Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Automated remote sleep monitoring needs uncertainty quantification

Elisabeth R.M. Heremans[1] | Laura Van den Bulcke[2] | Nabeel Seedat[3] | Astrid Devulder[4] | Pascal Borzée[5] | Bertien Buyse[5] | Dries Testelmans[5] | Maarten Van Den Bossche[2] | Mihaela van der Schaar[3] | Maarten De Vos[1,6]

[1]Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium

[2]Neuropsychiatry, Department of Neurosciences, KU Leuven, Leuven, Belgium

[3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom

[4]Department of Neurology, University Hospitals Leuven, Leuven, Belgium

[5]Department of Pneumology, University Hospitals Leuven, Leuven, Belgium

[6]Department of Regeneration, KU Leuven, Leuven, Belgium

Correspondence
Corresponding author Elisabeth R.M. Heremans, KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.
Email: elisabeth.heremans@kuleuven.be

Funding Information
This research was supported by the Research Foundation Flanders, Bijzonder Onderzoeksfonds KU Leuven, Flemish Government

Abstract

Wearable electroencephalography (EEG) devices emerge as a cost-effective and ergonomic alternative to gold standard polysomnography, paving the way for better health monitoring and sleep disorder screening. Machine learning allows to automate sleep stage classification, but trust and reliability issues have hampered its adoption in clinical applications. Estimating uncertainty is a crucial factor in enhancing reliability by identifying regions of heightened and diminished confidence. In this study, we used an uncertainty-centered machine learning pipeline, U-PASS, to automate sleep staging in a challenging real-world dataset of single-channel EEG and accelerometry collected with a wearable device from an elderly population. We were able to effectively limit the uncertainty of our machine learning model and to reliably inform clinical experts of which predictions were uncertain to improve the machine learning model's reliability. This increased the five-stage sleep scoring accuracy of a state-of-the-art machine learning model from 63.9% to 71.2% on our dataset. Remarkably, the machine learning approach outperformed the human expert in interpreting these wearable data. Manual review by sleep specialists, without specific training for sleep staging on wearable EEG, proved ineffective. The clinical utility of this automated remote monitoring system was also demonstrated, establishing a strong correlation between the predicted sleep parameters and the reference PSG parameters, and reproducing known correlations with the apnea-hypopnea index. In essence, this work presents a promising avenue to revolutionize remote patient care through the power of machine learning by the use of an automated data processing pipeline enhanced with uncertainty estimation.

KEYWORDS
electroencephalography, wearable EEG, machine learning, uncertainty quantification, sleep staging

1 | INTRODUCTION

An estimated 40 - 70 % of the elderly population suffers from chronic sleep problems, with insomnia and obstructive sleep apnea (OSA) being amongst the most frequently reported (Miner & Kryger, 2017; Irwin & Vitiello, 2019). Sleep disturbances not only impact the quality of life, but are also related to a variety of health conditions such as stroke (Siengsukon, Al-Dughmi, Al-Sharman, & Stevens, 2015), diabetes (Pallayova, Donic, Gresova, Peregrim, & Tomori, 2010), and depression (Tsuno, Besset, & Ritchie, 2005). Sleep is also related to cognitive function (Leng, McEvoy, Allen, & Yaffe, 2017; Pase et al., 2023; Bubu et al., 2020; Winer et al., 2021) and dementia (D'Rozario et al., 2020; Bubu et al., 2020; Lucey et al., 2021). Therefore, the objective measurement of sleep serves a dual purpose: diagnosing sleep disorders and early detection of associated health risks.

 © 2023 Copyright Holder Name

Nowadays, the gold standard for the objective measurement of sleep is in-hospital polysomnography (PSG) with manual sleep staging. However, the limited hospital bed capacity and the labor-intensive and costly nature of PSG limit its widespread application in large populations. Therefore, the objective measurement of sleep with wearable devices in a home-based setting has become an increasingly important topic of study. Many types of wearable devices have been developed for sleep monitoring, capturing physiological signals including movement, heart rate, and respiration (Radha et al., n.d.; Sridhar et al., 2020; Martin & Hakim, 2011; Qi, Soh, Gunawan, Low, & Thomas, 2014). Although these are valuable in consumer settings, research shows that wearable electroencephalography (EEG) devices are the most reliable option to perform complete five-stage sleep staging (Perez-Pozuelo et al., n.d.). The advent of wearable EEG measurement devices presents a vast opportunity for cost-effective and long-term at-home sleep monitoring (Arnal et al., 2020; K. B. Mikkelsen et al., 2019; Finan et al., 2016). However, their capacity to collect large amounts of data necessitates the utilization of accurate and reliable automated sleep scoring methods. This pressing urgency is aggravated further by the challenge that trained clinicians face in interpreting wearable EEG signals recorded from non-standard anatomical locations (K. B. Mikkelsen et al., 2019). Without the help of automated annotation, large-scale at-home sleep monitoring remains a far-fetched fantasy.

Fortunately, the fast-growing field of machine learning offers a solution, with a plethora of deep learning methods for sleep staging that have emerged in recent years (Phan, Andreotti, Cooray, Chén, & De Vos, 2018; Phan et al., 2020; Guillot & Thorey, 2021; Perslev et al., 2021). These models have demonstrated a close-to-human expert accuracy on public datasets, but their adoption in daily clinical practice remains very limited. The reasons for this are multiple. First, technology used in the medical field needs to meet very high performance and reliability standards. Machine learning models for sleep staging have mostly been trained on datasets consisting of healthy young individuals. Thus, their efficacy diminishes when applied to datasets comprising individuals with fragmented sleep patterns, such as the elderly or diseased – the population of greatest interest (Phan & Mikkelsen, 2022; Mander, Winer, & Walker, 2017; D'Rozario et al., 2020; Zhang et al., 2022; Jordan, McSharry, & Malhotra, 2014). Moreover, sleep staging models trained on data acquired with a certain measurement setup or device typically do not generalize well to data acquired with a different measurement protocol, because of the well-known distribution shift problem (Phan, Chén, et al., 2021; Heremans et al., 2022). This problem is aggravated for wearable devices, which record data from diverse non-standard locations. Second, the lack of transparency in machine learning models makes clinicians reluctant to adopt these methods in clinical practice (Phan & Mikkelsen, 2022; Qayyum, Qadir, Bilal, & Al-Fuqaha, 2021; Miotto, Wang, Wang, Jiang, & Dudley, 2017). In any decision-making or prediction process based on real-world data, there is inevitably some level of uncertainty. Quantifying this uncertainty helps to gain insight into the decision-making process and to inspect uncertain decisions, thereby addressing trust issues. In addition, identifying the uncertainty is the first step towards removing or reducing it, which in turn improves robustness and reliability. In conclusion, providing reliable uncertainty estimates is essential for the deployment of machine learning in daily clinical practice.

