Article

Not peer-reviewed version

From Pulses to Sleep Stages: Towards Optimised Sleep Classification Using Heart-Rate Variability

Pavlos I. Topalidis , Sebastian Baron , Dominik P. J. Heib , Esther-Sevil Eigl , Alexandra Hinterberger , Manuel Schabus *

Posted Date: 10 October 2023

doi: 10.20944/preprints202310.0592.v1

Keywords: automatic sleep-staging; machine learning; CNN; wearables; Polar; hear-rate varaiblity; Inter-beat intervals

Article

From Pulses to Sleep Stages: Towards Optimised Sleep Classification Using Heart-Rate Variability

Pavlos I. Topalidis [1], Sebastian Baron [2,4], Dominik P. J. Heib [1,3], Esther-Sevil Eigl [1], Alexandra Hinterberger [1] and Manuel Schabus [1,*]

[1] Laboratory for Sleep, Cognition and Consciousness Research, Department of Psychology and Centre for Cognitive Neuroscience Salzburg (CCNS), Paris-Lodron University of Salzburg, Austria;
[2] Department of Mathematics, Paris-Lodron University of Salzburg, Austria;
[3] Institut Proschlaf, Austria;
[4] Department of Artificial Intelligence and Human Interfaces (AIHI), Paris-Lodron University of Salzburg, Austria;
* Correspondence: manuel.schabus@sbg.ac.at; Tel.: +43-662-8044-5113.

Abstract: More and more people quantify their sleep using wearables and are getting obsessed in their pursuit of optimal sleep ("orthosomnia"). However, it is criticized that many of these wearables are giving inaccurate feedback and even can lead to negative daytime consequences. Acknowledging these facts, we here aim to extend previous findings Topalidis *et al.* [1] in a new sample of 136 self-reported poor sleepers by implementing optimization procedures to minimize erroneous classification when ambulatory sensing sleep. Firstly, here, we introduce an advanced interbeat-interval (IBI) quality control using a random forest method to account for wearable recordings in naturalistic and more noisy settings. We further aim to improve sleep classification by opting for a loss function model instead of overall epoch-by-epoch accuracy to avoid model biases towards the majority class (i.e., "light sleep"). Using these implementations, we compare the classification performance between the optimized (loss function model) and the accuracy model. We use signals derived from PSG, one-channel ECG, and two consumer wearables: the ECG breast belt Polar® H10 (H10) and the Polar® Verity Sense (VS), an optical Photoplethysmography (PPG) heart-rate sensor. Results reveal high overall accuracy for the new model for ECG (86.3 %, κ=.79), H10 (84.4%, κ=.76), and VS (84.2%, κ=.75) with intended improvements in deep sleep and wake. In addition, the new optimized model displays moderate to high correlations and agreement with PSG on primary sleep parameters, while measures of reliability, expressed in intra-class correlations, suggest excellent reliability for most sleep parameters. Finally, it is demonstrated that the new model is still classifying sleep accurately in 4-classes in users taking heart-affecting and/or psychoactive medication which can be considered a prerequisite for use also in high age groups and/or with common disorders. Further improving and validating sleep stage classification algorithms with affordable wearables may resolve existing scepticism and open the door for such approaches in clinical practice.

Keywords: automatic sleep-staging; machine learning; CNN; wearables; Polar; hear-rate varaiblity; Inter-beat intervals

1. Introduction

As sleep is an important factor related to health, disease, overall quality of life and peak performance [2], more and more people monitor it using wearable devices. The pursuit of good sleep can sometimes go too far, creating an obsession with monitoring and quantifying sleep that can result in increased stress and discomfort [3], a condition termed orthosomnia (i.e., from the Greek "ortho" meaning straight or upright and the Latin "somnia"). Many users of wearables have been erroneously led to believe, however, that commercial wearables can accurately and reliably measure sleep, even though almost all of them lack scientific sound and/or independent validation studies [4] against the gold standard, Polysomnography (PSG), a combination of electroencephalography (EEG), electrooculography (ECG) and electromyography (EMG).

Erroneous feedback about one's sleep (e.g., substantial underestimation of deep sleep, or wrong wake classification at sleep onset or during the night) can yet have serious adverse effects enhancing people's misperception and even having negative daytime consequences [5]. Such wrong feedback on wearables may also lead to inappropriate suggestions for adjusting sleep habits and work against the aim of promoting better sleep health [6]. This is particularly worrisome for people with sleep problems and preoccupations who are especially sensitive to feedback on their sleep [7–9]. The potential adverse effects of inaccurate feedback in combination with the limited rigour of many validation studies against the PSG gold standard certainly justify the scepticism around the broad use of wearables in the clinical field [10]. However, the potential benefits of using accurate wearable devices that capture daily sleep changes in natural environments and outside of the laboratory, combined with low costs, are undeniable. Especially in light of recent studies showing that implementing such technologies together with sleep intervention protocols can have a beneficial effect on therapy outcomes [11–14]. It is our opinion that soon such technologies if optimized and carefully validated, will play a central role in research and clinical practice as they allow continuous sleep measurements (and feedback) in ecologically valid home environments and at affordable cost.

However, only a few of the wearable technologies that provide multiclass epoch-by-epoch sleep classification have been transparent considering their sensing technology and algorithms used, and are indeed validated against PSG in suitable large (and heterogenous) samples, using appropriate performance metrics (e.g., Cohes's κ, sensitivity and specificity, F1) rather than solely relying on mean values per nights or overall epoch-by-epoch accuracies across sleep stages. Among the few, Chinoy *et al.* [15] has recently compared the performance of seven consumer sleep-tracking devices to PSG and reported that the reviewed devices displayed poor detection of sleep stages compared to PSG, with consistent under- or overestimation of the amount of REM or deep sleep. Chee *et al.* [16] validated two widely used sensors, the Oura ring (v. 1.36.3) and Actiwatch (AW2 v. 6.0.9), against PSG in a sample of 53 participants with multiple recordings each. Compared to PSG, the Oura ring displayed an average underestimation of about 40 minutes for TST, 16 minutes for REM sleep, and 66 minutes for light sleep, while it overestimated Wake After Sleep Onset (WASO) by about 38 minutes and deep (N3) sleep by 39 minutes. Another study Altini and Kinnunen [17] examined 440 nights from 106 individuals wearing the Oura ring (v. Gen2M) and found a good 4-class overall accuracy of 79% when including various autonomic nervous system signals, such as heart-beat variability, temperature, acceleration and circadian features.

We have recently developed a 4-class sleep stage classifications model (wake, light, deep, REM) that reaches up to 81% accuracy and a κ of 0.69 [1], when using only data from low-cost Heart Rate Variability (HRV) sensors. Although such classification accuracies are approaching expert inter-rater reliability [1], there are edge cases where classification is particularly difficult (e.g, longer periods of rest and absent movement while reading as compared to falling asleep) and prone to errors and may result in erroneous feedback to the users. We suggest that to deal with these cases it is first important to implement advanced signal quality controls, that detect artefacts related to sensor detachment, or excessive movements during sleep that lead to insufficient signal quality (i.e. inter-beat interval estimations) for meaningful classification. In fact, some of the previously suggested models developed for one-channel ECG recordings (e.g., Mathunjwa *et al.* [18]; Habib *et al.* [19]; Sridhar *et al.* [20]) have already addressed such issues by implementing individual normalisation and simple outlier corrections (e.g. linearly interpolating too short/long inter-beat intervals). In Figure 1 we illustrate an example from our data and show how bad IBI signal quality can lead to erroneous sleep stage classification, and suggest incorporating advanced machine learning approaches for bad signal detection and better classification accuracy.

