Heart Rate During Sleep Measured Using Finger-, Wrist- and Chest-Worn Devices: A Comparison Study

Nouran Abdalazim[1(✉)], Joseba Aitzol Arbilla Larraza[1], Leonardo Alchieri[1], Lidia Alecci[1], Silvia Santini[1], and Shkurta Gashi[2]

[1] Università della Svizzera italiana (USI), Lugano, Switzerland
{nouran.abdalazim,joseba.aitzol.arbilla.larraza,leonardo.alchieri,lidia.alecci,silvia.santini}@usi.ch
[2] ETH AI Center, Zürich, Switzerland
shkurta.gashi@ai.ethz.ch

Abstract. Wearable heart rate (HR) sensing devices are increasingly used to monitor human health. The availability and the quality of the HR measurements may however be affected by the body location at which the device is worn. The goal of this paper is to compare HR data collected from different devices and body locations and to investigate their interchangeability at different stages of the data analysis pipeline. To this goal, we conduct a data collection campaign and collect HR data from three devices worn at different body positions (finger, wrist, chest): The Oura ring, the Empatica E4 wristband and the Polar chestbelt. We recruit five participants for 30 nights and gather HR data along with self-reports about sleep behavior. We compare the raw data, the features extracted from this data over different window sizes, and the performance of models that use these features in recognizing sleep quality. Raw HR data from the three devices show a high positive correlation. When features are extracted from the raw data, though, both small and significant differences can be observed. Ultimately, the accuracy of a sleep quality recognition classifier does not show significant differences when the input data is derived from the Oura ring or the E4 wristband. Taken together, our results indicate that the HR measurements collected from the considered devices and body locations are interchangeable. These findings open up new opportunities for sleep monitoring systems to leverage multiple devices for continuous sleep tracking.

Keywords: Heart Rate · Wearable Devices · Ring · Wristband · Chestbelt · Statistical Analysis · Sleep Monitoring · Sleep Quality Recognition

1 Introduction

Personal health monitoring systems have recently received significant attention. They are capable of providing continuous and real time feedback to users about

A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 18–32, 2023.
https://doi.org/10.1007/978-3-031-34586-9_2

their health and daily behaviour [42]. Such systems rely on different physiological signals, such as, e.g., heart rate, to assess users' health state.

Improvements in sensors, battery and storage of wearable devices make them more powerful, affordable and pervasive. These improvements increase their ability to capture various physiological signals which help in return the development of personal health monitoring systems [44]. Such evolution encourages their employment in many health domains. Most wearables are capable of capturing heart rate (HR) traces, which can be employed in many health related applications like monitoring human stress [30,43], recovery after exercise [17] and sleep behaviour [32,44]. Changes in HR and heart rate variability (HRV) reflect autonomic nervous system patterns [44] and have been correlated with sleep stages [45,46], stress [31,52] and affect [47].

The availability of several health monitoring devices makes finding the most convenient device very challenging both for researchers and end users. The rapid development of wearables created a gap between the available devices and their evaluation studies [15,50]. Therefore, a comparison is needed to determine whether the sensor readings are interchangeable between devices, placed on different body positions. While there exist a few studies that investigated the measurements of wearable devices [37,50], it is not clear whether the raw physiological data are exchangeable and how such sensor measurements perform in downstream tasks. This understanding would allow researchers to make informed decisions regarding the use of such devices in data collection studies and users to choose the device that matches their needs without hampering the quality of the measurements.

In this paper, we investigate the interchangeability of HR signals obtained from three body positions, namely, finger, wrist and chest during sleep, since one of the wearables used (Oura ring) is dedicated and provides data continuously only during dormancy. To this goal, we run a data collection campaign in the wild, to gather physiological HR data – along with self reports about sleep behavior – using three well known devices: Oura ring (generation 3), Empatica E4 wristband and Polar chestbelt. We make the dataset available to other researchers upon request and signature of a data sharing agreement. Then, we assess the interchangeability of HR collected from wearables worn at different body locations. We extensively analyze the collected data using statistical measures as well as a sleep quality recognition task, to explore the interchangeability of HR measures at the level of raw data, time-domain features and classification capability. The main contributions of this paper are as follows:

- We collect and provide to the research community a dataset[1], named **HeartS**[2] collected from five participants over 30 days in their natural environments. The dataset contains heart rate data collected using three wearable devices – Oura ring (third generation), Empatica E4 wristband and Polar chestbelt – and self-reports regarding sleep and wake up times as well as sleep quality.

[1] Please contact the corresponding author of the paper to make a request regarding the dataset.

[2] Heart Rate from multiple devices and body positions for Sleep measurement.

- We perform extensive statistical analysis on the raw HR data and show a high correlation between the data of the three devices, suggesting their interchangeability.
- We extract common HR features from the three devices and show that the majority of the features have statistically negligible differences in small window sizes and small differences in large window size, which may impact machine learning tasks that use such features.
- We develop a machine learning pipeline and investigate the effect of the extracted features in sleep quality recognition. Our results confirm the interchangeability of the considered devices for this task.