Two recent studies (Kang, DeYoung, Tantiongloc, Coleman, & Owens, 2021; Phan, Mikkelsen, et al., 2021) proposed uncertainty quantification for automated sleep staging based on PSG recordings. Both quantified the uncertainty of machine learning predictions using a simple uncertainty metric. Kang et al. (Kang et al., 2021) subsequently presented the identified uncertain PSG epochs to a clinician for manual review. In contrast, this study focuses on wearables and aims to enable remote sleep monitoring using an uncertainty-guided machine learning pipeline (U-PASS) (Heremans et al., 2024), as shown in Figure 1. Specifically, the study has the following objectives:

1. Improve sleep staging accuracy in a challenging real-world dataset, consisting of single-channel wearable EEG and accelerometry collected from an elderly population with suspected sleep apnea.
2. Enhance the machine learning pipeline by excluding low-quality training data during model training and flagging uncertain data during deployment. These steps, respectively, are aimed at improving the model and reducing mistakes.
3. Investigate the role of human annotators in remote sleep monitoring with wearables, as compared to in-hospital standardized PSG studies.
4. Verify that a machine learning model based on behind-the-ear wearable EEG picks up on relevant diagnostic features with enough accuracy to be clinically useful.

By addressing these aims, this study seeks to demonstrate the potential of integrating wearable technology into sleep monitoring practices, providing insights that could facilitate more reliable and accessible solutions for sleep staging in clinical settings.

2 | METHODS

2.1 | Dataset

This study used sleep data from the Leuven senior sleep dataset (Heremans et al., 2022). This dataset consists of recordings from 90 patients over the age of 60 and with suspicion for OSA based on symptoms mentioned by the patient or caregiver. These patients were referred to the sleep lab of the University Hospitals Leuven to undergo an overnight PSG from September 2019 until December 2022. Only referred patients without major cardio-vascular and neurological disorders were included in the study. 79% of the patients had an apnea-hypopnea index (AHI) $\geq$ 15. Details of the dataset are described in Table 1. All participants underwent a PSG at the sleep laboratory of the University Hospitals Leuven (UZ Leuven). In addition, for each patient, sleep was simultaneously recorded with a wearable behind-the-ear EEG device. The cross-head behind-the-ear EEG measurement was simulated with standard EEG

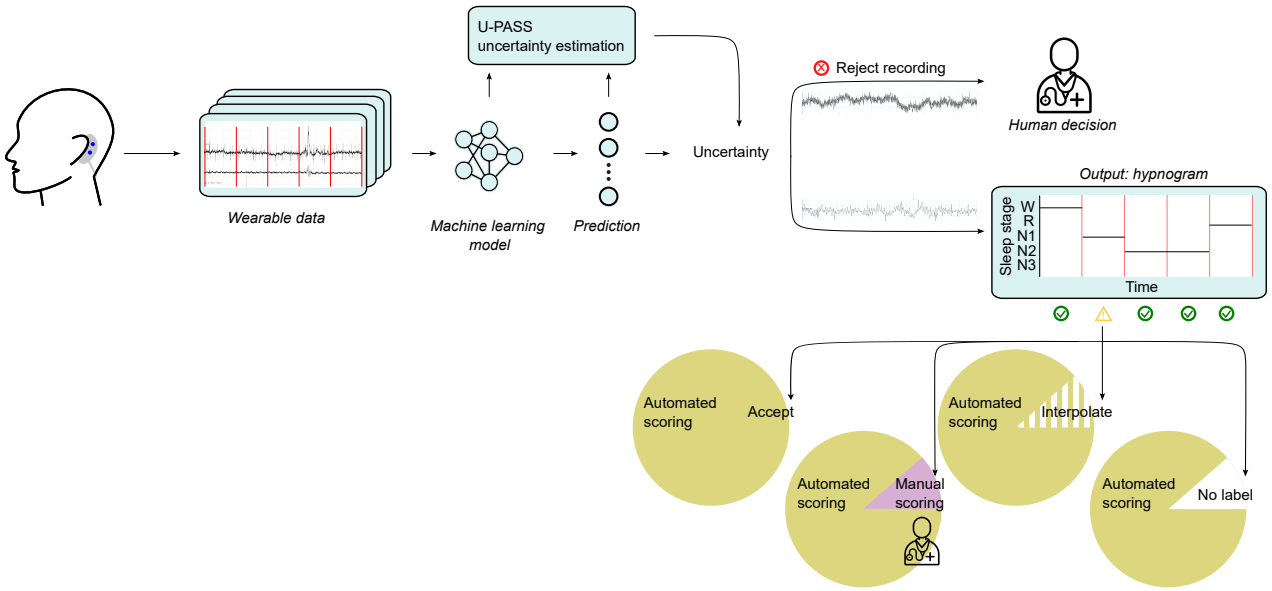


FIGURE 1 The proposed machine learning pipeline for automated sleep staging on wearable sleep recordings, enhanced with uncertainty quantification. Brain signals and potentially other physiological signals are measured from patients using a wearable device. After acquiring a dataset, a model is trained for sleep staging on these data. At training time, data uncertainty is tracked to select which channels to use, and to remove uninformative data. At evaluation time, the recordings with the most uncertainty are marked and presented to a clinician, who then decides whether to discard the recording or take further action. All other recordings are labeled by the machine learning model, and only the most uncertain samples of these recordings are marked. This work investigates what action is required for these uncertain samples: the four yellow circles show the four options we discuss.

equipment in the first 45 recordings, and measured with a real wearable device (SensorDot (Swinnen et al., 2021)) in the last 45 recordings. In both cases, the cross-head EEG measurements were obtained with one additional Ag/AgCl cup electrode placed approximately one centimeter under the A2 position on the right mastoid bone, referenced to A1. The wearable device measurements included a cross-head EEG channel, angular velocity, and acceleration. The sampling rates were 250 Hz for the wearable EEG recordings, and 50 Hz for the angular velocity and accelerometry. Every recording was trimmed to only retain the segment of the night where the lights were off, and pre-processed by resampling to 100 Hz and computing their short-time Fourier transform (STFT). The first 45 recordings were used as a training set, and the final 45 recordings were used as the test set. For each patient, the PSG was manually annotated according to the American Academy of Sleep Medicine (AASM) criteria (Berry, Quan, & Abreu, 2020), with each 30-second epoch labeled as Wake, REM, N1, N2, or N3. The sleep study was conducted in accordance with the Declaration of Helsinki, and the protocol with registration number S64190/B3222020000148 was approved by the Ethics Committee Ethische Commissie Onderzoek UZ/KU Leuven. The clinical trial number was NCT04755504.

TABLE 1 Sample characteristics.

	Train dataset	Test dataset
n	45	45
Age	65.96 ± 4.95	66.13 ± 5.02
BMI*	33.44 ± 14.84	30.60 ± 6.50
% female	35.6	42.2
AHI	33.38 ± 21.20	35.17 ± 25.47

Abbreviations: BMI, body mass index. AHI, apnea-hypopnea index.
*BMI was not recorded for two patients of the test dataset.
Data are reported as mean ± standard deviation unless not applicable.

2.2 | Sleep staging model

We trained SeqSleepNet (Phan et al., 2018), a state-of-the-art sleep staging model, to perform automated sleep staging on the wearable recordings in this dataset. This deep neural network has a sequence-to-sequence classification scheme. It transforms a sequence of adjacent raw data segments into a corresponding sequence of outputs, in this case sleep stages. The inputs are presented as time-frequency images, spectrograms of each 30-second segment. The sequence length is a parameter, which was fixed to 10 in this study. The SeqSleepNet architecture is composed of a block of layers that processes individual segments, and a block that works at the sequence level. The segment processing block is comprised of a number of frequency filters, followed

by an recurrent neural network (RNN) on the segment level. The outputs from the segment processing block are then concatenated into a sequence of inputs for the sequence processing block. A sequence-level RNN transforms this input sequence into an output sequence, which is mapped to a sequence of sleep stages by a softmax layer. The optimization and training parameters in this study were identical to those in the original paper (Phan et al., 2018): L2 regularization was applied, and the Adam optimizer was used with a learning rate of 10^{-4}. The training procedure is detailed in Appendix A.1.