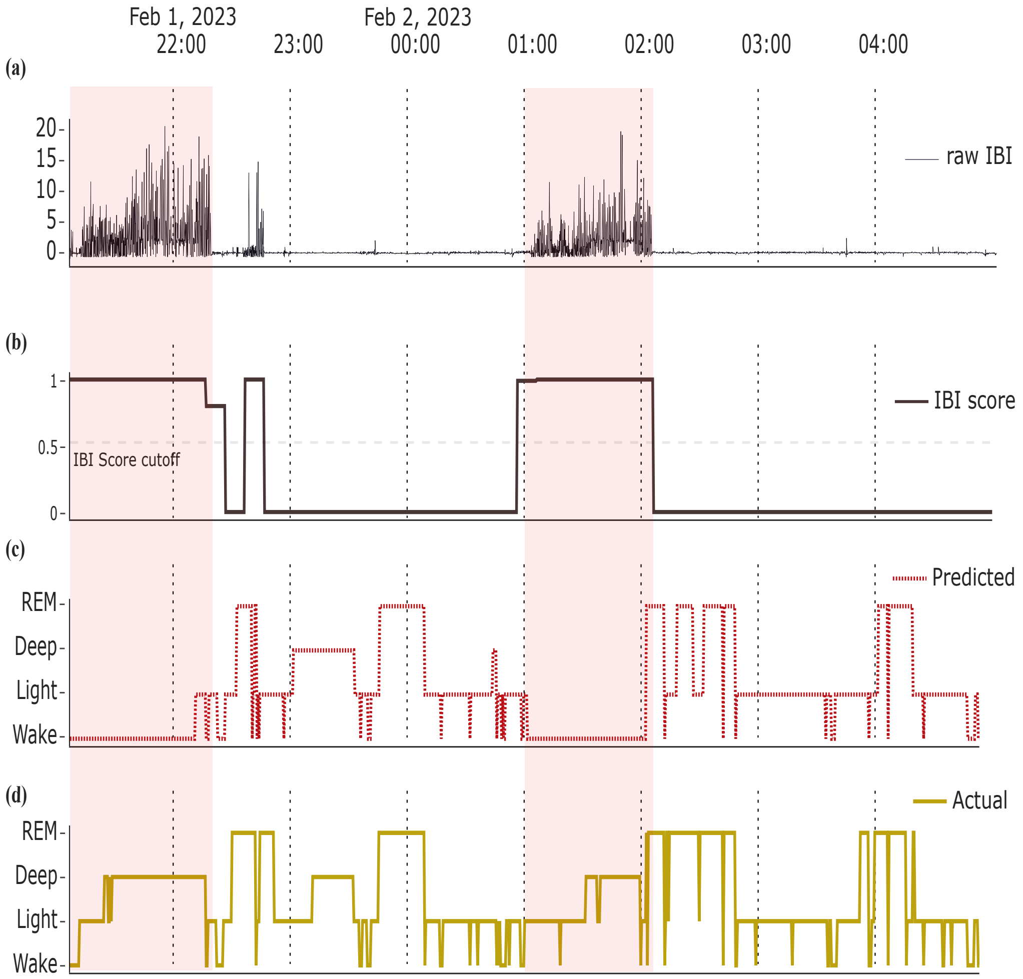


Figure 1. Bad IBI signal, if not detected, can lead to erroneous sleep stage classification. There are epochs in (**a**) the raw IBI signal that can be identified using (**b**) an advanced IBI signal quality control procedure based on a trained random forest model. Due to bad signal quality, sleep staging these epochs results in erroneous classification (**c**), which misrepresents the ground truth (**d**).

Additionally, to optimize sleep classification of edge cases we need to take into consideration the skewness in the distribution of sleep phases across the night (i.e., a disproportionate amount of sleep stages across the night with light sleep dominating). This becomes evident when one acknowledges that algorithms can reach epoch-by-epoch classification accuracies of up to 65-70% when an algorithm simply classifies "light sleep" (N1 and N2 [21]) throughout the night. A model that is trained for optimal overall classification accuracy, like the one suggested in [1], will display a bias towards the majority class (light sleep), resulting in poorer performance on less populated classes, such as wake and deep sleep. It is however crucial for the user that specific periods of waking and deep sleep are not misclassified as this substantially decreases the user's trust in the sleep analysis and consequently any potential sleep intervention. We suggest a model that opts for the minimum weighted cross-entropy loss function value, that encapsulates the skewness of the sleep stage distribution, thus resulting in unbiased classification. Figure 2 illustrates the bias of the accuracy model towards the majority class and how opting for a loss function model resolves this issue.

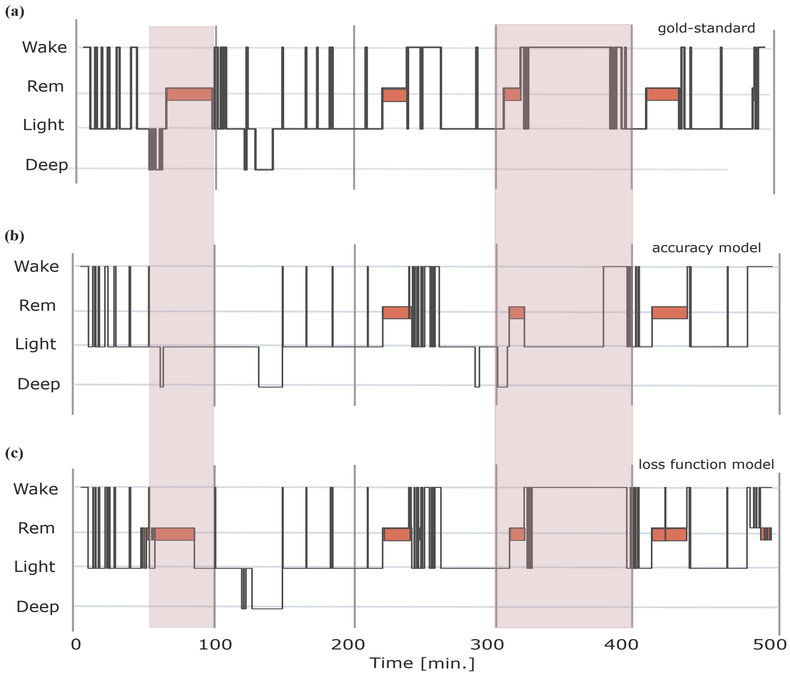


Figure 2. Example of a night where the accuracy model overestimates of majority class (light sleep). Panel (**a**) displays the actual PSG-based hypnogram as sleep staged automatically using the G3 sleepware gold-standard. Panel (**b**) displays sleep staging using the "accuracy model", while panel (**c**) displays the "loss function model". Note that the accuracy model displays a bias towards the majority class (e.g., epochs marked in red shading), as it strives to maximize overall classification accuracy, especially in cases where the model is unsure. In contrast, the loss function incorporates the skewed class distribution using categorical cross entropy weighted by class, correcting thus for a bias towards the majority class. In this example, both predicted hypnograms (**a;b**) use signals derived from the H10 sensor.

We here explore the benefits of applying an IBI-correction procedure, as well as opting for the loss function in classifying sleep. We tested our updated model on a new sample of 136 subjects, whose sleep was recorded using ambulatory PSG, in their homes, for one or more nights. As previously we are using an i) ECG gold-standard (providing IBIs), a ii) breast-worn belt measuring IBIs (Polar® H10) as well as and iii) a pulse-to-pulse interval device worn on the upper arm (Polar® Verity Sense -VS) as input signals. We put particular emphasis on the upper arm VS as it is a more comfortable sensor to sleep with than the H10 breast belt, without any notable compromise on sleep classification. Finally, we assess whether wearables such as the ones tested here are even accurate in classifying sleep in users with health issues and who, as a consequence, are taking central nervous system medication (psychoactive substances) and/or medication with expected effects on the heart (beta blockers, etc).

2. Methods

2.1. Participant

We recorded ambulatory PSG from 136 participants (female = 40; Mean Age = 45.29 SD = 16.23, range = 20 - 76) who slept in their homes, for one or more nights. In total (265) nights of recordings included the gold standard PSG and ECG. From these participants, 112 wore a Polar® H10 chest sensor (see materials) and 99 participants wore the Polar® Verity Sense, with 178 and 135 Nights respectively. This sample was part of an ongoing study, investigating the effects of online sleep training programs on participants with self-reported sleep complaints, which included participants with no self-reported acute mental or neurological disorders, capable of using smartphones. While we did not set strict exclusion criteria our sample naturally comes mainly with people having sleep difficulties. Specifically, 84.8% of our sample had a Pittsburgh Sleep Quality Index (PSQI) score above 5, with an average score of 9.36 (SD=3.21). For a subset of participants, we took a medical history and grouped them with those who were on psychoactive and/or heart medication (with medication, N= 17, Nnights= 40), and those with no medication (without medication, N= 39, Nnights= 87). The study was conducted according to the ethical guidelines of the Declaration of Helsinki.

2.2. Materials

We recorded PSG using the ambulatory varioport 16-channel EEG system (Becker Meditec®) with gold cup electrodes (Grass Technologies, Astro – Med GmbH®), according to the international 10-20 system, at frontal (F3, Fz, F4), central (C3, Cz, C4), parietal (P3, P4) and occipital (O1, O2) derivations. We recorded EOG using two electrodes (placed 1cm below the left outer canthus and 1cm above the right outer canthus, respectively), and the chin muscle activity from two EMG electrodes. Importantly, we recorded the ECG signal with two electrodes that we placed below the right clavicle and on the left side below the pectoral muscle at the lower edge of the left rib cage. Before actual sleep staging, we used the BrainVision Analyzer 2.2 (Brain Products GmbH, Gilching, Germany) to re-reference the signal to the average mastoid signal, filter according to the American Academy of Sleep Medicine (AASM) for routine PSG recordings (AASM Standards and Guidelines Committee, 2021) and downsampled to 128 Hz (original sampling rate was at 512 Hz). Sleep was then automatically scored in 30-second epochs using standard AASM scoring criteria, as implemented in the Sleepware G3 software (Sleepware G3, Koniklijke Philips N.V. Eindhoven, The Netherlands). The G3 software is considered to be non-inferior to manual human staging and can be readily used without the need for manual adjustment [22]. All PSG recordings in the current analysis have been carefully manually inspected for severe non-physiological artefacts in the EEG, EMG as well and EOG as such effects would render our PSG-based classification (serving as the gold standard) less reliable.