This paper is structured as follows. Section 2 presents an overview of the similar studies in the literature; Sect. 3 describes the data collection procedure, while Sect. 4 shows the comparison between HR signals from wearable devices. Follows Sect. 5, which presents the analysis of HR features and the machine learning task adopted for the comparison between wearable devices, and Sect. 6, which describes the limitations and future work that can be addressed. Finally, Sect. 7 presents the conclusion.

2 Related Work

Several researchers evaluate the performance of wearable devices by analyzing the provided sleep parameters, e.g., Total Sleep Time (TST), Total Wake Time (TWT), Sleep Efficiency (SE), Wake After Sleep Onset (WASO) [37,50]. Such studies are either conducted in controlled settings [44,48], or in unrestricted ones [15,37,50].

Roberts et al. [44], for instance, conduct a comparison study between consumer wearables (Oura ring, Apple watch)[3], two actigraphy wristbands and Polysomnography (PSG), used as ground truth. They report that data from commercial multi-sensor wearables are highly correlated and can be adopted in sleep-wake classification problem, competing with research-grade devices. Scott et al. [48] perform a comparison study between a new commercial smart ring (THIM) (See Footnote 3), two popular wearables (Fitbit and Actiwatch) (See Footnote 3) versus PSG. Their results show no significant differences between PSG and THIM. Other researchers evaluate the sleep-wake recognition capabilities, but do not compare neither the features nor the raw physiological signals [44,48].

Stone et al. [50] compare nine sleep tracking consumer devices positioned on wrist, finger or mattress-affixed monitors, using as ground truth Electroencephalography (ECG). Using sleep parameters, they show that Fitbit Ionic and Oura ring have the highest accuracy and minimum bias in calculating the TST,

[3] **Oura Ring**: https://ouraring.com; **Apple watch**: https://www.apple.com/watch/; **THIM ring:** https://thim.io; **Fitbit:** https://www.fitbit.com/; **Actiwatch:** https://www.usa.philips.com/healthcare/sites/actigraphy; **Samsung Gear Sport watch:** https://www.samsung.com/us/watches/galaxy-watch4/.

TWT and SE; while with sleep staging metrics they find no accurate result from commercial devices. Mehrabadi et al. [37] compare sleep parameters (e.g., TST, SE and WASO) from the Oura ring and the Samsung Gear Sport watch (See Footnote 3) versus a medically approved actigraphy device. They found significant correlation of both devices with actigraphy. However, neither of [37,50] applied comparisons over physiological data.

Table 1 provides a overview of the recently conducted studies that compare different wearable devices, where we can observe that only two are publicly available. The previously mentioned studies focus on specific sleep parameters, showing their interchangeability between different devices [37,48]. However, there is a gap with regards to the analysis of physiological signals, collected from various devices, and how these differences can impact sleep quality recognition task.

Table 1. Description of existing studies that compare different wearable devices

Study	Study Settings	Number of Participants	Study Duration	Publicly Available
[15]	Home	21	7 nights	No
[38]	Laboratory	6	9 nights	Yes
[48]	Laboratory	25	1 night	No
[44]	Laboratory	8	4 nights	No
[37]	Home	45	7 nights	Yes
[50]	Home	5	98 nights	No

3 Data Collection Campaign

We conduct a data collection campaign using two commercial devices and one research-grade device, for the HeartS dataset. In this section, we describe the study participants, the adopted devices, the collected data and the data collection procedure. The study is reviewed and approved by our Faculty's delegate for Ethics.

3.1 Participants

We recruit five participants (three females and two males) of age from 24 to 29 years (avg: 26.2, std: 2.3). Participants wear, for 30 consecutive nights, three wearable devices: (1) The Oura ring (Generation 3) (See Footnote 3), which measures sleep with a Photoplethysmography (PPG) sensor, from which HR and HRV are extracted [3,14]; (2) The Empatica E4 wristband[4], which is a research-grade wristband that extracts HR via PPG sensor [49]; (3) The Polar chestbelt, equipped with an Electrocardiogram (ECG) sensor [28]. Both the Polar H07[5] and

[4] https://www.empatica.com/en-gb/research/e4/.

[5] https://support.polar.com/e_manuals/H7_Heart_Rate_Sensor/Polar_H7_Heart_Rate_Sensor_accessory_manual_English.pdf.

H10[6] releases are included in the study, since we have only two Polar chestbelt. E4 wristband and Polar chestbelt, along with previous generations of Oura ring, are adopted in several studies in the literature, e.g., in [3,5,6,11,12,14,26,27,50]. The devices contain also other sensors, for instance, the E4 is equipped with electrodermal activity, accelerometer and skin temperature sensors. In this study, we use only the HR measurements because it is the only common sensor in all the devices, which allows us to compare them.