2.3 | U-PASS and uncertainty measures

The sleep staging model was trained on the wearable data using the U-PASS uncertainty-guided pipeline (Heremans et al., 2024) (see Figure 1). We previously developed this pipeline to improve the training data and hence the model performance, and identify uncertainty in the test data at deployment time to increase reliability. At training time, U-PASS first determines the minimal measurements needed based on the amount of uncertainty. Second, it removes the most ambiguous training data to further improve the model performance. At test time, it flags uncertain data, increasing the model's reliability on the remaining predictions. In this work, this flagging at deployment time was performed at the level of whole recordings and at the sample level. First, particularly difficult recordings were marked and presented to the clinician for further decision-making. This step was a coarse preselection to discard recordings that were unreadable. Second, for the recordings that were considered reliable, the same uncertainty metric was deployed to flag individual segments for which the model prediction was uncertain.

To identify uncertainty in the training data and test data, U-PASS estimates uncertainty using predefined uncertainty measures. We used the measures defined in (Heremans et al., 2024), which represented dynamics computed over training, rather than just capturing a single snapshot at the end of training. To identify uncertainty in the training data, we used the data uncertainty defined as $v_{al} = \mathbb{E}_{\Theta}[\mathcal{P}(x,\Theta)(1 - \mathcal{P}(x,\Theta))]$ (Seedat, Crabbé, Bica, & van der Schaar, 2022; Heremans et al., 2024), where $\mathcal{P}(x,\theta)$ is the model's output probability for the true class, and Θ are the model's parameters over the model's training iterations. To identify uncertainty in the test data, we used the distance-weighted confidence of the neighboring training samples (Heremans et al., 2024). For the n nearest training samples, the average confidence of the model is $\mathbb{E}_{\Theta}[\mathcal{P}(x,\Theta)]$. For each of these training samples, the associated confidence was weighted with its distance to the given test sample. These two factors respectively identified out-of-distribution samples and samples situated near uncertain regions, such as decision boundaries. Out-of-distribution samples are samples that deviate from the training data, such as samples corrupted by a new type of noise. Samples near decision boundaries typically exhibit features indicative of different sleep stages, for example, because of a sleep stage transition within an epoch.

In general, the uncertainty in sleep staging emerges from biological factors such as disease and age, measurement technique-related factors like segmentation and interference (van Gorp et al., 2022), and different possible interpretations of sleep scoring rules. These factors are inherent problems to sleep staging, leading to significant inter-rater variability as well as uncertainty in machine learning models (Lee, Lee, Cho, & Choi, 2022; Heremans et al., 2024) Appendix B visualizes some of these sources of uncertainty.

2.4 | Manual review

The U-PASS pipeline allows to flag uncertain predictions at deployment time, but we have not yet determined what to do with those flagged instances. Figure 1 visualizes how uncertain instances are detected, and which different actions could be taken for these uncertain predictions. When applying machine learning on PSG in hospital-based sleep studies, uncertain decisions can be deferred to human certified sleep scorers who are trained to review PSG data (Heremans et al., 2024; Kang et al., 2021). However, in the case of monitoring with wearables, the non-standard acquisition setups make the signals hard to read for clinicians (K. Mikkelsen & De Vos, 2018). In this study, we performed a small-scale experiment of preliminary nature to investigate the ability of a human scorer to review wearable EEG and accelerometry acquired. We first tested the human scorer's ability to independently score the full wearable recordings without access to any machine learning labels. Then, we tested the ability of the scorer to improve the machine learning predictions for the 10% most uncertain areas, given the full wearable recording and the machine learning predictions for the rest of the recording. We compared the human scorer's results to two automated baselines: one using all the machine learning predictions, even the most uncertain ones, and the other interpolating uncertain predictions with surrounding sleep stages. The experimental setup is detailed in Appendix A.2.

2.5 | Reliability assessment

The sleep staging accuracy should not be the sole focus for an automated sleep monitoring tool. Indeed, sleep staging serves as a means to an end in a sleep study. The aggregated sleep stages form sleep parameters, which, along with clinical data, lead to the diagnosis of sleep disorders (Shrivastava, Jung, Saadat, Sirohi, & Crewson, 2014). We calculated a total of 26 sleep parameters, and compared the sleep parameters predicted on the wearable with the reference parameters of the manually scored PSG, using Pearson's correlation test and a paired t-test.

To further assess the quality of these parameters, we investigated their clinical significance. We computed the Pearson correlation between a few important sleep parameters and the AHI, to verify that the relation between the AHI and automatically estimated sleep stage parameters based on our EEG wearable were consistent with relations found in previous clinical studies on PSG. This experiment was not aimed at proving the diagnostic power of wearable EEG toward sleep

apnea diagnosis, as there are more suitable respiration-based devices for this purpose. Instead, it aimed to validate the accuracy of AI-based sleep analysis using wearables to replicate clinical findings from hospital studies. If these parameters are reliable enough, such technology could be employed in home-based sleep studies, for instance, to monitor treatment outcomes long-term. The experimental setup for this experiment is detailed in Appendix A.3.

3 | RESULTS

3.1 | Uncertainty-based training data curation

U-PASS refined the training data and consequently improved the sleep staging performance in two key ways. It first determined the minimal measurements needed based on the amount of uncertainty. As the gyroscope measurements did not contribute to decreasing the data uncertainty, it was removed, and only the cross-head EEG channel and the acceleration data were used. Second, the 1% most ambiguous training samples were removed, improving downstream performance of the trained model (Heremans et al., 2024). These two steps improved the sleep staging model with an overall accuracy improvement from 63.9±0.9% to 64.5±0.9% (mean ± standard deviation).

3.2 | Detecting uncertainty at deployment time

Figure 2 shows how uncertainty measures were applied to flag uncertain test data at deployment time, both at the level of whole recordings and at the sample level.

3.2.1 | Detecting uncertain recordings

Figure 2a demonstrates the recording-wise flagging, plotting, for each recording, the two factors of the uncertainty measure. Those factors are the confidence of the closest training data and the distance to these data. A decision boundary was used to identify abnormal recordings. With 14% of the data (7 out of 45 recordings) rejected in this way, the sleep staging accuracy increased from 64.5±0.9% to 68.2±0.9%. The average sleep staging accuracy on the 7 rejected recordings was 43.5±1.8%.