Having conducted extensive research on multiple sensors, we decided on two Polar® (Polar® Electro Oy, Kempele, Finland) sensors as they came with the most accurate signal as compared to gold-standard ECG [23–25]: the H10 chest strap (Model: 1W) and the Verity Sense (VS, Model: 4J). Next to good signal quality, both sensors have good battery life (about 400 hours for H10, 30 hours for VS Sense), and low overall weight and volume, which makes them comfortable to sleep with.

2.3. Data synchronization and missing data handling

For each recording, sleep staging was computed using PSG and the G3 software (see materials) that served as the gold standard. As the beginning and end of PSG recordings and wearable recordings were manually set and therefore do not perfectly overlap in time, inter-beat interval time series provided by the wearable sensors were time synchronized to the inter-beat interval estimations from the ECG channels of the PSG recording using a windowed cross-correlation approach before sleep classification. We excluded recordings with the sensors for the following reasons: i) where signal quality was poor (25% criterion), ii) in cases the sensors failed to load the data, and iii) recordings where the synchronization between sensor and PSG was not possible.

2.4. Model optimization

Model (accuracy model) training has been described in Topalidis *et al.* [1]. We have further optimized this model by accounting for model biases towards the majority class (light sleep). We aim to oppose such biases even more by selecting the final model based on the loss function value which already incorporates the skewed class distribution (loss function model). In addition, we have implemented an IBI quality control procedure: we trained a random forest model on a subset of IBI windows which were manually labelled in terms of good and bad IBI segments and calculated a set of IBI features on a fixed time window of 10 minutes that were used as input. We started with the feature set used by Radha *et al.* [26] but reduced it to a set of 7 features based on permutation feature importance values. Furthermore, we adjusted the threshold of the output value to account for the model's ability to deal with minor levels of noise and distortion. The IBI quality control was applied after the actual sleep staging, and sleep stage labels of 30-second segments including bad IBIs were then replaced with the surrounding scorings of clean segments (i.e., segments without bad IBIs). In case more than 25% of all 30-second segments of a single night include bad IBIs, then the whole night was characterised as unstageable.

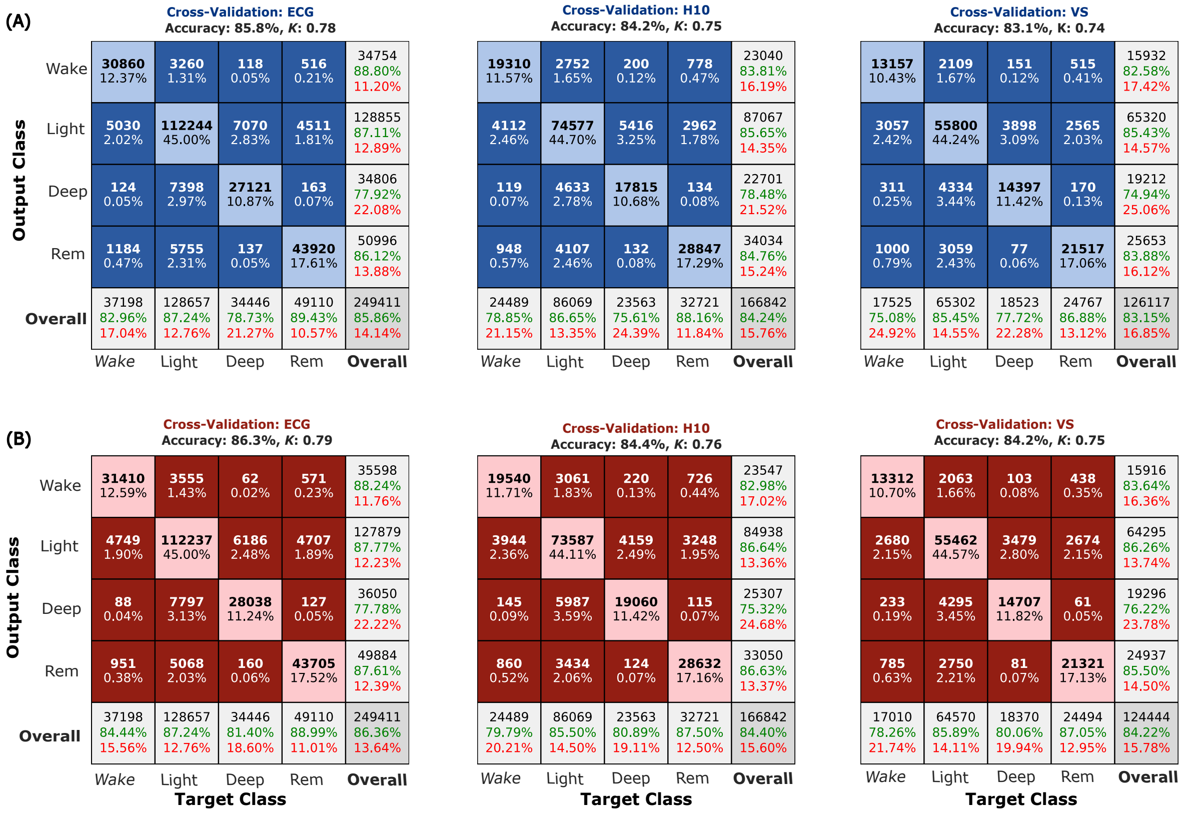
2.5. Sleep Parameters

We chose a few key sleep parameters extracted from PSG and the wearable sensors to explore their relationship and the agreement. We focused primarily on the following objective sleep variables: Sleep Onset Latency (SOL), defined as the time from lights out until the start of the first epoch of any stage of sleep (an epoch of N1, N2, N3, or R), Sleep Efficiency (SE: total sleep time/ time in bed*100), Wake After Sleep Onset (WASO), Total Sleep Time (TST), as well as Deep and REM sleep measured in minutes.

2.6. Model performance & Statistical Analysis

We used standard measures for estimating model performance, such as overall classification accuracy, Cohen's κ [27], as well per-class recall and precision values, that are displayed in the confusion matrix (see Figure 3). Note that, the performance metrics summarized in Figure 3, are computed by averaging all aggregated epochs. We further explored the performance of each sensor using a Wilcoxon signed-rank test where data for both the gold standard and each sensor existed. In addition, we examined the performance of the two models for each sleep stage separately by computing the F1 score and comparing them with a one-tailed non-parametric Wilcoxon signed-rank test expecting higher F1 scores for the loss function model. A Wilcoxon signed-rank test was also used to compare the classification accuracies between the recordings of participants on psychoactive or heart medication and those with no medication, as well as to see how these two groups differ in age and PSQI scores.

To explore the relationship between the PSG and the two sensors on sleep parameters, we conducted Spearman's Rank-Order correlation (denoted as ρ) and visualized the linear trends using scatterplots. We determined the agreement between the PSG and wearables sleep parameters using Bland-Altman plots, reporting thus biases, Standard Measurement Error (SME), Limits of Agreement (LOA), Minimal Detectable Change [28] (MDC: $SME \cdot 1.96 \cdot \sqrt{2}$), as well as absolute interclass correlation (ICC: two-way model with agreement type [29]). The MDC, or else smallest detectable change, refers to the smallest change detected by a method that exceeds measurement error [28]. The intraclass correlation of reliability encapsulates both the degree of correlation and agreement between measurements [29]. According to Koo and Li [29], values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively. All data and statistical analysis was performed in R (R Core Team [30]; version 4).


