, 0.01]	0.99 ± 0.01 [0.99, 1.00]	0.84 ± 0.10 [0.80, 0.87]	-0.01 ± 0.04 [-0.02, 0.01]
Light sleep	0.92 ± 0.07 [0.89, 0.94]	0.18 ± 0.09 [0.15, 0.22]	0.55 ± 0.10 [0.51, 0.59]	0.14 ± 0.12 [0.09, 0.19]
Deep sleep	0.06 ± 0.09 [0.02, 0.09]	0.98 ± 0.03 [0.97, 1.00]	0.85 ± 0.08 [0.82, 0.88]	0.07 ± 0.11 [0.03, 0.11]
Global	Balanced accuracy		0.33 ± 0.04 [0.31, 0.35]	
	Accuracy		0.53 ± 0.10 [0.49, 0.58]	
	Matthews CC		0.18 ± 0.10 [0.14, 0.22]	

Epoch-by-epoch concordance for the random forest using the features extracted from Empatica on common patients. Values presented as mean ± standard deviation and 95% confidence interval. CC = correlation coefficient

3.3.2 Fully convolutional neural network

The 133 patients having valid Empatica data were divided into 3 sets: training, evaluation, and test. The demographics of each set are shown in Table 3.11. The test set represents the holdout set, which is not used in any way during training but is subsequently used to measure the model's performance on new data. The 27 patients (approximately 20% of all Empatica patients) with valid data across all devices were assigned to the test set to facilitate comparisons of this model's results with those of other devices. Of the remaining 106 patients, approximately 15% (15 patients) were randomly assigned to the evaluation set, with the remaining 91 patients comprising the training set. The data in the training set are utilized directly by the model for learning, while those in the evaluation set are employed to assess the model's performance during training. In all experiments using the temperature signal, the 6 patients with faulty temperature sensors were excluded, resulting in 86 patients for training and 26 patients for testing with an unchanged evaluation set. Cross-validation was not conducted in this case due to the resource-intensive nature of network training. Repeating the training for different folds for each cycle would have required too much time.

Table 3.11: Population characteristics of each set used in Empatica model training

	Empatica	Train set	Evaluation set	Test set
Valid patients: n	133	91	15	27
Females: n (%)	73 (54.89)	51 (56.04)	8 (53.33)	14 (51.85)
Age (years): mean ± SD [range]	45.32 ± 16.23 [18.2, 84.6]	46.23 ± 16.37 [18.2, 84.6]	44.57 ± 15.70 [18.6, 73.7]	42.24 ± 16.15 [19.0, 72.8]
Diagnosis: %				
Breathing disorders	57.14	59.34	60.00	51.86
Hypersomnolence	15.79	12.09	20.00	22.23
Parasomnias	4.51	4.40	6.67	3.70
Insomnia	6.02	6.59	6.67	3.70
Circadian disorders	1.50	2.20	0	0
Movement disorders	2.26	2.20	0	3.70
Other disorders	3.01	4.40	0	0
Healthy controls	4.51	3.29	0	11.11
Missing diagnosis	5.26	5.49	6.67	3.70

SD = standard deviation

Several experiments were conducted using this model. Two modes of input were tested: directly using the raw signals or using the spectrograms of the signals. Subsequent experiments were conducted using both modes.

The initial experiment focused on the number of epochs within the model

input window. The lengths of 128, 256, 512, and 1024 epochs were examined, equating to approximately 1 hour, 2 hours, 4 hours, and 8 and a half hours, respectively. The results indicated that, using both raw signals and spectrograms, the best performance was obtained at the maximum length of 1024 epochs, which is coherent with the outcomes in the original model paper [9].

The next experiment was to choose which signals to use as input. The initial study [9] utilized accelerometer and PPG, also known as BVP, data gathered from a wearable device. Empatica additionally offers various other signals, including EDA and skin temperature, as well as measurements derived from the raw signals, such as HR, which is derived from BVP. The various combinations of the five signals (accelerometer, BVP, EDA, temperature, and HR) were evaluated using fixed model parameters to facilitate a comparison of the results. The results of the most interesting combinations are shown in Table 3.12 using the raw signals and Table 3.13 using the spectrograms. The tables present the balanced accuracy which was chosen as the overall metric to compare the performance of the different experiments because, unlike accuracy, it is robust to class imbalance and is easier to interpret than MCC. Using raw signals as direct input, the table shows that heart rate is the most effective among all single signals, followed by the accelerometer, EDA, BVP, and finally, temperature. While combining signals does lead to improvements over some single signals, HR alone still remains the most effective. The difference in performance between BVP and HR is noteworthy because HR is derived from BVP, so BVP contains more information and is also higher frequency, but can also contain more noise than HR. Instead, spectrogram results (Table 3.13) demonstrate that the most effective signals are BVP and accelerometer, followed by EDA, HR, and temperature. Combining signals substantially increases balanced accuracy in this case, with the best result obtained by using BVP and accelerometer simultaneously while adding more signals seems to have diminishing returns. In general, it is evident that using spectrograms as input yields better performance than using raw data, as repeatedly observed in all experiments. Therefore, only spectrograms were utilized in the subsequent step.