This two-factor uncertainty measure not only flagged problematic recordings, but also gave reasons for the uncertainty. The first factor, the distance to the closest training data, identified so-called 'out-of-distribution' data, recordings that deviate significantly from the training data. Such 'anomalous' signals can be caused by technical artefact or atypical brain activity due to biological differences or a pathology. Figure 2a, recording 1, demonstrates that the recording with the highest distance to the training data is corrupted by large amounts of noise. The second factor, the confidence of the closest training data, identified recordings with an abnormal amount of epochs close to sleep stage decision boundaries. This entails the presence of lots of sleep stage changes or ambiguous brain activity which can not clearly be assigned to a single sleep stage, both of which are telltales of sleep disorders (Shrivastava et al., 2014; Jordan et al., 2014). Figure 2a, recording 2, shows an example with an abnormally high proportion of sleep stage changes and sleep stage N1. N1 is a transitory sleep stage, which is notoriously ambiguous and hard to classify (Rosenberg & Hout, 2013). In conclusion, the two factors in this uncertainty metric guide the clinician to decide which action is required. This could be repeating the measurement for that patient, ignoring the recording, replacing the hardware, or asking for a follow-up study at the hospital if the data looks worrisome.

3.2.2 | Detecting uncertain samples

After flagging the most uncertain recordings, uncertain segments were flagged in the remaining recordings. With 10% of the data flagged as uncertain, the sleep staging accuracy on the remaining data increased from 68.2±0.9% to 71.2±0.9%. Figure 2b shows two segments of data with flagged samples, demonstrating that areas of high uncertainty often corresponded with errors. These errors were primarily attributed to ambiguities arising from sleep stage changes (K. B. Mikkelsen et al., 2017) and the sometimes unclear distinction between classes, for example between N2 and REM. The classification of REM-sleep was generally challenging with the cross-head EEG wearable, without EOG or EMG signals available.

3.3 | Does manual review help for uncertain data?

Figure 3 compares our machine learning model to a clinician in five-stage sleep staging on behind-the-ear wearable EEG data. When independently scoring the full wearable recordings, the clinician only achieved 48.7% accuracy, which was much poorer than the SeqSleepNet model's performance of 66.7% accuracy. When we only used the model's 90% most certain sleep stage predictions, it achieved a higher accuracy of 69.5% on that portion of the data. We then tested whether a human scorer could still improve upon the machine learning decisions for the most uncertain areas. Therefore, a human scorer reviewed the incompletely scored wearable data, filling in the blanks by scoring the remaining 10% most uncertain areas of each recording. The total accuracy then became 66.4%, which was still lower than the performance with fully automated scoring. The accuracy achieved by the human scorer was exactly the same as if we would simply interpolate the sleep stage labels using the surrounding sleep stages. We concluded that this human scorer, not trained specifically to interpret wearable behind-the-ear EEG data, could not contribute to achieving more accurate sleep staging with this type of wearable device. Importantly, this was a limited-scale experiment with a single clinician and one wearable device, so caution

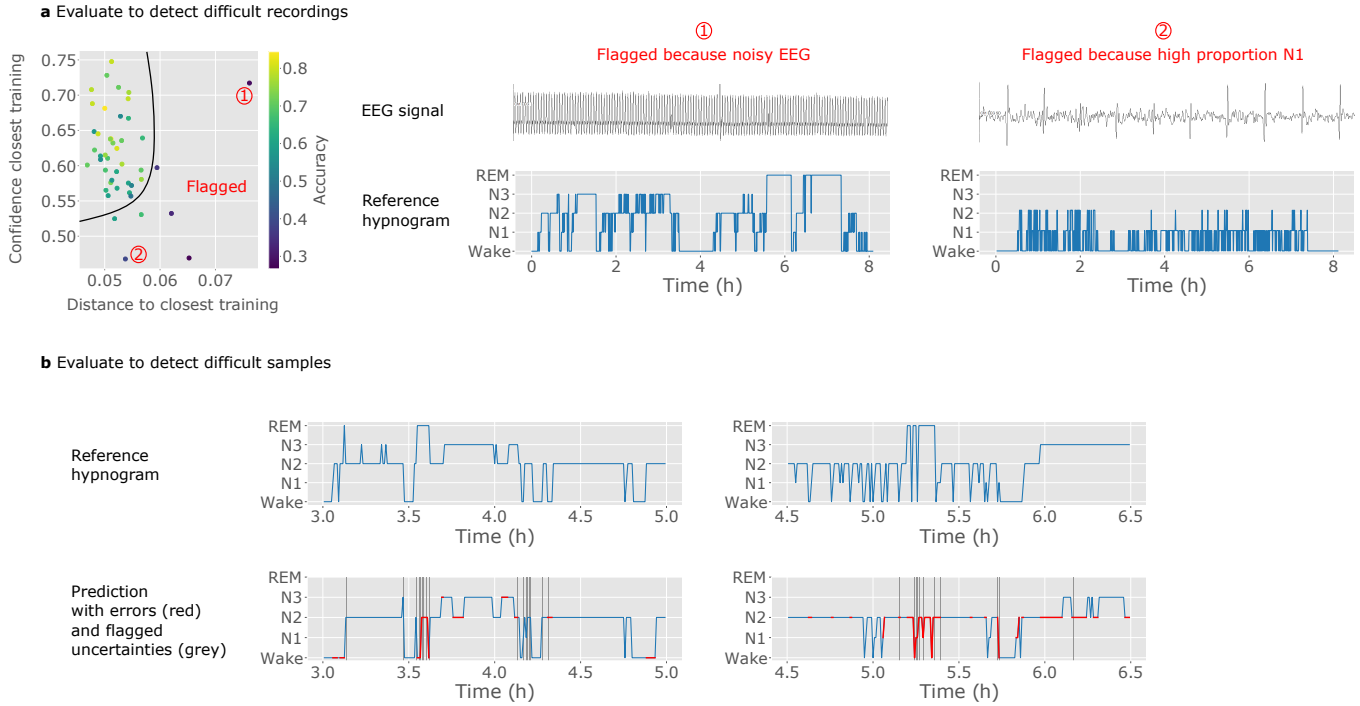


FIGURE 2 U-PASS applied to sleep data captured by a wearable device. (a) In the first evaluation step, recordings with the most prediction uncertainty are rejected as a whole. There can be different reasons for rejecting a recording. The used uncertainty measure takes into account the confidence of the closest training data and the distance from these data. On the left, both factors are visualized for each recording, showing how these influence the accuracy. On the right, two recording snippets and the recordings' hypnograms are shown. The first recording was rejected because of corruption by noise, and the second one was rejected because of the high proportion of sleep stage changes and of the transitory sleep stage N1. (b) In the second evaluation step, uncertainty-based rejection of individual samples can further increase the test accuracy. Samples with uncertainty values higher than the preset threshold were flagged. Two example hypnograms are shown, with the reference sleep stages (top) and the predicted sleep stages (bottom). The samples flagged as uncertain (shown by gray lines) correspond with areas where errors are made (marked in red) and areas with sleep stage changes.

should be exercised when generalizing these results to all clinicians and all wearable devices. In our experiment, the best sleep staging accuracies were achieved with pure automated scoring, when discarding the most uncertain data based on an uncertainty threshold. However, this would of course result in an incomplete hypnogram, as shown in Figure 3.

3.4 | Is automated sleep staging on wearable EEG reliable enough?