3.2 Data Collection Procedure

To design the study and collect the data, we follow similar procedures to the literature (e.g., [26,45,50]). At the beginning of the data collection procedure, all participants sign an informed consent form. We provide the devices, pen-and-paper diaries, and the instructions needed to set up the designated synchronization applications for obtaining the raw data from devices. We instruct participants to wear the devices on their left hand, since small lateral differences might be present if choosing difference sides [1]. Every night all participants wear the Oura ring and the E4 wristband, whereas only two wear the Polar chestbelt (since we have only two Polar chestbelts available). The participants wear the devices one hour before sleep and log the bed-time. The next day, the participants complete the self report about the sleep quality of the previous night and the wake up time then take off the devices one hour after waking up. During the day, the participants synchronize the collected data during the previous night from each device and charge the devices. To make sure of the quality and quantity of the collected data, we systematically monitor the compliance with the data collection.

3.3 Collected Data

We collect two types of data, *physiological data* using three wearable devices and *self-reports* using pen-and-paper diaries described as follows.

Physiological Data. The Oura ring provides one HR data point every five minutes during sleep as well as the *bed-time start* and the *bed-time end*. Participants use the Oura mobile application to synchronize the collected data to the Oura cloud dashboard. The Empatica E4 wristband provides HR values every second. Participants use the E4 manager desktop application[7] to synchronize the collected data to the pre-created study on the E4 website. The Polar chestbelt integrates with a third party mobile application named Polar Sensor Logger[8] to provide an HR value per second. The application stores the collected data on the device.

[6] https://www.polar.com/en/sensors/h10-heart-rate-sensor.

[7] https://support.empatica.com/hc/en-us/articles/206373545-Download-and-install-the-E4-manager-on-your-Windows-computer.

[8] https://play.google.com/store/apps/details?id=com.j_ware.polarsensorlogger&hl=en\&gl=US.

We collect data of 105 sleep sessions. One participant did not wear the E4 wristband on the left hand so we discard the corresponding sessions to be consistent among the participants. In total we have 98 sessions. We collect 9,038 HR data points from the Oura ring which are equivalent to about 753 h of data. For the E4 wristband, we collect 3,192,319 points (about 886 h), with mean ($\pm$ standard deviation) 56.09 ± 14.58 bpm and 61.81 ± 12.42 bpm respectively. From the Polar chestbelt, we collect 656,038 HR data points, equivalent to approximately 182 h with mean 59.26 ± 12.23 bpm.

Self Reports. Participants use the pen-and-paper diaries to provide daily self reports about: their bed and wake up time, latency (i.e., the estimated time until the participant fall asleep), number of awakenings and sleep quality level every night, similar to [26,45]. They report sleep quality on a five level Likert scale [33]: *very poor, poor, normal, good, excellent* following [10]. One of the participants stopped logging self reports after the first week of the study. The dataset thus contains 80 sleep sessions labelled with the sleep behaviour.

4 Comparison of 5-Minutes Averaged HR Signals

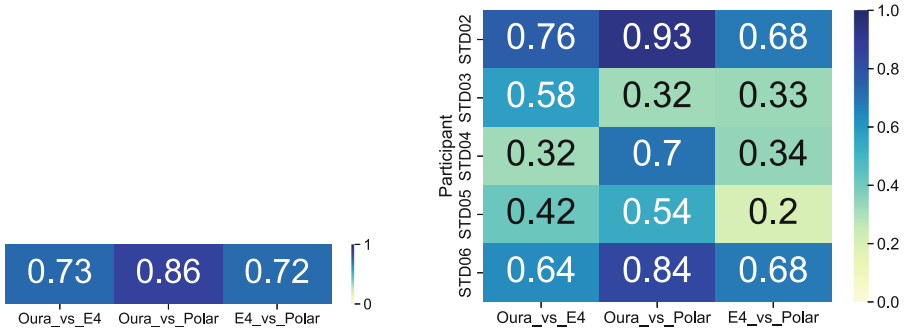
In this section we report the analysis performed using the HR signals collected as described in Subsect. 3.3. In particular, we describe the pre-processing steps, correlation and bias analysis.

4.1 Data Pre-processing

Since the Oura ring provides an average HR value every five minutes, for the current signal analysis, we down-sample the HR measurements to the same sampling frequency to obtain the same data granularity. In particular, we average the HR data of the E4 and Polar devices over five-minutes window. Given that Oura only provides the HR data during sleep, we use the *bed-time start* and *end* provided by Oura to define the sleep period and to segment the data of the E4 wristband and Polar chestbelt. We refer to the obtained traces as **averaged HR**.

4.2 Correlation Analysis

We use Shapiro-Wilk normality test to evaluate the parametric characteristic of the averaged HR [21]. We observe that the HR data, from all devices, is not normally distributed (p-value < 0.05). Based on that, we use Spearman's ρ rank correlation coefficient [35] to quantify the association between the HR signals. We conduct the analysis in two steps: first, we compute the correlation between each pair of devices using the averaged HR from all participants stacked together; then we compute the correlation using averaged HR, for each pair of device, per participant. Figure 1a shows the obtained correlation coefficients between averaged HR from every pair of devices. We find a high positive correlation



(a) Per devices Spearman's ρ correlation

(b) Per participant Spearman's ρ correlation

Fig. 1. Spearman's ρ correlation results between raw HR

between the averaged HR data from the three devices (>0.7), with the highest correlation between the Oura ring and the Polar chestbelt (0.86). All reported results are statistically significant, tested using an initial threshold of $\alpha = 0.05$ and Bonferroni [4] corrected to $\alpha_n = 0.01$, $n = 3$, as suggested in [25]. We also observe in Fig. 1b that correlation results by participant and device are similar to those by device (Fig. 1a), since they are all positive (>0.2). From this experiment, we conclude that there is a high positive correlation between averaged HR data from the three devices and the results are statistically significant ($\alpha = 0.05$, $\alpha_n = 0.003$, $n = 15$). The correlation analysis suggests interchangeability of the HR data across the three devices.