The final step was hyperparameter tuning to find the best parameters for the model using the spectrograms of BVP and accelerometer signals as input. The parameters used were:

- M, the number of encoders and decoders, with a value between 8, 10, 12, 14;
- K, the kernel height of the 2D convolutions that have kernel size (K, 3), with a value between 4, 8, 16, 32;
- initial filter number, the number of filters used in the 2D convolution of the first encoder, which determines the number of its output features and is incremented for each successive encoder level, with a value between 4, 8, 16, 32;

Table 3.12: Balanced accuracies using different combinations of raw signals in input to the Empatica network

Signals	Balanced accuracy
BVP	0.32 ± 0.04 [0.30, 0.34]
ACC	0.39 ± 0.07 [0.37, 0.42]
EDA	0.38 ± 0.07 [0.36, 0.41]
TEMP	0.36 ± 0.07 [0.33, 0.39]
HR	**0.47** ± 0.09 [0.44, 0.51]
BVP+ACC	0.41 ± 0.07 [0.38, 0.43]
BVP+ACC+EDA+HR	0.39 ± 0.06 [0.37, 0.42]
BVP+ACC+TEMP	0.40 ± 0.06 [0.38, 0.43]
BVP+ACC+EDA+TEMP+HR	0.36 ± 0.05 [0.34, 0.38]

Values presented as mean ± standard deviation followed by 95% confidence interval.
BVP = blood volume pulse, ACC = accelerometry, EDA = electrodermal activity, TEMP = skin temperature, HR = heart rate

Table 3.13: Balanced accuracies using different combinations of signals spectrograms in input to the Empatica network

Signals	Balanced accuracy
BVP	0.47 ± 0.08 [0.44, 0.50]
ACC	0.47 ± 0.08 [0.44, 0.50]
EDA	0.43 ± 0.09 [0.40, 0.47]
TEMP	0.33 ± 0.06 [0.30, 0.35]
HR	0.42 ± 0.07 [0.39, 0.45]
BVP+ACC	**0.52** ± 0.08 [0.48, 0.55]
BVP+ACC+EDA	0.48 ± 0.07 [0.46, 0.52]
BVP+ACC+TEMP	0.42 ± 0.07 [0.39, 0.44]
BVP+ACC+EDA+TEMP	0.45 ± 0.08 [0.42, 0.49]

Values presented as mean ± standard deviation followed by 95% confidence interval.
BVP = blood volume pulse, ACC = accelerometry, EDA = electrodermal activity, TEMP = skin temperature, HR = heart rate

- learning rate, with a value between 0.01, 0.001, 0.0001.

The search was performed using HyperBand [18], which allows for faster parameter search than traditional approaches such as grid search. Early stopping was also used: each learning procedure was stopped when there was no improvement in accuracy on the evaluation set for 25 consecutive epochs, as done in [9]. The HyperBand was run for up to 250 epochs for each configuration, and the one with

better accuracy on the evaluation set was selected. The parameters of the best model are: M=10, K=16, initial filters=16, learning rate = 0.001; interestingly, these are the same values found in the paper [9] as a result of optimization. The EBE concordance metrics on the test set, which consists of the patients with valid data from all devices, are shown in Table 3.14. The results indicate that the model struggles slightly to recognize wake and REM epochs (lower sensitivity), classifying many epochs as light sleep when they are not (lower specificity), but in general it succeeds in achieving good performance, with the MCC showing a moderate correlation for all of them. These results are remarkably higher than those of the random forest classifier with feature extraction (Table 3.10), suggesting that the poor performance was not due to the low amount of informative signals, but rather to the overly restrictive feature extraction method. Comparing these scores with those of the other devices on the same patients (Table 3.9), the model generally performs better than Fitbit and Emfit, but Somnofy is still slightly better, for example when comparing global metrics. The achieved accuracy of 67% aligns with the 69% accuracy obtained in both the original paper describing the model architecture [9] and the previous paper that tested sleep stages classification with Empatica [8]. However, it is essential to observe that the studies used different datasets with varying populations, precluding a direct comparison of the results.

The sleep measures obtained from Empatica are presented in Table 3.15, demonstrating significant differences only for sleep onset latency and REM latency. This suggests that Empatica is more reliable than the other devices for these measures. These findings are also reflected in the Bland-Altman plots in Figure 3.6, where sleep onset latency and REM latency have larger average differences of approximately 6 and 41 minutes of underestimation, respectively. The total sleep time exhibits a higher average overestimation difference of nearly 12 minutes when compared to the others, which display less difference.

Figure 3.7 displays the distributions of consecutive durations of sleep stages, where the similarity between the distributions of PSG and Empatica is noticeable, especially for WASO and deep sleep. The distribution of REM is more concentrated on short durations than PSG, thus overestimating the fragmentation of REM, which is also the case to a lesser extent for light sleep. The Empatica distributions exhibit a greater similarity than those of the other devices (Figure 3.5), but it must be taken into account that this model was trained on this specific population.

Table 3.14: Sleep staging classification performance of the best optimized convolutional network from Empatica signals spectrograms on common patients

Sleep stage	Sensitivity	Specificity	Accuracy	Matthews CC
Wake	0.59 ± 0.22 [0.50, 0.68]	0.94 ± 0.05 [0.92, 0.96]	0.88 ± 0.06 [0.85, 0.90]	0.54 ± 0.18 [0.47, 0.62]
REM	0.55 ± 0.22 [0.46, 0.64]	0.93 ± 0.04 [0.92, 0.95]	0.88 ± 0.05 [0.86, 0.90]	0.46 ± 0.21 [0.38, 0.54]
Light sleep	0.71 ± 0.07 [0.68, 0.74]	0.67 ± 0.12 [0.62, 0.71]	0.70 ± 0.06 [0.67, 0.72]	0.38 ± 0.13 [0.33, 0.43]
Deep sleep	0.70 ± 0.21 [0.62, 0.78]	0.92 ± 0.06 [0.90, 0.94]	0.89 ± 0.03 [0.87, 0.90]	0.61 ± 0.12 [0.56, 0.65]
Global	Balanced accuracy		0.64 ± 0.08 [0.60, 0.67]	
	Accuracy		0.67 ± 0.07 [0.64, 0.70]	
	Matthews CC		0.49 ± 0.12 [0.44, 0.53]	