As individual sleep parameters are informative for studying long-term sleep evolution and treatment effects, we tested if our machine learning method could reliably predict these parameters based on a one-channel EEG wearable. Based on the predicted sleep stages, we calculated 26 sleep parameters, as shown in Table 2. We then computed the correlation of these estimates with the sleep parameters from the manually scored PSG. Uncertain sleep stage predictions were not discarded in this experiment. The reference and prediction were significantly ($p < 0.01$) correlated for 18 out of 26 sleep parameters, demonstrating that trends and inter-patient differences were accurately captured with the wearable and machine learning approach. A paired t-test showed that despite these many significant correlations, the distributions were still significantly different in 17 of the parameters. Hence, although relative differences and trends were well preserved, the absolute values of most of the estimated parameters did not perfectly correspond to the true parameters. In Appendix C, we additionally show the agreement between some of the most important sleep parameters using Bland-Altman plots.

In addition, we also evaluated if the predicted parameters had the same correlations with the AHI as the reference parameters. Figure 4 shows the correlation between a few important sleep parameters and the AHI. The automatically predicted and reference sleep parameters are both depicted, with a linear trend line and Pearson's correlation coefficient showing how they correlate with the AHI score. The observed correlations between the sleep parameters and AHI score were largely consistent with previous clinical studies and were preserved by automated estimation (extended results in Table D.1).

Specifically, three previous studies were found that directly related sleep apnea to sleep architecture parameters (Basunia et al., 2016; Bianchi, Cash, Mietus, Peng, & Thomas, 2010; Shahveisi et al., 2018;

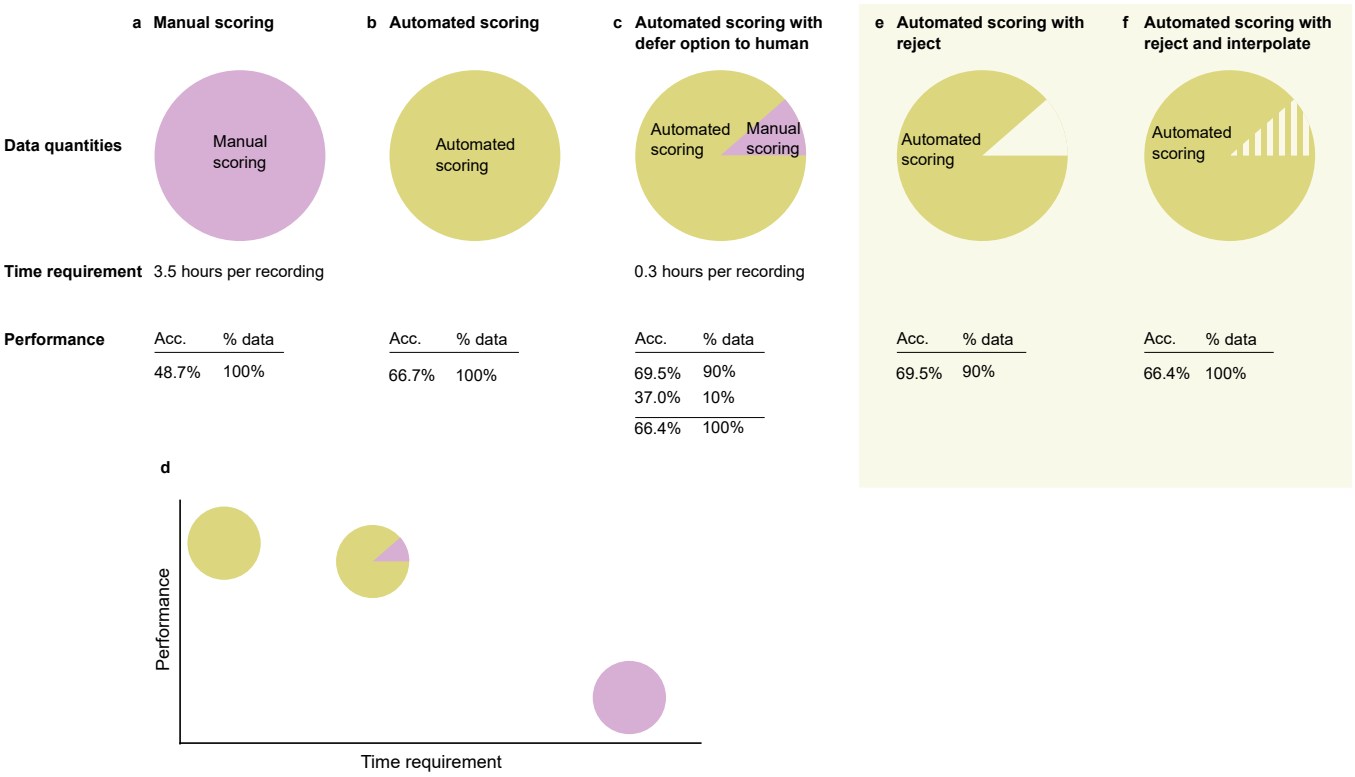


FIGURE 3 Comparison of automated sleep scoring to manual scoring for behind-the-ear wearable EEG recordings. (a) With fully manual scoring, the clinician spent 3.5 hours scoring each recording. (b) With fully automated scoring by our trained sleep model, the performance was higher without requiring any time from a clinician. (c) When only the 10% most uncertain predictions of the model were deferred to and scored by the clinician, the performance was still slightly worse than with fully automated scoring. (d) The performance and time requirement of each scenario is presented visually. (e) When a clinician can not score these uncertain samples better than a machine learning model, what can we do with the uncertain predictions? One option is to simply not predict anything, omitting these most uncertain predictions. When leaving out the 10% most uncertain predictions, the remaining 90% were scored much more accurately. This, however, results in an incomplete hypnogram. (f) Another option is to use simple interpolation of these most uncertain sleep stage labels with the surrounding sleep stages. This approach led to the same accuracy as the review by a clinician, and hence, worse results than taking the original model predictions.

Redline et al., 2004). The study by Basunia et al. (Basunia et al., 2016) was the only that directly correlated the parameters with the AHI. The study by Redline et al. (Redline et al., 2004) was conducted with the respiratory disturbance index (RDI), another sleep apnea severity score. The other two studies stratified patients into three groups for a more coarse analysis. Zooming in on a few important parameters, Basunia et al. (Basunia et al., 2016) found that sleep efficiency was negatively correlated with the AHI score, as confirmed by our results (both in the predicted and reference sleep parameters). In the relative times spent in each sleep stage, they found significant correlations for every sleep stage. For N1, we found the same positive correlation with AHI, based on both the predicted and reference sleep stages. For N3, surprisingly, the reference sleep parameters did not find a significant negative correlation like all three previous studies (Basunia et al., 2016; Bianchi et al., 2010; Shahveisi et al., 2018), but our predicted sleep parameter did. For REM, the reference and predicted parameters were not significantly negatively correlated with the AHI, unlike in (Basunia et al., 2016). For other sleep parameters, we did not find conclusive results in the literature, but our predicted and reference parameters agreed on most correlations, like the significant positive correlation of AHI with the amount of awakenings, sleep stage transitions and amount of wake after sleep onset (WASO).

One noteworthy difference between the predicted and reference parameters was the sleep onset as the time to the first N2 segment. The N2 sleep onset was significantly negatively correlated with the AHI score based on the reference sleep stages, but not based on the predicted sleep stages. The results from Basunia et al. (Basunia et al., 2016) agreed with the predicted sleep stages. The fact that the predictions and reference did not entirely agree for this measure was not completely unexpected, given the fact that the sleep onset is hard to reliably estimate, as it is often characterized by multiple transitions between Wake, N1 and N2. Another notable difference between the reference and predicted parameters was the amount of transitions to N3, negatively

correlating with the AHI score based on the predicted sleep stages, but not based on the reference sleep stages.