4.3 Bias Analysis

To assess the average difference between the devices, i.e., *bias* [15], we use a modified version of the Bland Altman Plot [7]. This plot measures the absolute difference between two distributions against the pair-wise averages. From the plot in Fig. 2, we observe that the data from Oura ring and the Polar chestbelt have the least average absolute difference (2.17), while the data from E4 wristband and the Polar chestbelt have the highest (5.01). These results confirm the higher correlation found between Oura's and Polar's HR data, as shown in Subsect. 4.2. In general, these results confirm the findings of the correlation analysis presented above.

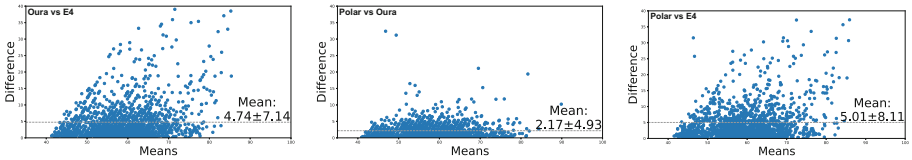


Fig. 2. Bias analysis results between every pair of devices. We report the average absolute differences and standard deviations.

5 Comparison of Features Extracted from Raw HR Signals

Many human activity recognition tasks rely on features extracted from the HR signals. This is in particular the case for sleep monitoring applications [32]. Thus, we extract time-domain HR features from each device using different window sizes and compare them. We execute the analysis in two steps. First, we analyze statistical differences in the extracted features. Then, we compare the performance obtained by machine learning (ML) classifiers for a sleep quality recognition task that use these features as input.

5.1 Data Pre-processing and Cleaning

In this part of data analysis, we rely on raw HR data collected from Oura ring and E4 wristband only. This is because we obtained only 18 sleep sessions for the Polar chestbelt, as opposed to approximately 80 for the other devices. We define *sleep sessions*, based on the *bed-time start* and *end* provided by Oura. We extract time-domain HR features, specifically *mean, standard deviation, range, median, variance, minimum, maximum, difference, slope*, over different window sizes, similarly to [26,39,46]. We employ three non-overlapping window sizes of 5, 10 and 60 min, and a window corresponding to the whole sleep session, similar to [25,39].

5.2 Effect Size Quantification of HR Features

To assess the difference between features extracted over each window from these devices, we employ Cliff's δ effect size [16], which allows us to determine the degree of difference between two samples. Cliff's δ values range between $[-1, 1]$, where 0 means that the two distributions are not different, while -1 and 1 indicate no distribution overlap [36]. We show the results in Fig. 3. We observe that for small window sizes, i.e., 5 and 10 min, most of the features show negligible or small differences. Also, it is noticeable that large differences are present with some Oura features due to its limited sampling rate in the designated windows. Such features rely on the data variability, e.g., the standard deviation is always 0 in a 5 min window for all Oura's data. By increasing the window size, we can

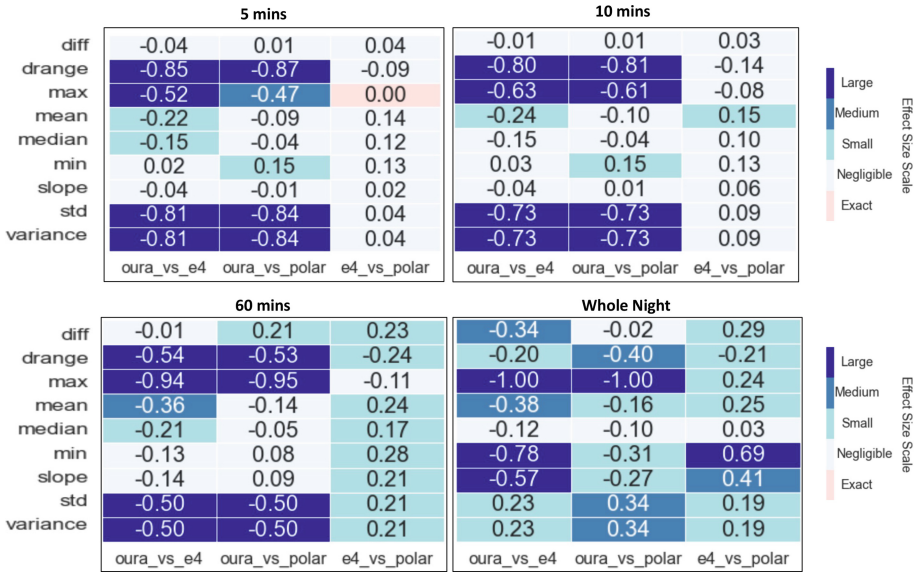


Fig. 3. Cliff's δ effect size or HR features over different window sizes

observe that the majority of the features have small differences and the differences in the features that require data variability decrease. Such observations motivate the next experiment, where we evaluate the impact of the features' differences on a sleep recognition task. These results, similarly to the correlation experiment results in Sect. 4, suggest limited differences and interchangeability between the devices.