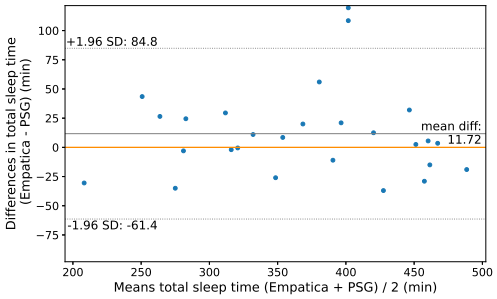
Epoch-by-epoch concordance for the best convolutional neural network using as input the Empatica BVP and accelerometer signals spectrograms as a result of hyperparameter tuning. Values presented as mean ± standard deviation followed by 95% confidence interval. Results calculated using only patients with all devices as a test so that they could be compared with those of other devices.
CC = correlation coefficient

Table 3.15: Sleep measures of the best Empatica model on common patients

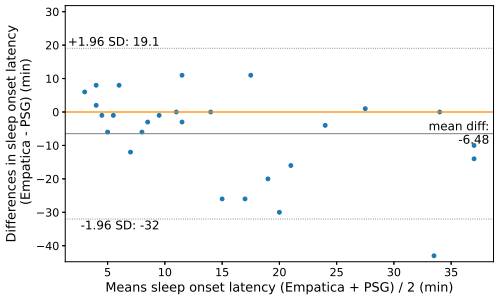
Sleep measure		PSG	Empatica
TST (min)	Mean ± SD	363.19 ± 79.58	374.91 ± 80.09
	95% CI mean	[331.70, 394.67]	[343.23, 406.59]
	P value t-test	-	.23*
SOL (min)	Mean ± SD	18.63 ± 14.94	12.15 ± 9.73
	95% CI mean	[12.72, 24.54]	[8.30, 16.00]
	P value t-test	-	.029*
REML (min)	Mean ± SD	132.65 ± 74.91	93.94 ± 56.47
	95% CI mean	[103.02, 162.29]	[71.61, 116.28]
	P value t-test	-	.033*
SE (%)	Mean ± SD	74.29 ± 13.55	76.36 ± 12.17
	95% CI mean	[68.93, 79.65]	[71.54, 81.17]
	P value t-test	-	.14
$WASO_d$ (min)	Mean ± SD	59.93 ± 40.87	56.07 ± 34.07
	95% CI mean	[43.76, 76.09]	[42.60, 69.55]
	P value t-test	-	.79*
REM_d (min)	Mean ± SD	62.78 ± 33.85	60.43 ± 31.14
	95% CI mean	[49.39, 76.17]	[48.11, 72.74]
	P value t-test	-	.69
$Light_d$ (min)	Mean ± SD	223.83 ± 59.75	233.02 ± 61.45
	95% CI mean	[200.20, 247.47]	[208.71, 257.33]
	P value t-test	-	.23
$Deep_d$ (min)	Mean ± SD	76.57 ± 32.13	81.46 ± 38.38
	95% CI mean	[63.86, 89.28]	[66.28, 96.65]
	P value t-test	-	.44

Underlined P values represent statistically significant differences.
Empatica = Empatica measures using the best model from hyperparameters tuning using as input the BVP and accelerometer spectrograms, TST = total sleep time, SOL = sleep onset latency, REML = REM latency, SE = sleep efficiency, $WASO_d$ = WASO duration, REM_d = REM duration, $Light_d$ = light sleep duration, $Deep_d$ = deep sleep duration, min = minutes, SD = standard deviation, CI = confidence interval, * = differences not normally distributed

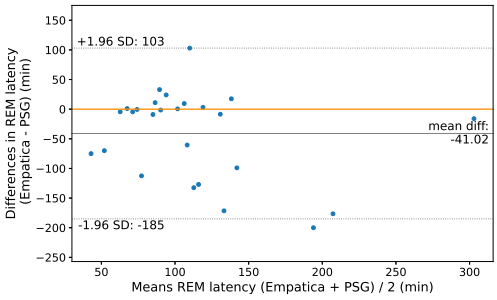
Figure 3.6: Bland-Altman plots sleep measures of the best Empatica model on common patients