Apart from a few exceptions, we conclude that the machine learning predictions based on wearable EEG resulted in reliable estimates of individual sleep parameters, and that the estimated parameters had the same correlations with the AHI as the the parameters based on the clinician's evaluation of the PSG.

4 | DISCUSSION

To our knowledge, this is the first study that used uncertainty to improve the automated analysis of wearable physiological signals. With our novel framework U-PASS, we showed that machine learning can reliably perform sleep staging based on wearable behind-the-ear EEG measurements from elderly patients. We compared the ability of our state-of-the-art machine learning model to a clinician in reading these non-standard physiological data, demonstrating that our machine learning model outperformed the human EEG reader who was not trained to interpret such wearable EEG. The sleep parameters derived from algorithmically determined sleep stages agreed with the reference values derived from the expert-annotated PSG as indicated by Pearson's correlation test. This means that inter-patient differences and trends were well preserved with the wearable AI approach, although significant differences were found in the absolute values of the parameters. We further investigated the potential clinical efficacy of the automated pipeline by confirming known correlations of these sleep parameters with a clinical metric, the AHI. We showed that the automated pipeline could indeed reproduce these findings, proving its clinical relevance. Therefore, relative changes in sleep parameters should be regarded as more reliable indicators than absolute values.

With U-PASS, our uncertainty-guided pipeline (Heremans et al., 2024), we introduced the following improvements to the standard machine learning framework. At training time, we decided which channels to use based on uncertainty, and discarded the training data with the most data uncertainty to improve the model. The optimal performance was achieved when 1% of the training data were discarded. These improvements introduced small improvements to the model performance, consistent with results in different datasets (Heremans et al., 2024). At deployment time, we used a combination of detecting uncertainty related to decision boundaries and out-of-distribution or 'anomalous' data to flag uncertainties. On a coarse level, we flagged recordings with high uncertainty overall, and on a finer level, we flagged uncertain samples. We investigated whether manual review could help to correct these uncertain samples with a clinician-in-the-loop approach, and concluded that this was not true in this particular case.

More specifically, this study revealed that a sleep expert was unable to manually correct sleep stages from a wearable device with a behind-the-ear cross-head EEG setup in an elderly population. We have to emphasize that this experiment was limited to a single clinician and one wearable device setup, so we do not have conclusive evidence that this finding can be extrapolated to all sleep experts and all wearables. There is a possibility that sleep scoring on behind-the-ear signals is more difficult to score than from other wearables. However, another study confirmed that machine learning outperforms human scorers on a sleep staging task using around-the-ear EEG, this time in healthy participants (K. B. Mikkelsen et al., 2019). The further complete lack of studies with wearable EEG being scored by clinicians confirms the inherent difficulty. In any case, sleep staging rules are only defined for standard EEG/PSG channels, and so, for any wearable deviating from this standard, the rules would need to be adapted and clinicians would need to be extensively retrained to obtain reliable scoring with minimal interrater variability. We argue that it would be prohibitively costly and time consuming to achieve this for the plethora of wearables that are becoming available.

On the contrary, for PSG recordings, manual review of automated sleep scoring proves beneficial and feasible. Two studies that investigated a clinician-in-the-loop approach for automated PSG scoring are the study by Kang et al. (Kang et al., 2021), experimentally demonstrating that human scorers can correct automated PSG scoring, and U-PASS (Heremans et al., 2024), where we showed how clinicians can even iteratively improve a sleep staging model. Hence, a future research direction for wearables could be to further elevate their signal quality and to rescale and transform the wearable EEG signals to better resemble standard EEG derivations, such that a clinician-in-the-loop approach becomes feasible.

Overall, the major advantage of using our uncertainty-guided framework for remote monitoring applications is the increase in reliability. The accuracy on the test set was improved by an absolute 0.6% through the improvements of the training set. It was then improved by an absolute 3.7% by discarding the seven most problematic recordings of the test set. Using only 90% of the most certain predictions at a segment-level led to another improvement of an absolute 3.0%. Remote monitoring with wearables is prone to noise and variability in data quality, and manual review of data quality in large samples is typically not feasible. Therefore, detecting such uncertainties is crucial to ensure the trustworthiness and usefulness of an automated data processing system. Moreover, automatic sleep staging systems typically perform worse in elderly and diseased populations (Phan & Mikkelsen, 2022; Baumert, Hartmann, & Phan, 2023; Patanaik, Ong, Gooley, Ancoli-Israel, & Chee, 2018), where sleep disturbances and neurodegenerative disorders result in more fragmented sleep which is difficult to score. We demonstrated that applying uncertainty measures can increase the accuracy in such populations of interest, which is essential to adopt machine learning in clinical practice. In conclusion, frameworks like U-PASS improve the reliability of automated machine learning systems, which is crucial for the success of long-term at-home wearable sleep monitoring systems.

A second advantage of our method is that it is completely model-agnostic. The U-PASS pipeline can be used for any model which returns output probabilities, as it only needs access to those predicted probabilities. As such, it could even be deployed in other applications such as

TABLE 2 A total of 26 sleep parameters are calculated based on the hypnograms obtained with our deep neural network and the reference hypnograms, as annotated by a clinician based on PSG. Pearson's correlation coefficient (ρ) is calculated between the sleep parameters calculated with both approaches, and the t-test for related samples is performed to identify significant differences between both distributions.

	Predicted[a]	Reference[a]	Pearson's ρ[b]		t-test[c]
			ρ	p-value	p-value
TST (hr)	5.09 ± 1.14	5.50 ± 1.49	**0.91**	0.00	**0.00**
Sleep efficiency (%)	0.65 ± 0.09	0.70 ± 0.14	**0.85**	0.00	**0.00**
W (hr)	2.84 ± 1.10	2.44 ± 1.29	**0.86**	0.00	**0.00**
N1 (hr)	0.09 ± 0.10	0.72 ± 0.63	**0.63**	0.00	**0.00**
N2 (hr)	3.74 ± 0.92	2.72 ± 0.90	**0.55**	0.00	**0.00**
N3 (hr)	0.63 ± 0.42	1.13 ± 0.88	**0.52**	0.00	**0.00**
REM (hr)	0.64 ± 0.47	0.93 ± 0.52	**0.58**	0.00	**0.00**
N1 rel (% TST)	0.02 ± 0.02	0.14 ± 0.13	**0.61**	0.00	**0.00**
N2 rel (% TST)	0.74 ± 0.12	0.51 ± 0.14	0.27	0.10	**0.00**
N3 rel (% TST)	0.12 ± 0.07	0.19 ± 0.13	**0.42**	0.01	**0.00**
REM (% TST)	0.12 ± 0.08	0.16 ± 0.07	**0.53**	0.00	**0.00**
Awakenings	33.76 ± 16.85	36.16 ± 22.89	**0.79**	0.00	0.30
Sleep stage transitions	90.38 ± 40.29	135.45 ± 69.69	**0.81**	0.00	**0.00**
Transitions to REM	8.41 ± 4.83	7.87 ± 5.23	0.25	0.13	0.59
Transitions to N3	5.03 ± 2.50	8.11 ± 9.09	0.22	0.18	**0.04**
Time between REM (hr)	0.56 ± 0.29	0.61 ± 0.39	0.14	0.43	0.53
Time between N3 (hr)	0.78 ± 0.40	0.71 ± 0.52	0.16	0.40	0.50
Sleep onset N1 (hr)	1.63 ± 0.67	0.68 ± 0.74	0.30	0.07	**0.00**
Sleep onset N2 (hr)	0.42 ± 0.36	0.62 ± 0.74	**0.63**	0.00	**0.05**
WASO (hr)	1.78 ± 0.86	1.54 ± 1.02	**0.77**	0.00	**0.03**
Awakenings (/TST)	0.06 ± 0.03	0.06 ± 0.04	**0.73**	0.00	0.74
Sleep stage transitions (/TST)	0.15 ± 0.06	0.21 ± 0.12	**0.68**	0.00	**0.00**
Transitions to REM (/TST)	0.01 ± 0.01	0.01 ± 0.01	0.09	0.57	0.45
Transitions to N3 (/TST)	0.01 ± 0.00	0.01 ± 0.01	0.12	0.48	0.10
WASO (/TST)	0.37 ± 0.20	0.32 ± 0.27	**0.71**	0.00	0.12
REM latency (hr)	2.03 ± 1.11	2.08 ± 1.37	**0.74**	0.00	0.76