(A)
(B)
Output Class
Target Class

(A) Cross-Validation: ECG
Accuracy: 85.8%, *K*: 0.78

	Wake	Light	Deep	Rem	Overall
Wake	30860 12.37%	3260 1.31%	118 0.05%	516 0.21%	34754 88.80% 11.20%
Light	5030 2.02%	112244 45.00%	7070 2.83%	4511 1.81%	128855 87.11% 12.89%
Deep	124 0.05%	7398 2.97%	27121 10.87%	163 0.07%	34806 77.92% 22.08%
Rem	1184 0.47%	5755 2.31%	137 0.05%	43920 17.61%	50996 86.12% 13.88%
Overall	37198 82.96% 17.04%	128657 87.24% 12.76%	34446 78.73% 21.27%	49110 89.43% 10.57%	249411 85.86% 14.14%

Cross-Validation: H10
Accuracy: 84.2%, *K*: 0.75

	Wake	Light	Deep	Rem	Overall
Wake	19310 11.57%	2752 1.65%	200 0.12%	778 0.47%	23040 83.81% 16.19%
Light	4112 2.46%	74577 44.70%	5416 3.25%	2962 1.78%	87067 85.65% 14.35%
Deep	119 0.07%	4633 2.78%	17815 10.68%	134 0.08%	22701 78.48% 21.52%
Rem	948 0.57%	4107 2.46%	132 0.08%	28847 17.29%	34034 84.76% 15.24%
Overall	24489 78.85% 21.15%	86069 86.65% 13.35%	23563 75.61% 24.39%	32721 88.16% 11.84%	166842 84.24% 15.76%

Cross-Validation: VS
Accuracy: 83.1%, *K*: 0.74

	Wake	Light	Deep	Rem	Overall
Wake	13157 10.43%	2109 1.67%	151 0.12%	515 0.41%	15932 82.58% 17.42%
Light	3057 2.42%	55800 44.24%	3898 3.09%	2565 2.03%	65320 85.43% 14.57%
Deep	311 0.25%	4334 3.44%	14397 11.42%	170 0.13%	19212 74.94% 25.06%
Rem	1000 0.79%	3059 2.43%	77 0.06%	21517 17.06%	25653 83.88% 16.12%
Overall	17525 75.08% 24.92%	65302 85.45% 14.55%	18523 77.72% 22.28%	24767 86.88% 13.12%	126117 83.15% 16.85%

(B) Cross-Validation: ECG
Accuracy: 86.3%, *K*: 0.79

	Wake	Light	Deep	Rem	Overall
Wake	31410 12.59%	3555 1.43%	62 0.02%	571 0.23%	35598 88.24% 11.76%
Light	4749 1.90%	112237 45.00%	6186 2.48%	4707 1.89%	127879 87.77% 12.23%
Deep	88 0.04%	7797 3.13%	28038 11.24%	127 0.05%	36050 77.78% 22.22%
Rem	951 0.38%	5068 2.03%	160 0.06%	43705 17.52%	49884 87.61% 12.39%
Overall	37198 84.44% 15.56%	128657 87.24% 12.76%	34446 81.40% 18.60%	49110 88.99% 11.01%	249411 86.36% 13.64%

Cross-Validation: H10
Accuracy: 84.4%, *K*: 0.76

	Wake	Light	Deep	Rem	Overall
Wake	19540 11.71%	3061 1.83%	220 0.13%	726 0.44%	23547 82.98% 17.02%
Light	3944 2.36%	73587 44.11%	4159 2.49%	3248 1.95%	84938 86.64% 13.36%
Deep	145 0.09%	5987 3.59%	19060 11.42%	115 0.07%	25307 75.32% 24.68%
Rem	860 0.52%	3434 2.06%	124 0.07%	28632 17.16%	33050 86.63% 13.37%
Overall	24489 79.79% 20.21%	86069 85.50% 14.50%	23563 80.89% 19.11%	32721 87.50% 12.50%	166842 84.40% 15.60%

Cross-Validation: VS
Accuracy: 84.2%, *K*: 0.75

	Wake	Light	Deep	Rem	Overall
Wake	13312 10.70%	2063 1.66%	103 0.08%	438 0.35%	15916 83.64% 16.36%
Light	2680 2.15%	55462 44.57%	3479 2.80%	2674 2.15%	64295 86.26% 13.74%
Deep	233 0.19%	4295 3.45%	14707 11.82%	61 0.05%	19296 76.22% 23.78%
Rem	785 0.63%	2750 2.21%	81 0.07%	21321 17.13%	24937 85.50% 14.50%
Overall	17010 78.26% 21.74%	64570 85.89% 14.11%	18370 80.06% 19.94%	24494 87.05% 12.95%	124444 84.22% 15.78%

Figure 3. Confusion Matrices of accuracy (**upper**) and loss function models (**lower**). The IBIs were extracted from gold-standard ECG (**left**), chest-belt H10 (**middle**), and the PPG VS (**right**), and were classified using the two models. In each confusion matrix, rows represent predicted classes (Output Class) and columns represent true classes (Target Class). Cells on the diagonal indicate correct classifications, while off-diagonal cells represent incorrect classifications. Each cell displays the count and percentage of classifications. Precision ($truePositives/truePositives + falsePositives$) is displayed on the gray squares on the right, while recall ($truePositives/(truePositives + falseNegatives)$) is displayed at the bottom. The number of epochs has been equalized for between the two models for a more fair comparison. Note that next to the small improvement in the overall accuracy compared to the accuracy model, the loss function model displays an increase in the recall of wake and deep sleep stages. This is arguably enough to address few of the nights that are difficult to classify.

3. Results

3.1. Comparison of the accuracy and loss function models and performance after optimization

Figure 3 illustrates the confusion matrices for both models and each sensor. When looking at the overall performance, we did observed small (in the magnitude of 1 to 2%) increase in the overall classification accuracy. This was significant for the ECG (p<0.001), but not for the H10 (p=0.094) and VS (p=0.13). When looking at κ values a significant increase was observed for the ECG (p<0.001) and H10 (p=0.036), but not VS (p=0.14). Figure 4a displays the average model κ values for each sensor separately. In addition, when using the loss function model κ values for ECG (M=0.798, SD=0.08) were significantly higher (p<0.001) than the H10 (M=0.772, SD=0.09) and VS (M=0.772, SD=0.1), while the two sensors did not significantly differ. Compared to classifying only light sleep for the whole recording, considered here as the chance level, the loss function model achieved significantly higher classifications accuracies (p<0.001) for all sensors. Figure 4b illustrates the classification accuracies achieved for every recording, using the loss function model, and the corresponding accuracies when staging only the majority class. Finaly, we tested whether F1 scores, incorporating both the precision and recall of the model, improve when considering each class separately. When looking at the ECG sensor we observed small but significant classification accuracies in all sleep stages (wake: p<0.001; REM: p=0.004; light: p=0.003; deep: p=0.001). F1 scores in H10 were also higher for the loss function in all sleep stages, but light sleep (p=0.083), while only wake sleep showed a benefit from the loss function model in VS (p=0.75).