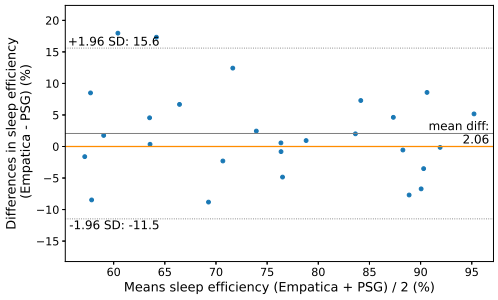
(a) Total sleep time



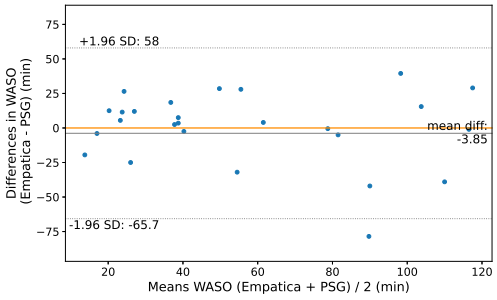
(b) Sleep onset latency



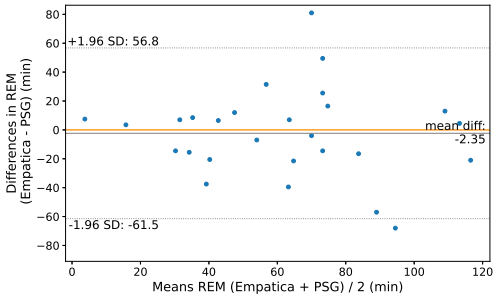
(c) REM latency



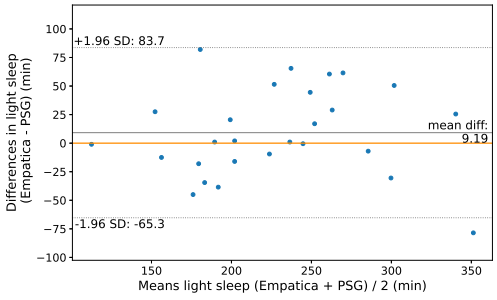
(d) Sleep efficiency



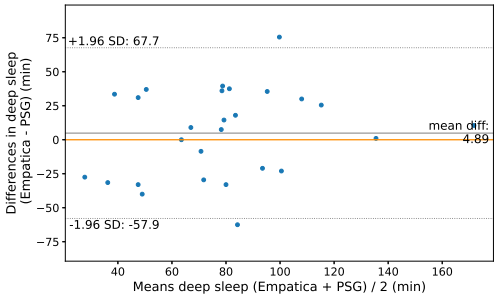
(e) WASO duration



(f) REM duration



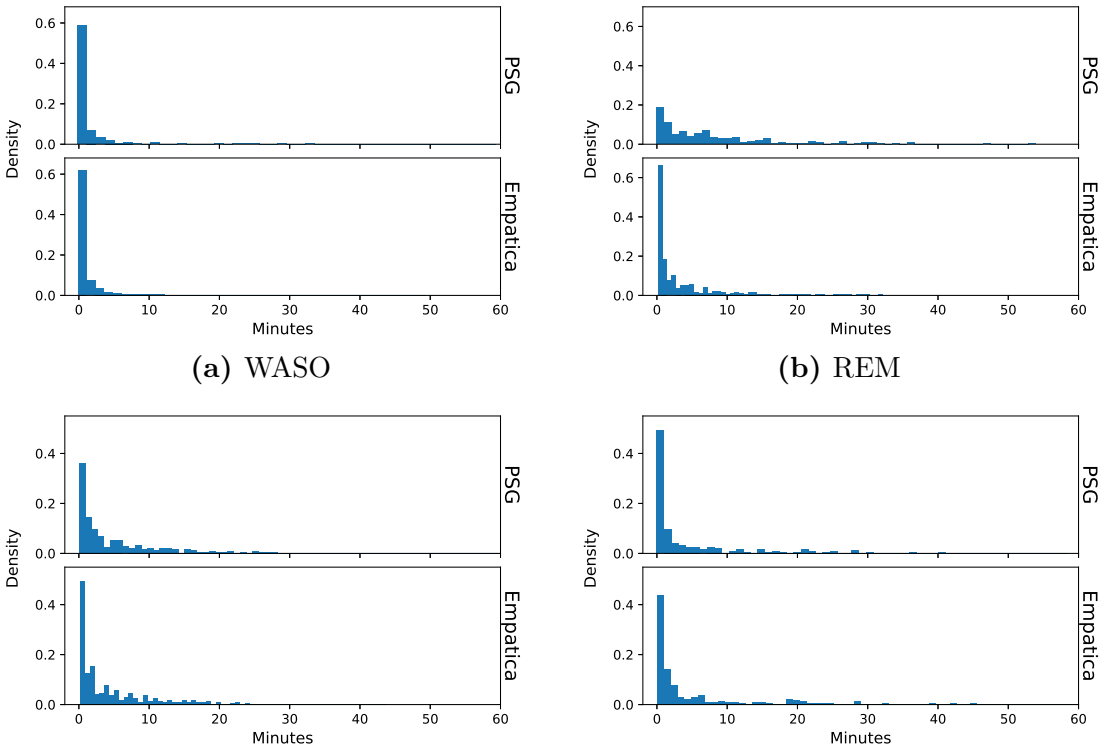
(g) Light sleep duration



(h) Deep sleep duration

Differences between the sleep measures of the best Empatica model and PSG on common patients. min = minutes, SD = standard deviation

Figure 3.7: Distributions of sleep stages consecutive durations of the best Empatica model on common patients



(a) WASO **(b)** REM

(c) Light sleep **(d)** Deep sleep

Comparison of the estimated distributions of consecutive durations for each sleep stage between the PSG and the best Empatica model from hyperparameters tuning using as input the BVP and accelerometer spectrograms. The distributions are calculated for the common patients. Each plot is normalized so that the area under the histogram integrates to 1.

3.4 Fusion of devices sleep stages

The sleep stages provided by Somnofy, Fitbit, Emfit, and the fully convolutional network from Empatica data are combined using a random forest algorithm. The Empatica model used is the one obtained from hyperparameter tuning. Only the 27 patients with valid data for all devices were used to avoid missing sleep stages and to use only patients in the Empatica test dataset. This approach ensures that the results can be directly compared with those of individual devices.