[a]Represented as mean $\pm$ standard deviation.
[b]Significant correlations with $p \leq 0.01$ are marked in bold.
[c]Significant differences with $p \leq 0.05$ are marked in bold.
Abbreviations: WASO: Wake after sleep onset, TST: total sleep time (always expressed in hours), hr: hours.

remote seizure monitoring or cardiac monitoring. The uncertainty metrics can also be changed to any other post-hoc uncertainty metric. We decided to use a distance-weighted confidence of neighboring training samples as the uncertainty metric (see Section 2.3), but this could be swapped for the Shannon entropy or other uncertainty estimation methods such as Monte Carlo dropout (Gal & Ghahramani, 2015). Depending on the application and needs, the framework could be adapted in multiple ways. Either the recording-wise uncertainty estimation or the sample-wise uncertainty estimation could be excluded, only performing uncertainty estimation on one of the two levels. In some applications, five-class sleep staging may not be needed. For example, if screening for insomnia, the primary parameters of interest could be the total sleep time and time spent awake, requiring only a two-class classification between sleep and wake.

A third benefit of our method is the fact that it can easily be integrated in a clinical workflow with minimal adjustments and requirements. The training requirements of U-PASS do not impact clinical practice, since model training can be done in advance, outside of clinical routine. This ensures minimal disruption to clinical workflows. At test time, U-PASS only requires about 200 MB of working memory to compute the uncertainty measures of a test recording (in case of a training dataset of 50 recordings, and scaling linearly with the amount of training data). Based on these minimal requirements and the complete model-agnostic nature of our method, U-PASS can feasibly be integrated into any AI-based automated sleep scoring system in daily clinical practice.

A fourth benefit of using our uncertainty-guided framework is that the identified uncertainty in itself can also provide useful information for the diagnosis of sleep disorders. We used an uncertainty metric composed of two factors, with a distance factor detecting out-of-distribution data and a confidence factor detecting closeness to decision boundaries. These two factors give a reason for uncertainty, which can be taken into account to diagnose sleep disorders. For example, we showed that a fragmented sleep, typically associated with OSA, leads to more uncertainty as more segments lie on boundaries between different sleep stages (see hypnogram 2 in Figure 2a, and Figure B.1). A low confidence factor in this uncertainty metric could be a relevant marker when screening for OSA. On the other hand, the distance factor will point to atypical, out-of-distribution data. This could be useful for detecting abnormal EEG activity, which can be present in epilepsy or dementia.

We showed that we could identify uncertain predictions such that the remaining ones were more accurate. However, this does not answer the question what we should do with the identified uncertainties, especially if a clinician may not be able to correct them. As already discussed, the ideal solution we envisage would be to transform the