5.3 Sleep Quality Recognition Task

In this part of the analysis, we detail the comparison between devices in a machine learning task, with the aim of assessing the impact of the observed differences in the HR features from the devices. We employ a sleep quality recognition task, given that Oura ring is indeed dedicated to sleep behavior monitoring. As ground truth, we used the subjective sleep quality scores from the self reports, described in Subsect. 3.3.

Classification Procedure. We define this problem as a binary classification task, using normalised sleep quality labels (range [0,1]), with a threshold of 0.5 to discriminate between positive (high) and negative (low) class, similar to [26,54]. From this we obtain the following distribution: 61% of the data on the positive class and 39% on the negative class. Since the data is not completely balanced, we employ, only during training, synthetic minority over-sampling technique (SMOTE) [13]. Since the collected data provides one score per sleep session, when using 5, 10 and 60 min windows we train by assigning the label to each window. However, at validation time, we evaluate only one score per sleep sessions: to

obtain this, we apply majority voting over the window-level predictions. When training using the whole sleep session, majority voting is not used.

Classification Models. We adopt 10 classification models in our experiments: Decision Tree (DT) [51], Gaussian Naïve Bayes (NB) [20], Support Vector Machine (SVM) [19], Multilayer Perceptron (MLP) [24], k-nearest neighbour (KNN) [41], Random Forest (RF) [8], XGBoost [23], AdaBoost [22], Quadratic Discriminant Analysis (QDA) [29] and Gaussian Process (GP) [53]. We use the implementation of such algorithms from the Scikit-Learn Python library [40].

Evaluation Methodology. We perform evaluation of the chosen models using two *cross validation* paradigms. We first evaluate the models' capability to generalize to a new participant, this is achieved with *Leave One Participant Out* cross validation, in which we train each model using all but one user's data, and use the remaining user's data as test set. Second, we evaluate the ability of the models to recognize the sleep quality score for an already existing participant using the *Leave One Session Out* cross validation. This procedure allows to train on all sleep sessions but one, and test on the remaining one. We use two baseline classifiers, to identify if our models are capable of learning patterns from the input data [34]. The first baseline is the Random Guess classifier (RG), which makes sleep quality predictions by extracting randomly the positive and negative labels from a uniform distribution. The second baseline, denominated "a-priori", always predict a constant value, chosen as the majority class, which in this case is the positive class. We adopt the *balanced accuracy* as the evaluation metric for the experiments, given the imbalance in the class distributions when testing [9]. We compare the performance of these baseline classifiers with the other models using the Wilcoxon signed-rank statistical significance test [18,21] with a threshold of 0.05.

Classification Results. For the *Leave One Participant Out* cross validation, we show results in Table 2. While small variations in performance are present across window size and device, all classifiers do not achieve accuracies higher than 0.65, with most models not higher then the baselines (0.5 RG and 0.6 a-priori). Indeed, only one model (AdaBoost, whole night, E4) achieves a balanced accuracy higher than 0.6. However, all models are not statistically different (p-value threshold $\alpha = 0.05$) from the a-priori baseline. The results suggest that both interpersonal variability and the limited number of participants do not allow to achieve significant performance, with respect to the baselines, when testing on an unseen participant [2].

For the *Leave Out Session Out* cross validation paradigm, we report results in Table 3. From these, we see that models trained on the whole night achieve a lower performance than models trained over smaller windows. The results show that for the 5 min window most of the models are able to recognize the sleep quality for an already existing participant, surpassing the baselines (>0.62). A model trained on the 60 min window, using Oura features, has the highest overall average accuracy (0.76). For the whole night, one of the models can reach a performance of 0.66 using the E4 features. Accordingly, the performance

Table 2. Average balanced accuracy, with standard errors, for three devices in sleep quality recognition task using different window size and *Leave One Participant Out* cross validation. The a-priori baseline always predicts the positive class, while the Random Guess (RG) uses a uniform distribution to make predictions