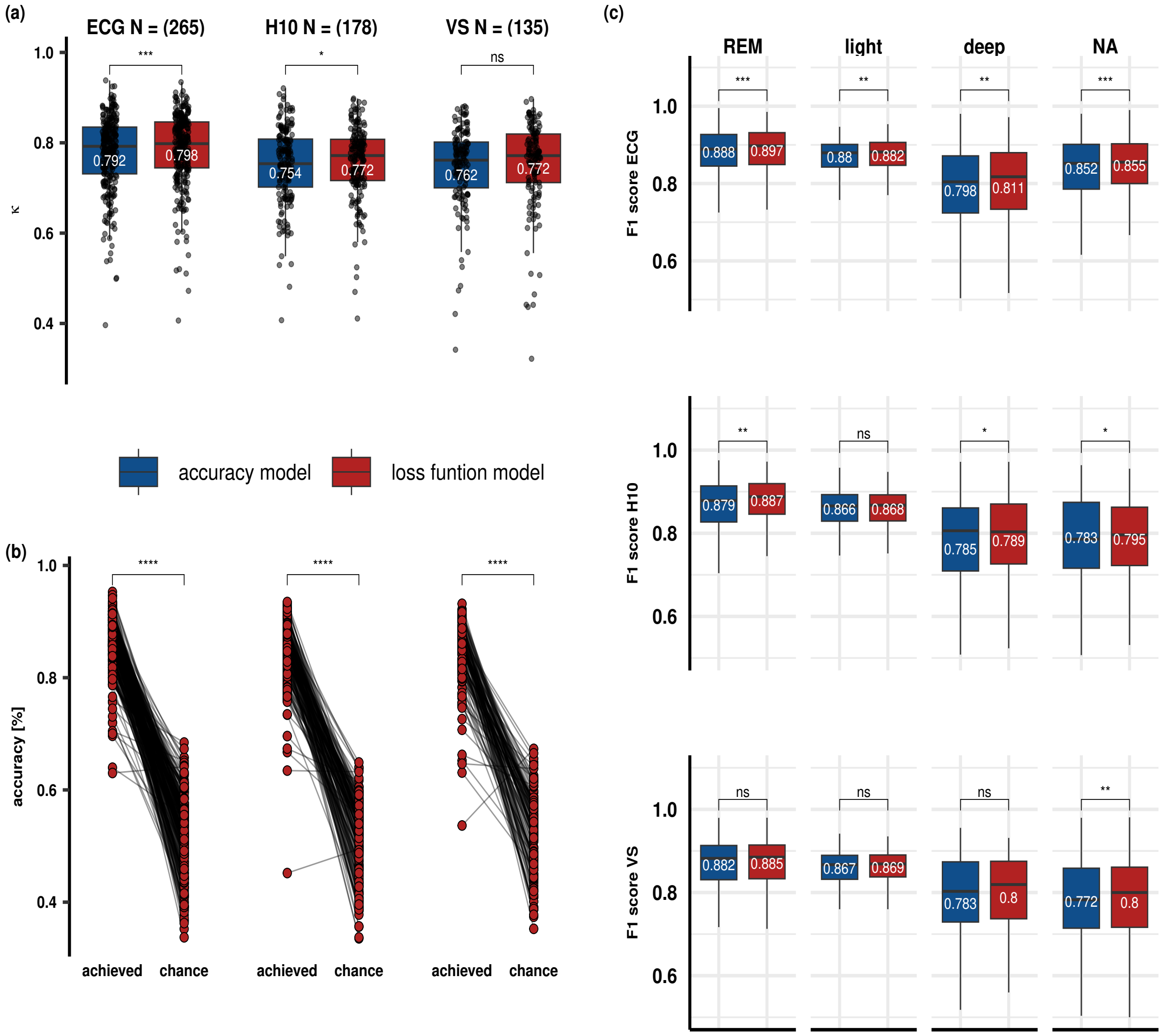
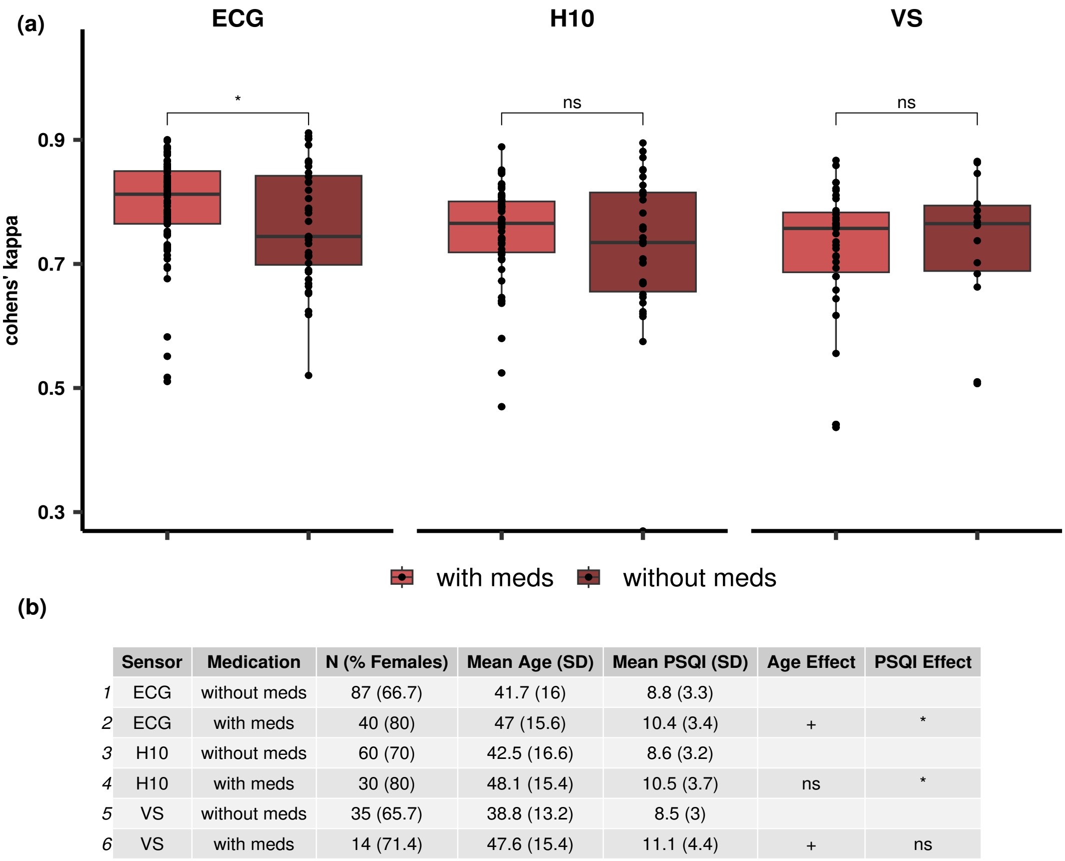


Figure 4. Performance metrics of the accuracy and loss function models as computed for ECG (**left**), H10 (**middle**), and VS (**right**). (**a**) The loss function model yield a small but significant increase in the κ values for all sensors. (**b**) The loss function model displayed significant higher classification accuracies (**b**) compared to staging the majority class for the whole recording. (**c**) When considering the performance of the two models separately in each class, as reflected in per class F1 scores, we observed small but significant increase in the performance of the loss function model. ns - not significant, <.1 - +, <.05 - *, <.01 - **, <.001 - ***, <.0001 - ****.

3.2. Effects of pshycactive and/or heart affecting medication on classification performance

We explored how medication can influence the classification performance by comparing the performance metrics, accuracy and κ, between the recordings of participant with and without medication. We found significant effects of medication for the ECG recordings (p=0.033), but such difference did not reach statistical significance for the H10 (p=0.52) and VS (p=0.87). The same effects were observed when using accuracy as a performance metric. Note that there was a trend for age differences in ECG and VS recordings (ECG: p=0.06; VS: p=0.06), while the two groups differ also significantly in their PSQI scores for the ECG (p=0.023) and H10 (p=0.027), but not the VS recordings. Figure 5a illustrates differences in κ values between the two groups for each sensor, while the desctipives are summarized in Figure 5b.



	Sensor	Medication	N (% Females)	Mean Age (SD)	Mean PSQI (SD)	Age Effect	PSQI Effect
1	ECG	without meds	87 (66.7)	41.7 (16)	8.8 (3.3)		
2	ECG	with meds	40 (80)	47 (15.6)	10.4 (3.4)	+	*
3	H10	without meds	60 (70)	42.5 (16.6)	8.6 (3.2)		
4	H10	with meds	30 (80)	48.1 (15.4)	10.5 (3.7)	ns	*
5	VS	without meds	35 (65.7)	38.8 (13.2)	8.5 (3)		
6	VS	with meds	14 (71.4)	47.6 (15.4)	11.1 (4.4)	+	ns

Figure 5. The effects of heart or psychoactive medication on sleep stage classification using the optimized model. (**a**) Comparison of κ values between recordings obtained from people with and without medication, for each sensors separately. (**b**) Group descriptives including number of subjects, age and PSQI group averages, as well as statistical effects of Age and PSQI. Note that there significant difference between the two groups in the ECG recordings, but at the same time a trend for age and a significant difference in the PSQI scores. ns - not significant, <.1 - +, <.05 - *, <.01 - **, <.001 - ***, <.0001 - ****.

3.3. Correlation and agreement between the gold-standard PSG and the two wearble devises on primary sleep parameters

We observed significant (p <.001) high Spearman correlations between all PSG and VS derived sleep parameters of interest. More specifically, between H10 and PSG we found a high correlation in Sleep Onset Latency (ρ = 0.82, p<0.001), Sleep efficiency (ρ = 0.88, p<0.001), and Total Sleep Time (ρ = 0.96, p<0.001), as well as, Wake After Sleep Onset (ρ = 0.89, p<0.001), Deep (ρ = 0.6, p<0.001) and REM sleep (ρ = 0.76, p<0.001).

Similar correlations were observed between VS and PSG: Sleep Onset Latency (ρ = 0.82, p<0.001); Sleep efficiency (ρ = 0.84, p<0.001), and Total Sleep Time (ρ = 0.94, p<0.001)'; Wake After Sleep Onset (ρ = 0.8, p<0.001); Deep (ρ = 0.6, p<0.001); REM sleep (ρ = 0.78, p<0.001). Figure 6 displays the correlation between VC and PSG.