To observe the impact of each device, all combinations of two, three, and finally all four devices were tested. Table 3.16 displays the outcomes for each combination after leave-one-group-out cross-validation. The accuracies for each sleep stage are predominantly high, all close to 90 percent, except for light sleep, which is around 70 percent. Considering that the highest balanced accuracy achieved so far in common patients was 68% for Somnofy, in the two-device combinations only Somnofy plus Empatica manages to perform better than Somnofy alone. On the other hand, all the combinations with three devices that include the Somnofy display enhanced balanced accuracy relative to single devices. The combination of all four devices does not seem to further improve performance and has very similar results to the 3-device combinations. The most effective combination appears to be Somnofy, Fitbit, and Empatica.

Other experiments have been done, including giving in input to the random forest the predecessor and subsequent epochs in addition to the current epoch, but no improvement in performance was found.

3.5 Comparison results of the devices and models

A summary comparison between the results of Somnofy, Fitbit, Emfit, and the best results with Empatica and devices fusion on common patients is shown in Table 3.17 for individual sleep stage metrics and Table 3.18 for global metrics. For Empatica, the results of the best model obtained with hyperparameter tuning are shown, and for device fusion, the combination of Somnofy, Fitbit, and Empatica is reported.

Summarizing the results, the tables confirm how Somnofy performs considerably better than Fitbit, which in turn performs substantially better than Emfit. The Empatica algorithm manages to perform better than Fitbit, but not Somnofy. The fusion of the different devices, on the other hand, proves to be effective and performs better than Somnofy alone, with a better accuracy both in individual sleep stages and overall.

Table 3.19 presents a further comparison of overall metrics for common patients categorized by diagnosis to demonstrate variations in performance among different

sleep disorders. The findings indicate that commercial devices exhibit superior performance on healthy patients, perhaps due to their model training data being predominantly composed of this subgroup, but are limited when dealing with disorders. On the other hand, the Empatica and fusion devices models exhibit consistent performance across all sleep disorders with less noticeable variation between different diagnoses. However, their outcomes for healthy patients are inferior compared to Somnofy.

Table 3.16: Performance of sleep stages fusion for all devices combinations

Somnofy	Fitbit	Emfit	Empatica	Accuracy				Balanced accuracy	Matthews CC
				Wake	REM	Light	Deep		
x	x			0.89 ± 0.07 [0.87, 0.92]	0.89 ± 0.06 [0.86, 0.91]	0.69 ± 0.08 [0.66, 0.72]	0.87 ± 0.06 [0.84, 0.89]	0.65 ± 0.10 [0.61, 0.69]	0.49 ± 0.14 [0.43, 0.55]
x		x		0.90 ± 0.07 [0.87, 0.93]	0.90 ± 0.05 [0.88, 0.92]	0.71 ± 0.09 [0.67, 0.74]	0.87 ± 0.07 [0.84, 0.90]	0.67 ± 0.10 [0.63, 0.70]	0.51 ± 0.15 [0.45, 0.57]
x			x	0.90 ± 0.07 [0.87, 0.93]	0.90 ± 0.05 [0.88, 0.92]	0.72 ± 0.08 [0.68, 0.75]	0.89 ± 0.05 [0.87, 0.91]	0.71 ± 0.09 [0.68, 0.75]	0.51 ± 0.14 [0.46, 0.57]
	x	x		0.89 ± 0.08 [0.86, 0.92]	0.87 ± 0.07 [0.84, 0.90]	0.68 ± 0.09 [0.64, 0.71]	0.86 ± 0.07 [0.83, 0.89]	0.64 ± 0.11 [0.59, 0.68]	0.43 ± 0.16 [0.37, 0.49]
	x		x	0.89 ± 0.06 [0.86, 0.91]	0.89 ± 0.06 [0.87, 0.92]	0.70 ± 0.06 [0.67, 0.72]	0.88 ± 0.04 [0.86, 0.90]	0.66 ± 0.08 [0.63, 0.70]	0.48 ± 0.11 [0.44, 0.52]
		x	x	0.88 ± 0.07 [0.86, 0.91]	0.88 ± 0.04 [0.86, 0.90]	0.69 ± 0.06 [0.67, 0.71]	0.88 ± 0.04 [0.86, 0.90]	0.63 ± 0.08 [0.60, 0.66]	0.47 ± 0.11 [0.43, 0.51]
x	x	x		0.90 ± 0.08 [0.87, 0.93]	0.90 ± 0.05 [0.88, 0.92]	0.70 ± 0.10 [0.66, 0.74]	0.88 ± 0.06 [0.85, 0.90]	0.70 ± 0.11 [0.66, 0.74]	0.49 ± 0.15 [0.43, 0.55]
x	x		x	**0.91** ± 0.07 [0.88, 0.93]	**0.91** ± 0.05 [0.88, 0.93]	**0.73** ± 0.08 [0.69, 0.76]	**0.90** ± 0.05 [0.88, 0.91]	**0.73** ± 0.09 [0.69, 0.76]	**0.54** ± 0.13 [0.48, 0.59]
x		x	x	0.90 ± 0.07 [0.87, 0.93]	**0.91** ± 0.04 [0.89, 0.92]	0.71 ± 0.08 [0.68, 0.74]	0.89 ± 0.05 [0.86, 0.91]	0.72 ± 0.09 [0.69, 0.76]	0.51 ± 0.12 [0.46, 0.56]
	x	x	x	0.89 ± 0.08 [0.86, 0.92]	0.89 ± 0.05 [0.87, 0.91]	0.69 ± 0.08 [0.66, 0.72]	0.88 ± 0.05 [0.86, 0.90]	0.67 ± 0.08 [0.64, 0.71]	0.47 ± 0.13 [0.41, 0.52]
x	x	x	x	0.90 ± 0.07 [0.87, 0.93]	**0.91** ± 0.04 [0.89, 0.92]	0.72 ± 0.08 [0.69, 0.76]	0.89 ± 0.06 [0.87, 0.92]	0.72 ± 0.09 [0.69, 0.76]	0.53 ± 0.13 [0.48, 0.59]