Window size Device/ Model	5 mins		10 mins		60 mins		whole night	
	Oura	E4	Oura	E4	Oura	E4	Oura	E4
DT	0.53 ± 0.13	0.45 ± 0.05	0.46 ± 0.03	0.50 ± 0.09	0.44 ± 0.05	0.43 ± 0.09	0.42 ± 0.07	0.48 ± 0.03
NB	0.49 ± 0.01	0.50 ± 0.00	0.47 ± 0.03	0.49 ± 0.01	0.42 ± 0.05	0.47 ± 0.03	0.45 ± 0.02	0.47 ± 0.03
SVM	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.51 ± 0.01	0.45 ± 0.05	0.50 ± 0.00	0.5 ± 0.03	0.50 ± 0.00
MLP	0.50 ± 0.00	0.47 ± 0.03	0.50 ± 0.00	0.49 ± 0.02	$\mathbf{0.55 \pm 0.05}$	$\mathbf{0.52 \pm 0.02}$	0.42 ± 0.06	0.54 ± 0.07
RF	0.46 ± 0.04	0.49 ± 0.01	0.52 ± 0.03	$\mathbf{0.55 \pm 0.07}$	0.42 ± 0.06	0.41 ± 0.06	0.36 ± 0.06	0.43 ± 0.05
XGBoost	$\mathbf{0.55 \pm 0.02}$	0.48 ± 0.02	0.52 ± 0.03	0.50 ± 0.04	0.52 ± 0.02	0.52 ± 0.03	0.38 ± 0.07	0.5 ± 0.05
AdaBoost	0.48 ± 0.06	0.49 ± 0.04	$\mathbf{0.53 \pm 0.05}$	0.50 ± 0.03	0.53 ± 0.06	0.51 ± 0.05	0.43 ± 0.05	$\mathbf{0.64 \pm 0.05}$
QDA	0.54 ± 0.06	0.5 ± 0.00	0.49 ± 0.01	0.45 ± 0.06	0.49 ± 0.01	0.51 ± 0.04	$\mathbf{0.51 \pm 0.07}$	0.55 ± 0.04
KNN	0.53 ± 0.04	$\mathbf{0.53 \pm 0.05}$	0.44 ± 0.04	0.51 ± 0.03	0.50 ± 0.07	0.45 ± 0.03	0.41 ± 0.1	0.45 ± 0.03
GP	0.48 ± 0.02	0.49 ± 0.01	0.5 ± 0.01	0.45 ± 0.04	0.50 ± 0.04	0.49 ± 0.11	0.45 ± 0.1	0.47 ± 0.1
RG	0.39 ± 0.10		0.62 ± 0.04		0.44 ± 0.09		0.62 ± 0.06	
a-priori	0.50 ± 0.0		0.50 ± 0.0		0.50 ± 0.0		0.5 ± 0.0	

of the models diminishes when using the whole sleep session, compared to smaller window sizes. The results also suggest that there is no real advantage between models trained with data from Oura ring or E4 wristband, with all window experiments achieving accuracies higher then 0.7 with at least one model (best baseline 0.61). From the experiments, we can conclude that both devices achieve comparable performance in the sleep quality recognition task. With our results in Sect. 4 and Subsect. 5.2, these classification task supports that the devices used are interchangeable with respect to heart rate data. We also find that it is better to adopt a windowed data, as opposed to using the whole night session, when performing sleep quality recognition. However, it is worth noting how, given the limited amount of data, the standard errors evaluated are quite large. This means that no result is statistically significant with respect to the a-priori baseline, increasing the available data would allow to mitigate this problem.

6 Limitations and Future Work

The main limitation of our work is the small number of participants in the dataset (five). We also only collected an average of 16 nights per participants. In future work, performing a data collection with more participants and for more nights could lead to further insights. This is especially true for the Polar chestbelt, since we do not use in the subjective sleep recognition task given the limited number (18) of sleep sessions collected with this device. The use of more HR tracking wearable devices could be explored. As suggested by [2], the use of a subjective sleep quality score can also hinder a machine learning task, as such we are considering exploring the devices performance compared to an additional objective measure.

Table 3. Average balanced accuracy, with standard errors, for three devices in sleep quality recognition task using different window size and *Leave One Session Out* cross validation. The a-priori baseline always predicts the positive class, while the Random Guess (RG) uses a uniform distribution to make predictions.

Window size Device/ Model	5 mins		10 mins		60 mins		whole night	
	Oura	E4	Oura	E4	Oura	E4	Oura	E4
DT	0.62 ± 0.05	0.71 ± 0.05	0.71 ± 0.05	$0.71 \pm .05$	0.74 ± 0.05	0.68 ± 0.05	0.53 ± 0.05	0.60 ± 0.05
NB	0.71 ± 0.05	0.68 ± 0.05	0.71 ± 0.05	0.68 ± 0.05	0.60 ± 0.06	0.66 ± 0.05	0.54 ± 0.05	0.63 ± 0.05
SVM	0.50 ± 0.06	0.53 ± 0.06	0.42 ± 0.06	0.53 ± 0.06	0.51 ± 0.06	0.44 ± 0.06	0.52 ± 0.05	0.55 ± 0.05
MLP	0.57 ± 0.06	0.64 ± 0.05	0.56 ± 0.06	0.59 ± 0.06	0.59 ± 0.06	$\mathbf{0.72 \pm 0.05}$	0.55 ± 0.05	0.65 ± 0.05
RF	0.69 ± 0.05	$\mathbf{0.72 \pm 0.05}$	$\mathbf{0.72 \pm 0.05}$	$\mathbf{0.74 \pm 0.05}$	0.74 ± 0.05	0.65 ± 0.05	0.61 ± 0.05	$\mathbf{0.66 \pm 0.05}$
XGBoost	0.66 ± 0.05	$\mathbf{0.72 \pm 0.05}$	$\mathbf{0.72 \pm 0.05}$	0.72 ± 0.05	$\mathbf{0.76 \pm 0.05}$	0.65 ± 0.05	0.60 ± 0.05	0.63 ± 0.05
AdaBoost	$\mathbf{0.72 \pm 0.05}$	$\mathbf{0.72 \pm 0.05}$	$\mathbf{0.72 \pm 0.05}$	0.72 ± 0.05	0.64 ± 0.05	$\mathbf{0.72 \pm 0.05}$	0.51 ± 0.05	0.63 ± 0.05
QDA	0.64 ± 0.05	0.64 ± 0.05	0.59 ± 0.06	0.62 ± 0.05	0.50 ± 0.06	0.64 ± 0.05	0.50 ± 0.05	0.52 ± 0.05
KNN	0.66 ± 0.05	0.70 ± 0.05	0.71 ± 0.05	0.68 ± 0.05	0.66 ± 0.05	0.65 ± 0.05	0.52 ± 0.05	0.59 ± 0.05
GP	0.70 ± 0.05	0.72 ± 0.05	0.71 ± 0.05	0.71 ± 0.05	0.70 ± 0.05	0.55 ± 0.06	$\mathbf{0.64 \pm 0.05}$	0.59 ± 0.05
RG	0.61 ± 0.05		0.57 ± 0.06		0.39 ± 0.05		0.41 ± 0.05	
a-priori	0.61 ± 0.05		0.61 ± 0.06		0.61 ± 0.05		0.59 ± 0.05	