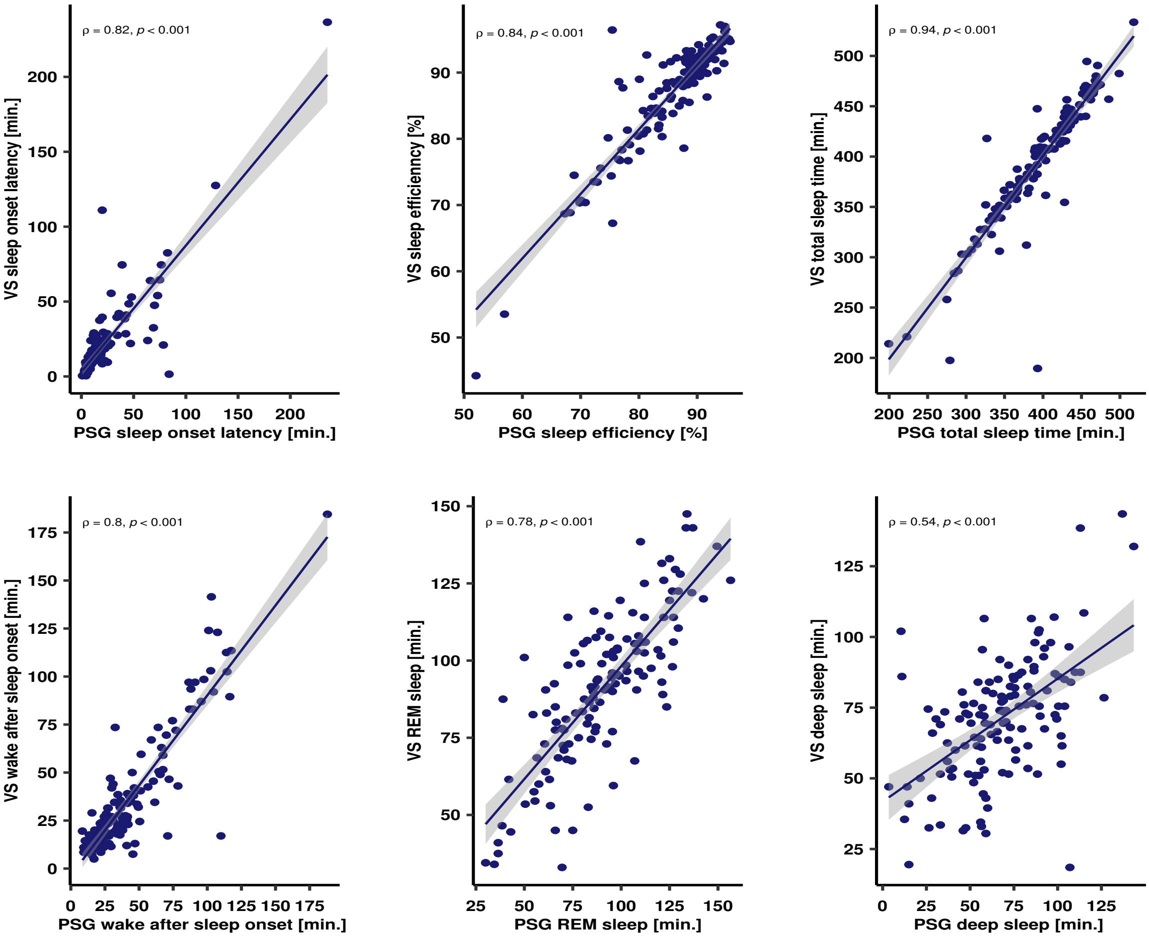
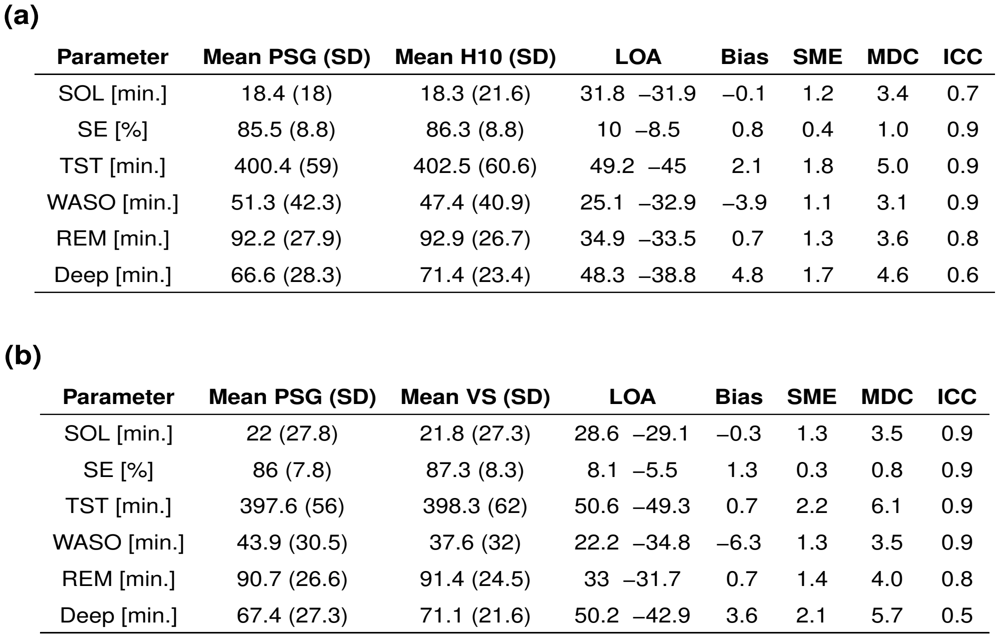


Figure 6. Correlations between VS and PSG sleep parameters as computed with Spearman's rank correlations (ρ). PSG sleep parameters are ploted on x′ axis while VS metrics on y′ axis. Individual points reflect each recording. The solid line reflects the corresponding linear model and the shaded areas the 95% confidence intervals. Note that all sleep metrics correlate highly with PSG, with deep sleep showing the weakest positive correlation.

We explored the extent of the agreement between the gold standard and both sensors using Bland-Altman plots on the sleep parameters of interest. Figure 7 summarizes the extent of agreement between the gold standard, PSG, and the two wearable devices, (**a**) H10 and (**b**) VS on the sleep parameters of interest. Figure 8 illustrates the agreement between PSG and VS through Bland-Altman plots.



(a)

Parameter	Mean PSG (SD)	Mean H10 (SD)	LOA		Bias	SME	MDC	ICC
SOL [min.]	18.4 (18)	18.3 (21.6)	31.8	−31.9	−0.1	1.2	3.4	0.7
SE [%]	85.5 (8.8)	86.3 (8.8)	10	−8.5	0.8	0.4	1.0	0.9
TST [min.]	400.4 (59)	402.5 (60.6)	49.2	−45	2.1	1.8	5.0	0.9
WASO [min.]	51.3 (42.3)	47.4 (40.9)	25.1	−32.9	−3.9	1.1	3.1	0.9
REM [min.]	92.2 (27.9)	92.9 (26.7)	34.9	−33.5	0.7	1.3	3.6	0.8
Deep [min.]	66.6 (28.3)	71.4 (23.4)	48.3	−38.8	4.8	1.7	4.6	0.6

(b)

Parameter	Mean PSG (SD)	Mean VS (SD)	LOA		Bias	SME	MDC	ICC
SOL [min.]	22 (27.8)	21.8 (27.3)	28.6	−29.1	−0.3	1.3	3.5	0.9
SE [%]	86 (7.8)	87.3 (8.3)	8.1	−5.5	1.3	0.3	0.8	0.9
TST [min.]	397.6 (56)	398.3 (62)	50.6	−49.3	0.7	2.2	6.1	0.9
WASO [min.]	43.9 (30.5)	37.6 (32)	22.2	−34.8	−6.3	1.3	3.5	0.9
REM [min.]	90.7 (26.6)	91.4 (24.5)	33	−31.7	0.7	1.4	4.0	0.8
Deep [min.]	67.4 (27.3)	71.1 (21.6)	50.2	−42.9	3.6	2.1	5.7	0.5

Figure 7. Table of Agreement. Agreement between the gold standard PSG and and wearable devices (**a**) H10 and (**b**) VS in measuring the sleep parameters of interest. Note that there is good to excellent reliability on all sleep parameters, but deep sleep which shows a moderate reliability. LOA = Limits of Agreement, upper - lower; SME = Standard Measurement Error; MDC = Minimal Detectable Change; ICC = Intra-Class Correlation.