Epoch-by-epoch concordance accuracy for each sleep stage and overall balanced accuracy and MCC for the fusion of sleep stages from different devices using a random forest. All possible device combinations are displayed, with each "x" indicating that the device was used for that combination. Values presented as mean ± standard deviation followed by 95% confidence interval. Results calculated using only patients with all devices.
CC = correlation coefficient

Table 3.17: Epoch-by-epoch concordance comparison of single stage metrics between devices and proposed models on common patients

Sleep stage	Model	Sensitivity	Specificity	Accuracy	Matthews CC
Wake	Somnofy	0.64 ± 0.21	0.94 ± 0.07	0.89 ± 0.07	**0.61** ± 0.15
	Fitbit	0.38 ± 0.17	0.98 ± 0.03	0.88 ± 0.08	0.46 ± 0.13
	Emfit	0.09 ± 0.11	**0.99** ± 0.03	0.85 ± 0.10	0.16 ± 0.18
	Empatica	0.59 ± 0.22	0.94 ± 0.05	0.88 ± 0.06	0.54 ± 0.18
	Fusion	**0.79** ± 0.18	0.91 ± 0.08	**0.91** ± 0.07	0.54 ± 0.15
REM	Somnofy	**0.71** ± 0.29	0.93 ± 0.05	0.90 ± 0.05	**0.57** ± 0.28
	Fitbit	0.68 ± 0.29	0.89 ± 0.07	0.86 ± 0.07	0.48 ± 0.26
	Emfit	0.48 ± 0.22	0.82 ± 0.07	0.77 ± 0.06	0.24 ± 0.20
	Empatica	0.55 ± 0.22	0.93 ± 0.04	0.88 ± 0.05	0.46 ± 0.21
	Fusion	0.63 ± 0.26	**0.94** ± 0.05	**0.91** ± 0.05	0.51 ± 0.28
Light sleep	Somnofy	0.73 ± 0.09	0.72 ± 0.14	0.72 ± 0.08	**0.45** ± 0.17
	Fitbit	**0.77** ± 0.10	0.60 ± 0.16	0.68 ± 0.09	0.36 ± 0.16
	Emfit	0.64 ± 0.08	0.51 ± 0.08	0.58 ± 0.06	0.15 ± 0.13
	Empatica	0.71 ± 0.07	0.67 ± 0.12	0.70 ± 0.06	0.38 ± 0.13
	Fusion	0.70 ± 0.11	**0.76** ± 0.12	**0.73** ± 0.08	**0.45** ± 0.14
Deep sleep	Somnofy	0.65 ± 0.22	0.94 ± 0.05	0.89 ± 0.06	0.59 ± 0.22
	Fitbit	0.52 ± 0.24	**0.95** ± 0.05	0.87 ± 0.06	0.50 ± 0.24
	Emfit	0.40 ± 0.15	0.85 ± 0.05	0.77 ± 0.05	0.23 ± 0.14
	Empatica	0.70 ± 0.21	0.92 ± 0.06	0.89 ± 0.03	0.61 ± 0.12
	Fusion	**0.78** ± 0.22	0.92 ± 0.06	**0.90** ± 0.05	**0.62** ± 0.18

Epoch-by-epoch concordance of the devices providing sleep stages (Somnofy, Fitbit, and Emfit) and the best models proposed (fine-tuned convolutional model using spectrograms of BVP and accelerometer for Empatica, and the combination of Somnofy, Fitbit and Empatica for the fusion model). Sensitivity, specificity, accuracy, and Matthews correlation coefficient for each sleep stage are reported and they were calculated using only patients with all devices. Values presented as mean ± standard deviation.
CC = correlation coefficient

Table 3.18: Epoch-by-epoch concordance comparison of overall metrics between devices and proposed models on common patients

Model	Balanced accuracy	Accuracy	Matthews CC
Somnofy	0.68 ± 0.12	0.70 ± 0.09	**0.54** ± 0.15
Fitbit	0.59 ± 0.12	0.64 ± 0.11	0.44 ± 0.16
Emfit	0.40 ± 0.08	0.48 ± 0.09	0.19 ± 0.12
Empatica	0.64 ± 0.08	0.67 ± 0.07	0.49 ± 0.12
Fusion	**0.73** ± 0.09	**0.72** ± 0.09	**0.54** ± 0.13

Epoch-by-epoch concordance of the devices providing sleep stages (Somnofy, Fitbit, and Emfit) and the best models proposed (fine-tuned convolutional model using spectrograms of BVP and accelerometer for Empatica, and the combination of Somnofy, Fitbit, and Empatica for the fusion model). Overall balanced accuracy, accuracy, and Matthews correlation coefficient are reported and they were calculated using only patients with all devices. Values presented as mean ± standard deviation.
CC = correlation coefficient

Table 3.19: Comparison of overall epoch-by-epoch concordance metrics on common patients divided by diagnosis between devices and proposed models