7 Conclusion

We run a data collection campaign for 30 nights to collect HR data during sleep using Oura ring, Empatica E4 wristband and Polar chestbelt, in the wild, along with self reports about sleep behaviour. We provide the dataset to other researchers upon request to extend our data analysis. Then, we investigate the interchangeability of HR data collected from these wearables. To this goal, we run an extensive data analysis. We find that there is a high positive correlation between the HR data from the three devices based on Spearman's correlation coefficient. Using bias analysis, we also estimate that the Oura ring's HR signal has less variations with respect to the ECG-based Polar chestbelt, compared to data from the E4 wristband. We also assess the difference between time-domain features extracted from the three devices for different windows sizes, finding them negligible or small in most cases. Finally in order to evaluate the impact of such small differences, we employ these features in a machine learning task to predict subjective sleep quality. We find that, when testing on a new sleep session, there is not appreciable difference between models trained on features extracted from Oura ring's or E4 wristband's HR signals. We also find that a higher performance is achieved when separating the sleep session into non-overlapping windows, as opposed to using the whole night's data. In conclusion, our results suggest interchangeability among the devices. Even with the outlined limitations of our study, we believe that the three devices can be used in broader settings, e.g., health tracking, with similar outcomes.

Acknowledgement. This contribution is supported by the Swiss National Science Foundation (SNSF) through the grant 205121 _197242 for the project "PROSELF: Semi-automated Self-Tracking Systems to Improve Personal Productivity". Shkurta Gashi is supported by an ETH AI Center postdoctoral fellowship.

References

1. Alchieri, L., et al.: On the impact of lateralization in physiological signals from wearable sensors (2022)
2. Alecci, L., et al.: On the mismatch between measured and perceived sleep quality. In: Proceedings of the 2022 UbiComp (2022). https://doi.org/10.1145/3544793.3563412
3. Altini, M., et al.: The promise of sleep: a multi-sensor approach for accurate sleep stage detection using the oura ring. Sensors **21**(13) (2021)
4. Armstrong, R.A.: When to use the B onferroni correction. Ophthalmic Physiol. Opt. **34**(5) (2014)
5. Assaf, M., Rizzotti-Kaddouri, A., Punceva, M.: Sleep detection using physiological signals from a wearable device. In: Inácio, P.R.M., Duarte, A., Fazendeiro, P., Pombo, N. (eds.) HealthyIoT 2018. EICC, pp. 23–37. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-30335-8_3
6. Barika, R., et al.: A smart sleep apnea detection service. In: 17th International Conference on CM. The British Institute of NDT (2021)
7. Bland, J.M., et al.: Measuring agreement in method comparison studies. Stat. Methods Med. Res. **8**(2) (1999)
8. Breiman, L.: Random forests. Mach. Learn. **45**(1) (2001)
9. Brodersen, K.H., et al.: The balanced accuracy and its posterior distribution. In: 20th ICPR. IEEE (2010)
10. Buysse, D.J., et al.: The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res. **28**(2), 193–213 (1989)
11. Cakmak, A.S., et al.: An unbiased, efficient sleep-wake detection algorithm for a population with sleep disorders: change point decoder. Sleep **43**(8) (2020)
12. Carlozzi, N.E., et al.: Daily variation in sleep quality is associated with health-related quality of life in people with spinal cord injury. Arch. Phys. Med. Rehabil. **103**(2) (2022)
13. Chawla, N.V., et al.: Smote: synthetic minority over-sampling technique. JAIR **16** (2002)
14. Chee, N.I., et al.: Multi-night validation of a sleep tracking ring in adolescents compared with a research actigraph and polysomnography. Nat. Sci. Sleep **13** (2021)
15. Chinoy, E.D., et al.: Performance of four commercial wearable sleep-tracking devices tested under unrestricted conditions at home in healthy young adults. Nat. Sci. Sleep **14** (2022)
16. Cliff, N.: Dominance statistics: ordinal analyses to answer ordinal questions. Psychol. Bull. **114**(3), 494 (1993)
17. Cole, C.R., Blackstone, E.H., Pashkow, F.J., Snader, C.E., Lauer, M.S.: Heart-rate recovery immediately after exercise as a predictor of mortality. N. Engl. J. Med. **341**(18), 1351–1357 (1999)
18. Conover, W.J.: Practical Nonparametric Statistics, vol. 350. Wiley, Hoboken (1999)
19. Cortes, C., et al.: Support-vector networks. Mach. Learn. **20**(3) (1995)
20. Duda, R.O., et al.: Pattern Classification and Scene Analysis, vol. 3. Wiley, New York (1973)
21. Field, A., et al.: How to Design and Report Experiments. Sage (2002)
22. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)

23. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stati. (2001)
24. Gardner, M.W., et al.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. **32**(14–15) (1998)
25. Gashi, S., et al.: Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**(1), 1–19 (2019)
26. Gashi, S., et al.: The role of model personalization for sleep stage and sleep quality recognition using wearables. IEEE Pervasive Comput. **21**, 69–77 (2022)
27. Ghorbani, S., et al.: Multi-night at-home evaluation of improved sleep detection and classification with a memory-enhanced consumer sleep tracker. Nat. Sci. Sleep **14** (2022)
28. Gilgen-Ammann, R., et al.: RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. EJAP **119** (2019)
29. Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2. Springer, Heidelberg (2009)
30. Hellhammer, J., et al.: The physiological response to trier social stress test relates to subjective measures of stress during but not before or after the test. Psychoneuroendocrinology **37**(1), 119–124 (2012)
31. Hernandez, J., Morris, R.R., Picard, R.W.: Call center stress recognition with person-specific models. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011. LNCS, vol. 6974, pp. 125–134. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24600-5_16
32. Imtiaz, S.A.: A systematic review of sensing technologies for wearable sleep staging. Sensors **21**(5) (2021)
33. Joshi, A., et al.: Likert scale: explored and explained. Br. J. Appl. Sci. Technol. **7**(4) (2015)
34. Kelleher, J.D., Mac Namee, B., D'arcy, A.: Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press (2020)
35. Kendall, M.G., et al.: The Advanced Theory of Statistics. The Advanced Theory of Statistics, 2nd edn (1946)
36. Kromrey, J.D., et al.: Analysis options for testing group differences on ordered categorical variables: an empirical investigation of type I error control and statistical power. MLRV **25**(1) (1998)
37. Mehrabadi, M.A., et al.: Sleep tracking of a commercially available smart ring and smartwatch against medical-grade actigraphy in everyday settings: instrument validation study. JMIR mHealth uHealth **8**(11) (2020)
38. Miller, D.J., et al.: A validation study of a commercial wearable device to automatically detect and estimate sleep. Biosensors **11**(6) (2021)
39. Min, J.K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., Hong, J.I.: Toss'n'turn: smartphone as sleep and sleep quality detector. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 477–486 (2014)
40. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. JMLR **12** (2011)
41. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
42. Raskovic, D., et al.: Medical monitoring applications for wearable computing. Comput. J. **47**(4), 495–504 (2004)

43. Reinhardt, T., et al.: Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST). Psychiatry Res. **198**(1), 106–111 (2012)
44. Roberts, D.M., et al.: Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography. Sleep **43**(7) (2020)
45. Sano, A., et al.: Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In: Proceedings of the IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN 2015). IEEE (2015)
46. Sano, A., et al.: Multimodal ambulatory sleep detection using LSTM recurrent neural networks. IEEE J. Biomed. Health Inform. **23**(4), 1607–1617 (2019)
47. Schmidt, P., Reiss, A., Dürichen, R., Van Laerhoven, K.: Wearable-based affect recognition—a review. Sensors **19**(19), 4079 (2019)
48. Scott, H., et al.: The development and accuracy of the THIM wearable device for estimating sleep and wakefulness. Nat. Sci. Sleep **13** (2021)
49. Siirtola, P., et al.: Using sleep time data from wearable sensors for early detection of migraine attacks. Sensors **18**(5) (2018)
50. Stone, J.D., et al.: Evaluations of commercial sleep technologies for objective monitoring during routine sleeping conditions. Nat. Sci. Sleep **12** (2020)
51. Swain, P.H., et al.: The decision tree classifier: design and potential. IEEE Trans. Geosci. Electron. **15**(3) (1977)
52. Taylor, S.A., et al.: Personalized multitask learning for predicting tomorrow's mood, stress, and health. IEEE Trans. Affect. Comput. **11**, 200–213 (2017)
53. Williams, C.K., et al.: Gaussian Processes for Machine Learning, vol. 2. MIT Press, Cambridge (2006)
54. Yan, S., et al.: Estimating individualized daily self-reported affect with wearable sensors. In: 2019 IEEE ICHI (2019)