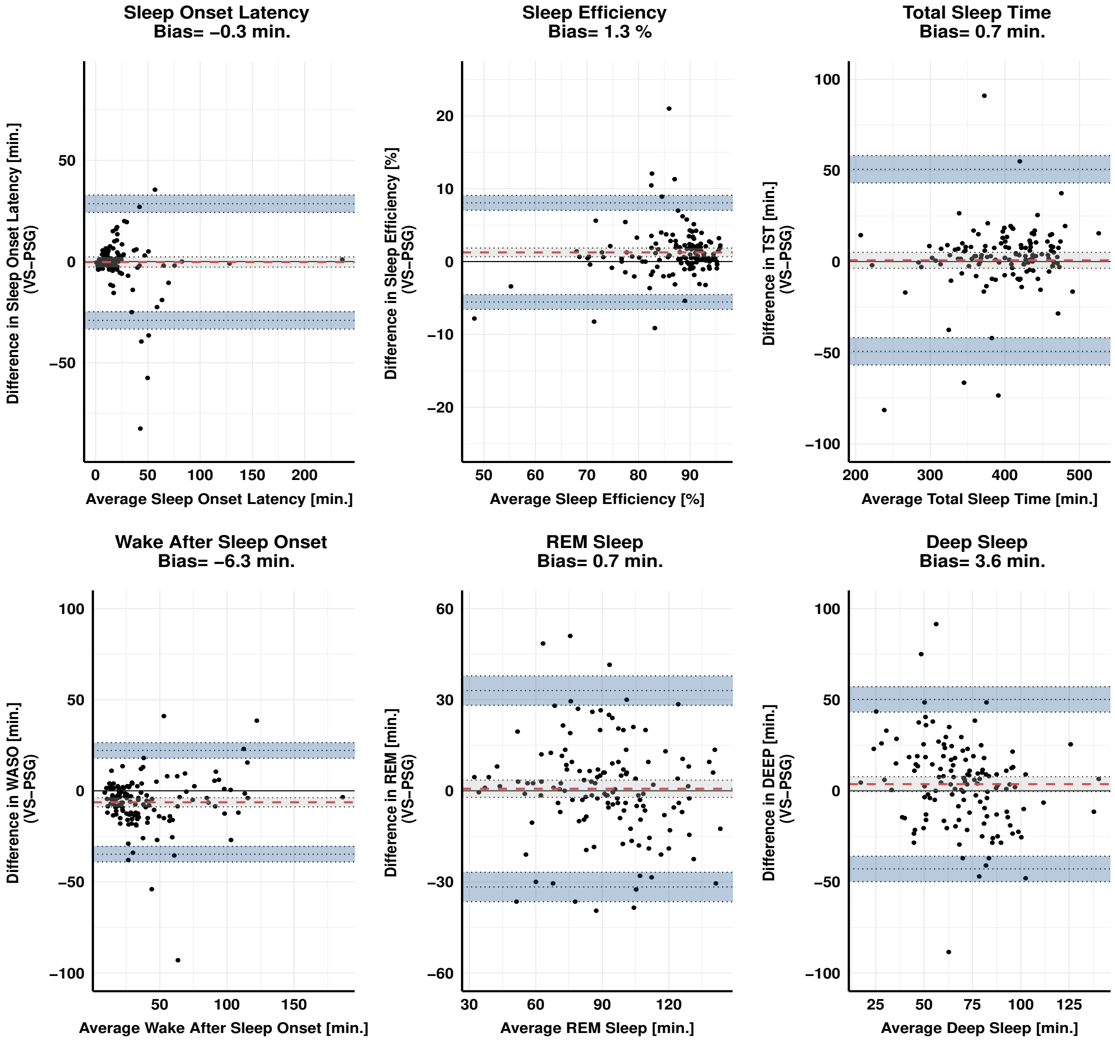


Figure 8. Agreement between the PSG-based sleep metrics and VS sensor, as visualized with Bland-Altman plots. The dashed red line represent the mean difference (i.e., bias) between the two measurements.The black solid line represents the point of equality (where the difference between the two devices is equal to 0), while the dotted lines represent upper and lower limits of agreement (LOA). The shaded areas indicate the 95% confidence interval of bias, lower and upper agreement limits. A positive bias value indicates an VS overestimation, while negative bias reflects an VS underestimation, using the gold standard, PSG, as point of reference (VS-PSG). Note that VS underestimates SOL and WASO, while it overestimates the rest of the sleep parameters. However, the degree of bias is here minimal.

4. Discussion

In the current study, we validated our previous findings in a new sample and showed that the model suggested in Topalidis *et al.* [1] can be optimized further by including advanced IBI signal control and opting for the loss function for choosing the final model. Results reveal statistically higher classification performance for the loss function model, compared to choosing the model with the highest overall accuracy during validation (accuracy model), although numerically the benefit is small. We discuss why even a small but reliable increase is relevant here while highlighting that we optimize a model that already displays very high classification performance. We further show that psychoactive and/or heart-affecting medication does not have a strong impact on sleep stage classification accuracy. Lastly, we evaluated our new optimized model for measuring sleep-related parameters and found that our model shows substantial agreement and reliability with PSG-extracted sleep parameters.

In the following section, we discuss the results, acknowledge the limitations of the current study and provide an outlook on sleep classification using wearables in sleep research and clinical practice for the near future.

When looking at the confusion matrices (in Figure 3) we observed a small increase in the overall accuracy for the loss function model, although the accuracy model uses the overall accuracy for choosing the best model in the training phase. Descriptively, there is an increase in wake and deep sleep recall in the loss function model (in the magnitude of 1% to 5 %), suggesting that it is indeed beneficial to opt for the loss function model for the model training. This becomes also apparent when comparing the κ values from the individual nights Figure 4, without epoch aggregation. In addition, we observed that the optimized model performs far above a model that would only predict the majority class (light sleep) for the entire recording. When comparing the two models statistically per class using F1 scores extracted per recording, we observed a benefit for the loss function model in all sleep stages when looking at ECG data. In H10 recordings we observed benefits for wake, REM, and deep sleep, while for the VS only classifying wake with the optimised model showed an improvement.

One could discuss whether these small accuracy increases are meaningful but one needs to acknowledge that the new model displays accuracy performance approaching human inter-rater agreement and critically improves in wake classification which is crucial as causing most irritation in the end-user if experienced as wrong. As it has been pointed out in [1] it is estimated that at four classes experts display an agreement of around 88% as compared to the 84% ($\kappa \approx .75$) on average in our study across the wearable devices which equals to 95.45% agreement of our algorithm with PSG-based human scorings. In the presented classification approach we implement the recommendation from [31], who systematically reviewed automatic sleep staging using cardiorespiratory signals, and suggests taking into account signal delays and implementing sequence learning models that are capable of incorporating temporal relations.

In the current study, we also explored the effects of psychoactive and heart medication on sleep stage classification performance. We reasoned that such medication could have a direct effect on PSG and ECG [32,33] and thereby affect relevant features of the signal, resulting in decreased classification performance. We observed a small but statistically significant decrease in κ values in people with and without medications for the ECG data, but no such drops in classification accuracy for the H10 or VS sensor. It is important to note that for the ECG data also age and PSQI differences were observed for the medication vs. non-medication group (on medication being older and having worse sleep quality) and thus the drop in classification performance could be driven by these differences. However, also here the median κ for all classified recordings from participants on medication groups was above .07 for all devices, suggesting a substantial classification agreement.

Furthermore, we evaluated the agreement and reliability of our model against PSG on primary sleep parameters. Particular emphasis was put on the VS sensor, as it is the more comfortable sensor to sleep with, while at the same time providing a nice indirect measure of heart-rate variability using the pulse-to-pulse intervals of the PPG. Surprisingly, only for a few sleep variables, the ECG-based H10 was found to be better in accuracy. For all stages but deep sleep (0.5), the key sleep parameters showed high correlations (0.7 to 0.94) between PSG and VS. The highest correlation was found for total sleep time and showed almost perfect agreement (r = .94). These correlational analyses were nicely complemented by intra-class correlations, which likewise indicated moderate (for deep sleep) to excellent reliability. Systematic bias of the VS against and the PSG gold standard was visualised using Bland-Altman plots and was found to be minimal. Keeping in mind that if the end-user has a sleep issue, "deep sleep" seems to be a crucial measure, as it has been related to physical and mental recovery, immune system strengthening or the disposal of brain waste products via the glymphatic system (e.g., Reddy and van der Werf [34]). Future emphasis shall thus be put on further increasing the classification accuracy for the "deep sleep" class. We also provided the minimal detectable change (MDC) metric, which indicates the smallest change detected by a method that exceeds the measurement error [28], and found that with both the H10 and VS there is a very good resolution that is accurate up to ±5 minutes across sleep parameters.

In summary, a trend in society and Western health systems goes towards an increasing adoption of digital health interventions including sleep (see Arroyo and Zawadzki [35] for a systematic review on mHealth interventions in sleep). Objective and reliable tracking of the effects of such interventions thus also becomes more and more relevant and allows ecologically valid and continuous measurements [10] in natural home settings. Recently, for example, Spina *et al.* [12] used sensor-based sleep feedback in a sample of 103 participants suffering from sleep disturbances and found that such sensor-based sleep feedback can already reduce some of the insomnia symptoms. Interestingly, such feedback alone was however not enough to induce changes in sleep–wake misperception which may need additional interventions (see Hinterberger *et al.* [36]). Given that people with sleep problems and preoccupations about their sleep are especially sensitive to such feedback there is a high ethical necessity to only provide certain and accurate feedback to patients to prevent negative side effects [5].

Author Contributions: PT: Conceptualization, Writing - original draft, Data Analysis, Visualization. DH: Machine Learning. BS: Machine Learning. EE and AH: Conceptualization, Data-Collection. MS: Supervision, Conceptualization, Writing - review and editing.

Funding: Pavlos I. Topalidis was supported by the Doctoral College "Imaging the Mind" (FWF, Austrian Science Fund:W 1233-B). We would also like to thank the funding grant (20102-F2002176-FPR) of the county of Salzburg.

Institutional Review Board Statement: All procedures performed were in accordance with the ethical standards of the institutional research committee (GZ 46/2020) and with the Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available from the University of Salzburg. Restrictions apply to the availability of these data, which were used under licence for this study.