Disorder	Metric	Somnofy	Fitbit	Emfit	Empatica	Fusion
Breathing disorders (14)	Bal. Acc.	0.64	0.55	0.40	0.64	**0.73**
	Accuracy	0.68	0.61	0.48	0.68	**0.70**
	MCC	**0.50**	0.39	0.19	0.49	**0.50**
Hypersomnolence (6)	Bal. Acc.	0.73	0.66	0.39	0.64	**0.75**
	Accuracy	0.75	0.71	0.52	0.67	**0.79**
	MCC	0.61	0.52	0.21	0.46	**0.64**
Healthy controls (3)	Bal. Acc.	**0.84**	0.67	0.45	0.69	0.75
	Accuracy	**0.81**	0.72	0.55	0.72	0.77
	MCC	**0.72**	0.56	0.26	0.56	0.62
Insomnia (1)	Bal. Acc.	0.64	0.56	0.43	0.53	**0.70**
	Accuracy	**0.64**	0.56	0.46	0.49	0.60
	MCC	0.48	0.38	0.22	0.31	**0.47**
Movement disorders (1)	Bal. Acc.	**0.73**	0.40	0.35	0.67	0.72
	Accuracy	0.69	0.50	0.32	**0.74**	0.67
	MCC	0.55	0.26	0.20	**0.63**	0.50
Parasomnias (1)	Bal. Acc.	0.49	0.57	0.39	0.48	**0.61**
	Accuracy	0.53	**0.61**	0.43	0.56	0.57
	MCC	0.30	**0.43**	0.13	0.35	0.33
Missing diagnosis (1)	Bal. Acc.	0.68	0.60	0.28	0.63	**0.69**
	Accuracy	**0.73**	0.70	0.42	0.63	**0.73**
	MCC	**0.56**	0.49	0.00	0.45	0.53

Overall epoch-by-epoch concordance divided per diagnosis of the patients with valid data of all devices. The performances are presented for the devices providing sleep stages (Somnofy, Fitbit, and Emfit) and the best models proposed (fine-tuned convolutional model using spectrograms of BVP and accelerometer for Empatica, and the combination of Somnofy, Fitbit, and Empatica for the fusion model). Values represent the mean and the number in parentheses following each diagnosis indicates the number of patients affected by the disorder.
Bal. Acc. = balanced accuracy, MCC = Matthews correlation coefficient

Chapter 4

Discussion and conclusion

This thesis evaluates the potential use of commercial devices as a support or alternative to PSG for sleep stages classification in patients with sleep disorders. The analysis is based on data from a prior study that included four commercial devices: Somnofy, Fitbit Inspire 2, Emfit, and Empatica E4, added to patients undergoing PSG. First, a comprehensive evaluation was conducted on Somnofy, Fitbit, and Emfit devices that automatically provide sleep stages. Then, customized algorithms were proposed, initially using one device and then combining multiple devices.

Sleep stages classifications from commercial devices have demonstrated statistically significant differences from PSG for sleep measures such as total sleep time and REM latency. As these measures are among the parameters used to diagnose sleep disorders, it is not advisable to make a diagnosis using these device-based measures. Emfit, in particular, performed inadequately, specifically failing to detect periods of wakefulness. Fitbit experienced some difficulty with this type of patients and was unable to recognize sleep stages of very short durations. Somnofy, on the other hand, was able to provide accurate scores across all sleep stages and was effective at detecting short duration sleep stages.

The Empatica device was selected to train a custom algorithm due to its high frequency raw data and diverse signal types. The algorithm, adapted from [9] and fine-tuned, was found to be effective on Empatica and in patients with sleep disorders, showing superior results to Fitbit, the other wearable device, but not reaching the performance of Somnofy. Different combinations and types of inputs were tested, revealing that PPG and accelerometer spectrograms were the most effective.

The novel proposed approach of fusing sleep stages from different devices has shown promising results, improving the overall accuracy of sleep stages classification in comparison to individual devices. It is important to point out that some of these devices are contactless, which minimizes the burden of using multiple devices

simultaneously and thus makes it a viable option.

In conclusion, this study demonstrates that current commercial devices can achieve accuracies of over 70% in classifying sleep stages in patients with sleep disorders. Comparing the accuracy of these devices to that of doctors manually scoring PSG, which was estimated at 88% [3], there is still a gap. However, the advantages of these devices, such as the ability to be used easily in one's own home for multiple nights, make them capable of supporting studies done with PSG. Nonetheless, limitations still exist, particularly regarding sleep measures, and further improvements are necessary to enable the diagnosis of sleep disorders using only these devices.

Limitations and future developments

The inter-rater agreement for PSG among doctors can be seen as a limitation since the gold standard used for both training and testing is unstable and the results obtained may vary significantly from one doctor to another. Therefore, Empatica and fusion models should be tested on additional datasets, preferably with a larger number of patients and diverse population characteristics, to obtain a more reliable performance estimation.

A limitation of this dataset was that there were few patients with valid data for all devices relative to the total number of patients. As a result, only a small number of patients were available to train and test the sleep stages fusion model from multiple devices.

As a future development in sleep stages fusion, it is possible to utilize raw signals directly without relying on the device's classification algorithm. This approach could potentially be more effective, and since several of the devices used in this study provide signals at relatively low frequencies (e.g., Emfit every 4 seconds), it could be explored with other commercial devices as well.

Bibliography

[1] American Academy of Sleep Medicine. *International Classification of Sleep Disorders*. 3rd ed. Darien, IL: American Academy of Sleep Medicine, 2014 (cit. on p. 2).

[2] Ulysses J. Magalang et al. «Agreement in the Scoring of Respiratory Events and Sleep Among International Sleep Centers». In: *Sleep* 36.4 (Apr. 2013), pp. 591–596. ISSN: 0161-8105. DOI: 10.5665/sleep.2552. eprint: https://academic.oup.com/sleep/article-pdf/36/4/591/26661198/aasm.36.4.591.pdf. URL: https://doi.org/10.5665/sleep.2552 (cit. on p. 3).