Conflicts of Interest: MS is co-founder of and CSO at the NUKKUAA GmbH. DH is also an employee of the Institut Proschlaf Austria (IPA).

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
IBI	Inter-beat-interval
HRV	Heart Rate Variability

References

1. Topalidis, P.; Heib, D.P.; Baron, S.; Eigl, E.S.; Hinterberger, A.; Schabus, M. The Virtual Sleep Lab—A Novel Method for Accurate Four-Class Sleep Staging Using Heart-Rate Variability from Low-Cost Wearables. *Sensors* **2023**, *23*, 2390.
2. Ramar, K.; Malhotra, R.K.; Carden, K.A.; Martin, J.L.; Abbasi-Feinberg, F.; Aurora, R.N.; Kapur, V.K.; Olson, E.J.; Rosen, C.L.; Rowley, J.A.; others. Sleep is essential to health: an American Academy of Sleep Medicine position statement. *Journal of Clinical Sleep Medicine* **2021**, *17*, 2115–2119.
3. Baron, K.G.; Abbott, S.; Jao, N.; Manalo, N.; Mullen, R. Orthosomnia: are some patients taking the quantified self too far? *Journal of Clinical Sleep Medicine* **2017**, *13*, 351–354.
4. Rentz, L.E.; Ulman, H.K.; Galster, S.M. Deconstructing commercial wearable technology: contributions toward accurate and free-living monitoring of sleep. *Sensors* **2021**, *21*, 5071.
5. Gavriloff, D.; Sheaves, B.; Juss, A.; Espie, C.A.; Miller, C.B.; Kyle, S.D. Sham sleep feedback delivered via actigraphy biases daytime symptom reports in people with insomnia: Implications for insomnia disorder and wearable devices. *Journal of sleep research* **2018**, *27*, e12726.
6. Ravichandran, R.; Sien, S.W.; Patel, S.N.; Kientz, J.A.; Pina, L.R. Making sense of sleep sensors: How sleep sensing technologies support and undermine sleep health. Proceedings of the 2017 CHI conference on human factors in computing systems, 2017, pp. 6864–6875.

7. Downey, R.; Bonnet, M.H. Training subjective insomniacs to accurately perceive sleep onset. *Sleep* **1992**, *15*, 58–63.
8. Tang, N.K.; Harvey, A.G. Correcting distorted perception of sleep in insomnia: a novel behavioural experiment? *Behaviour research and therapy* **2004**, *42*, 27–39.
9. Tang, N.K.; Harvey, A.G. Altering misperception of sleep in insomnia: behavioral experiment versus verbal feedback. *Journal of consulting and clinical psychology* **2006**, *74*, 767.
10. Roomkham, S.; Lovell, D.; Cheung, J.; Perrin, D. Promises and challenges in the use of consumer-grade devices for sleep monitoring. *IEEE reviews in biomedical engineering* **2018**, *11*, 53–67.
11. Aji, M.; Glozier, N.; Bartlett, D.J.; Grunstein, R.R.; Calvo, R.A.; Marshall, N.S.; White, D.P.; Gordon, C. The effectiveness of digital insomnia treatment with adjunctive wearable technology: a pilot randomized controlled trial. *Behavioral sleep medicine* **2022**, *20*, 570–583.
12. Spina, M.A.; Andrillon, T.; Quin, N.; Wiley, J.F.; Rajaratnam, S.M.; Bei, B. Does providing feedback and guidance on sleep perceptions using sleep wearables improves insomnia? Findings from Novel Insomnia Treatment Experiment ("NITE"), a randomised controlled trial. *Sleep* **2023**, p. zsad167.
13. Song, Y.M.; Choi, S.J.; Park, S.H.; Lee, S.J.; Joo, E.Y.; Kim, J.K. A real-time, personalized sleep intervention using mathematical modeling and wearable devices. *Sleep* **2023**, *46*, zsad179.
14. Murray, J.M.; Magee, M.; Giliberto, E.S.; Booker, L.A.; Tucker, A.J.; Galaska, B.; Sibenaller, S.M.; Baer, S.A.; Postnova, S.; Sondag, T.A.; others. Mobile app for personalized sleep–wake management for shift workers: A user testing trial. *Digital Health* **2023**, *9*, 20552076231165972.
15. Chinoy, E.D.; Cuellar, J.A.; Huwa, K.E.; Jameson, J.T.; Watson, C.H.; Bessman, S.C.; Hirsch, D.A.; Cooper, A.D.; Drummond, S.P.; Markwald, R.R. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep* **2021**, *44*, zsaa291.
16. Chee, N.I.; Ghorbani, S.; Golkashani, H.A.; Leong, R.L.; Ong, J.L.; Chee, M.W. Multi-night validation of a sleep tracking ring in adolescents compared with a research actigraph and polysomnography. *Nature and science of sleep* **2021**, pp. 177–190.
17. Altini, M.; Kinnunen, H. The promise of sleep: A multi-sensor approach for accurate sleep stage detection using the oura ring. *Sensors* **2021**, *21*, 4302.
18. Mathunjwa, B.M.; Lin, Y.T.; Lin, C.H.; Abbod, M.F.; Sadrawi, M.; Shieh, J.S. Automatic IHR-based sleep stage detection using features of residual neural network. *Biomedical Signal Processing and Control* **2023**, *85*, 105070.
19. Habib, A.; Motin, M.A.; Penzel, T.; Palaniswami, M.; Yearwood, J.; Karmakar, C. Performance of a Convolutional Neural Network Derived from PPG Signal in Classifying Sleep Stages. *IEEE Transactions on Biomedical Engineering* **2022**.
20. Sridhar, N.; Shoeb, A.; Stephens, P.; Kharbouch, A.; Shimol, D.B.; Burkart, J.; Ghoreyshi, A.; Myers, L. Deep learning for automated sleep staging using instantaneous heart rate. *NPJ digital medicine* **2020**, *3*, 1–10.
21. Iber, C. The AASM manual for the scoring of sleep and associated events: rules, terminology, and technical specification. *(No Title)* **2007**.
22. Bakker, J.P.; Ross, M.; Cerny, A.; Vasko, R.; Shaw, E.; Kuna, S.; Magalang, U.J.; Punjabi, N.M.; Anderer, P. Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring. *Sleep* **2023**, *46*, zsac154.
23. Schaffarczyk, M.; Rogers, B.; Reer, R.; Gronwald, T. Validity of the polar H10 sensor for heart rate variability analysis during resting state and incremental exercise in recreational men and women. *Sensors* **2022**, *22*, 6536.
24. Gilgen-Ammann, R.; Schweizer, T.; Wyss, T. RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. *European journal of applied physiology* **2019**, *119*, 1525–1532.
25. Hettiarachchi, I.T.; Hanoun, S.; Nahavandi, D.; Nahavandi, S. Validation of Polar OH1 optical heart rate sensor for moderate and high intensity physical activities. *PLoS One* **2019**, *14*, e0217288.
26. Radha, M.; Fonseca, P.; Moreau, A.; Ross, M.; Cerny, A.; Anderer, P.; Long, X.; Aarts, R.M. Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Scientific reports* **2019**, *9*, 1–11.
27. Viera, A.J.; Garrett, J.M.; others. Understanding interobserver agreement: the kappa statistic. *Fam med* **2005**, *37*, 360–363.
28. de Vet, H.C.; Terwee, C.B.; Ostelo, R.W.; Beckerman, H.; Knol, D.L.; Bouter, L.M. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and quality of life outcomes* **2006**, *4*, 1–5.

doi:10.20944/preprints202310.0592.v1

29. Koo, T.; Li, M. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016; 15 (2): 155–63, 2000.
30. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.
31. Ebrahimi, F.; Alizadeh, I. Automatic sleep staging by cardiorespiratory signals: a systematic review. *Sleep and Breathing* **2021**, pp. 1–17.
32. Doghramji, K.; Jangro, W.C. Adverse effects of psychotropic medications on sleep. *Psychiatric Clinics* **2016**, *39*, 487–502.
33. Symanski, J.D.; Gettes, L.S. Drug effects on the electrocardiogram: a review of their clinical importance. *Drugs* **1993**, *46*, 219–248.
34. Reddy, O.C.; van der Werf, Y.D. The sleeping brain: harnessing the power of the glymphatic system through lifestyle choices. *Brain sciences* **2020**, *10*, 868.
35. Arroyo, A.C.; Zawadzki, M.J. The implementation of behavior change techniques in mHealth apps for sleep: systematic review. *JMIR mHealth and uHealth* **2022**, *10*, e33527.
36. Hinterberger, A.; Eigl, E.S.; Schwemlein, R.N.; Topalidis, P.; Schabus, M. Investigating the subjective and objective efficacy of a cognitive behavioral therapy for insomnia (CBT-I)-based smartphone app on sleep: a randomized controlled trial **2023**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.