[3] Ståle Toften, Ståle Pallesen, Maria Hrozanova, Frode Moen, and Janne Grønli. «Validation of sleep stage classification using non-contact radar technology and machine learning (Somnofy®)». In: *Sleep Medicine* 75 (2020), pp. 54–61. ISSN: 1389-9457. DOI: https://doi.org/10.1016/j.sleep.2020.02.022. URL: https://www.sciencedirect.com/science/article/pii/S1389945720301027 (cit. on pp. 3, 4, 52).

[4] Su Eun Lim, Ho Seok Kim, Si Woo Lee, Kwang-Ho Bae, and Young Hwa Baek. «Validation of Fitbit Inspire 2(TM) Against Polysomnography in Adults Considering Adaptation for Use». In: *Nat Sci Sleep* 15 (2023), pp. 59–67 (cit. on p. 4).

[5] Kiran Kumar, Guruswamy Ravindran, Giuseppe Atzori, Damion Lambert, Hana Hassanin, Victoria Revell, and Derk-Jan Dijk. «Three Contactless Sleep Technologies Compared with Polysomnography and Actigraphy in a Heterogenous Group of Older Men and Women in a Model of Mild Sleep Disturbance». In: (2022). DOI: 10.21203/rs.3.rs-2251437/v1. URL: https://doi.org/10.21203/rs.3.rs-2251437/v1 (cit. on p. 4).

[6] Benjamin Stucky, Ian Clark, Yasmine Azza, Walter Karlen, Peter Achermann, Birgit Kleim, and Hans-Peter Landolt. «Validation of Fitbit Charge 2 Sleep and Heart Rate Estimates Against Polysomnographic Measures in Shift Workers: Naturalistic Study». In: *J Med Internet Res* 23.10 (Oct. 2021), e26476. ISSN: 1438-8871. DOI: 10.2196/26476. URL: http://www.ncbi.nlm.nih.gov/pubmed/34609317 (cit. on pp. 4, 11, 25).

[7] Mahnoosh Kholghi, Irene Szollosi, Mitchell Hollamby, Dana Kai Bradford, and Qing Zhang. «A validation study of a ballistocardiograph sleep tracker against polysomnography». In: *Journal of Clinical Sleep Medicine* 18 (4 Apr. 2022), pp. 1203–1210. ISSN: 15509397. DOI: 10.5664/jcsm.9754 (cit. on p. 4).

[8] Qiao Li, Qichen Li, Ayse S. Cakmak, Giulia Da Poian, Donald L. Bliwise, Viola Vaccarino, Amit J. Shah, and Gari D. Clifford. «Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables». In: *Physiological Measurement* 42 (4 Apr. 2021). ISSN: 13616579. DOI: 10.1088/1361-6579/abf1b0. URL: https://pubmed.ncbi.nlm.nih.gov/33761477/ (cit. on pp. 5, 14, 40).

[9] Mads Olsen, Jamie M. Zeitzer, Risa N. Richardson, Polina Davidenko, Poul J. Jennum, Helge B. D. Sørensen, and Emmanuel Mignot. «A Flexible Deep Learning Architecture for Temporal Sleep Stage Classification Using Accelerometry and Photoplethysmography». In: *IEEE Transactions on Biomedical Engineering* 70.1 (2023), pp. 228–237. DOI: 10.1109/TBME.2022.3187945 (cit. on pp. 5, 14–16, 38–40, 51).

[10] J. Martin Bland and Douglas G. Altman. «Statistical methods for assessing agreement between two methods of clinical measurement.» In: *Lancet (London, England)* 1 (8476 Feb. 1986), pp. 307–10. ISSN: 0140-6736 (cit. on p. 8).

[11] Rosario Delgado and Xavier-Andoni Tibau. «Why Cohen's Kappa should be avoided as performance measure in classification». In: *PLOS ONE* 14.9 (Sept. 2019), pp. 1–26. DOI: 10.1371/journal.pone.0222916. URL: https://doi.org/10.1371/journal.pone.0222916 (cit. on p. 10).

[12] Simon Föll, Martin Maritsch, Federica Spinola, Varun Mishra, Filipe Barata, Tobias Kowatsch, Elgar Fleisch, and Felix Wortmann. «FLIRT: A feature generation toolkit for wearable data». In: *Computer Methods and Programs in Biomedicine* 212 (2021), p. 106461. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2021.106461. URL: https://www.sciencedirect.com/science/article/pii/S0169260721005356 (cit. on pp. 12, 13).

[13] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. «U-Sleep: resilient high-frequency sleep staging». In: *npj Digital Medicine* 4.1 (Apr. 2021), p. 72. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00440-5. URL: https://doi.org/10.1038/s41746-021-00440-5 (cit. on p. 14).

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-Net: Convolutional Networks for Biomedical Image Segmentation». In: (2015). arXiv: 1505.04597 [cs.CV] (cit. on p. 14).

[15] Thorsten Falk et al. «U-Net: deep learning for cell counting, detection, and morphometry». In: *Nature Methods* 16.1 (Jan. 2019), pp. 67–70. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0261-2. URL: https://doi.org/10.1038/s41592-018-0261-2 (cit. on p. 14).

[16] Dan Hendrycks and Kevin Gimpel. «Gaussian Error Linear Units (GELUs)». In: (2023). arXiv: 1606.08415 [cs.LG] (cit. on p. 15).

[17] Sergey Ioffe and Christian Szegedy. «Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift». In: (2015). arXiv: 1502.03167 [cs.LG] (cit. on p. 15).

[18] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. «Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization». In: (2018). arXiv: 1603.06560 [cs.LG] (cit. on p. 39).