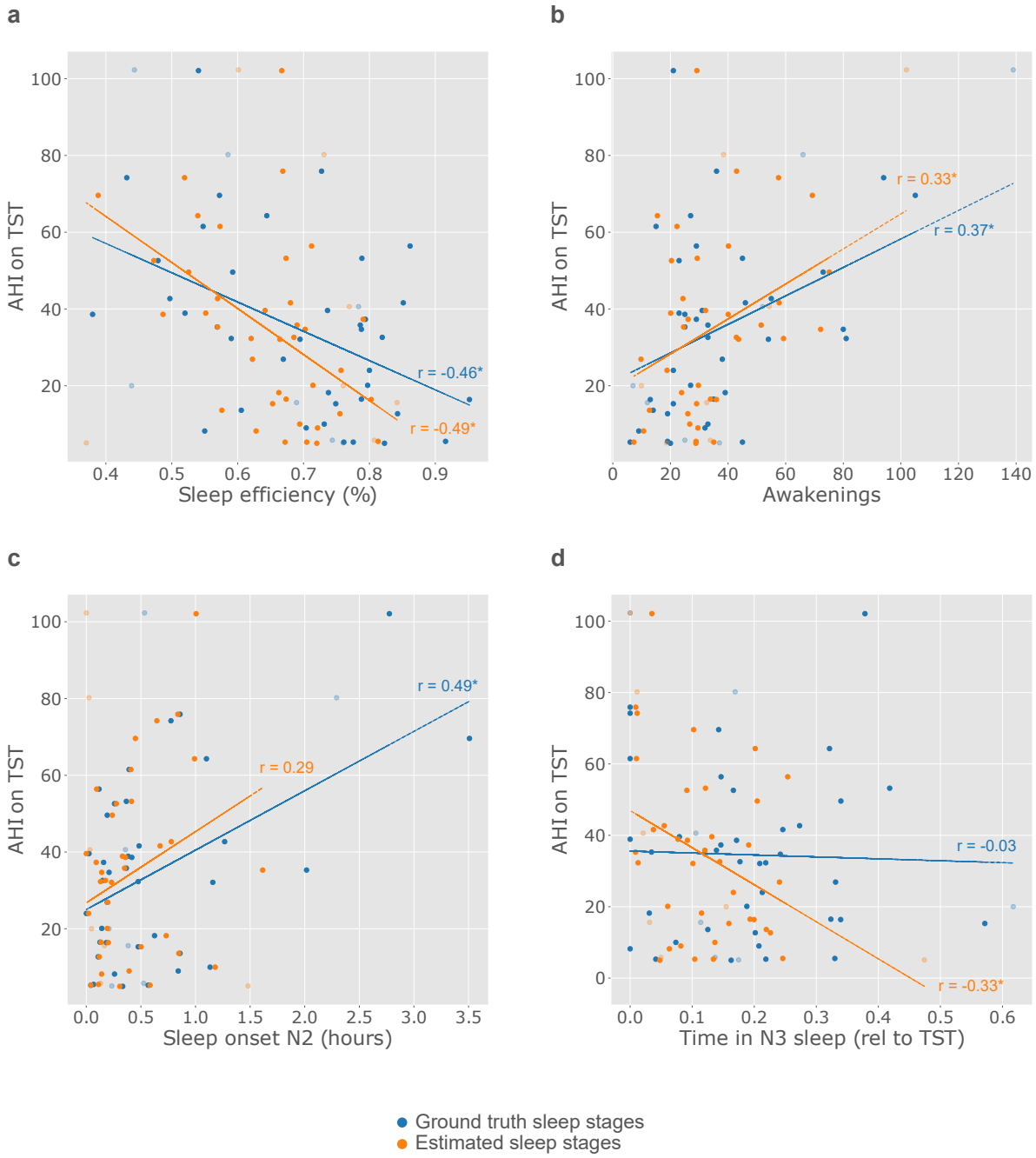


FIGURE 4 Clinical significance of sleep parameters derived from model predictions and PSG reference. These scatter plots show the relation between the apnea-hypopnea index on the total sleep time (AHI on TST) and some important sleep parameters with a linear trend line and Pearson's correlation coefficient (r). The light-colored dots represent recordings that were discarded because of high levels of uncertainty. The sleep parameters are both calculated based on the reference and the predicted sleep stages. Most sleep parameters that correlate well with the AHI score do so consistently, both when calculated based on the reference and predicted sleep stages. The sleep efficiency (a) and number of awakenings (b) are two examples with excellent concordance. Sleep onset (c) and relative time in N3 sleep (d) are two exceptions, for which the results are not consistent for the predicted and reference sleep stages. See Table D.1 for the full overview.

wearable EEG signals to closely resemble standard EEG signals, enabling clinicians to review the uncertain predictions. However, as this solution is not yet available to our best knowledge, uncertainty estimation methods like U-PASS are still, if not even more, invaluable for wearable sleep monitoring. We showed how U-PASS was able to identify bad-quality

recordings with overall high uncertainty. Since wearables have the advantage of being cheap recording methods, problematic recordings can be identified, discarded, and repeated. Consequently, we showed how U-PASS was also able to flag uncertain regions in the hypnogram. Discarding these uncertain predictions led to more accurate sleep staging, but an incomplete hypnogram. If the goal is to estimate sleep parameters, these uncertain predictions should not simply be discarded. As sleep parameters aggregate information from a whole night, small variations in sleep scoring generally do not have a large impact on these parameters. Hence, if the uncertainties are limited to small segments and sleep stage transitions, we would recommend to use fully automated sleep staging (the most reliable option as per Figure 3) and the resulting sleep parameters. On average, these proved to be relatively well aligned with the reference PSG-based sleep parameters (Table 2). If large segments of the recording are uncertain, the clinician can either decide to repeat the recording or perform in-hospital PSG to investigate whether the uncertainty stems from a sleep disorder. If not all sleep parameters are of interest, we should remark that repeating the recording may not be needed, depending on where the uncertainty is concentrated. For example, uncertainty around N2 and N3 sleep is unlikely to affect total sleep time and WASO, and can thus largely be disregarded for these parameters. Therefore, sample-level uncertainty can provide insights into the confidence level of specific sleep parameters. Future research should explore how uncertainty at a sample level can be mapped to confidence estimates for sleep parameters, adding value in the overall diagnostic process.

Two additional practical considerations are worth mentioning when it comes to implementing U-PASS in clinical practice. First, the cost aspect of sleep studies with wearables still needs to be investigated and optimized. Repeating a measurement is the most reliable way to handle a poor-quality recording, but funding schemes may not cover the costs of potential failed sleep studies. One potential strategy to mitigate the likelihood of failed recordings for diagnostic or screening sleep studies using wearables is to provide the patient with the device for an extended period, such as a week, allowing for multiple trials. After each night, uncertainty metrics can serve as direct feedback to the patient, indicating whether the recording was successful or additional recordings are needed. In case of multiple successful recordings, averaging the results of these subsequent measurements may provide a more robust and reliable estimate of sleep, and between-day variability might be informative as well.

A second practical consideration involves effectively communicating uncertainty to clinicians who are not AI experts. To visualize sample-level uncertainty, we recommend displaying the raw signals segmented with the corresponding hypnogram underneath. The hypnogram should be color-coded to indicate the reliability of the predictions: one color for high-quality, reliable predictions and other colors for medium or high uncertainty, similar to the grey areas in Figure 2b.

A limitation of the present study is the need for determining the uncertainty thresholds for marking uncertain regions. These thresholds are subject to adjustment based on specific application demands and precision requirements, as well as the need for scoring completeness. When marking complete recordings as uncertain, the uncertainty threshold can be set lower if it is cheap to replace bad quality recordings by new ones, or when the goal is to detect anomalous recordings with high sensitivity for anomalous brain activity screening. When marking individual samples as uncertain, the threshold depends on the utilization of the resulting hypnogram, and on which sleep parameters need be estimated. The thresholds can be advantageous in providing the flexibility to deploy this framework in different applications, but domain expertise is needed to choose the thresholds appropriately for every setting.

Our results have implications for the future of AI in sleep monitoring. The vast amount of data that wearable technologies can gather are quite certainly going to surpass the capacities of human review. In this work, we showed another problem, which is that sleep experts may not always be able to interpret data from wearables without being specifically trained for this. Therefore, we propose combining AI and wearables to conduct large-scale screenings of populations or for long-term sleep monitoring, particularly to monitor treatment effects. These are operations which are too costly and time-consuming to be achievable without the help of wearables and AI. We have shown that AI could enable such large-scale operations, by providing fine-grained labeling of the wearable data, and aggregating this information into sleep parameters. These parameters can then be presented in sleep reports for interpretation by human sleep experts. These parameters should be considered approximations of those observed in PSG and must be interpreted with caution. Nevertheless, relative differences between patients and populations are expected to be accurately reflected. Uncertainty estimation plays a crucial role in such an AI-based sleep monitoring approach. We have shown that improvements can be achieved through both uncertainty-informed refinement of the training data and flagging of uncertain data during deployment. In conclusion, an uncertainty-guided framework is essential to guarantee the reliability of AI-based analysis for wearable data in home-based sleep monitoring.

AUTHOR CONTRIBUTIONS

ERMH was the major contributor in the conceptualization of the paper, the code and experiments, and writing the manuscript. LVdB was the clinician who performed the manual review of the uncertain data. NS mainly contributed in the conceptualization of the uncertainty quantification techniques. PB, BB, DT, and AD performed the clinical study in which the data were collected. MVDB, MvdS, and MDV supervised the work and helped with conceptualization of the study. All authors read and approved the final manuscript.

FUNDING STATEMENT

This research was supported by the Research Foundation—Flanders (FWO; GrantNo. 1SC2921N) for ERMH; by the 'Bijzonder Onderzoeksfonds KU Leuven (BOF)' ('Prevalence of Epilepsy and Sleep Disturbances in Alzheimer Disease—C24/18/097' and 'Starting Grant: Artificial Intelligence (AI)-enabled mining of big longitudinal datasets collected with wearable sensors') for MDV; by the Flemish Government (AI Research

Program) for MDV; and by the Cystic Fibrosis Trust for NS. MDV and ERMH are affiliated to Leuven AI—KU Leuven institute for AI, B-3000, Leuven, Belgium.

DATA AVAILABILITY

The data analysed during the current study are not publicly available due to ethical reasons, but may be made available from the corresponding author on reasonable request. Requests for data shall be reviewed and responded to in a manner that is consistent with the informed consent form signed by the participants.

CODE AVAILABILITY

The code for this paper was made using Python. It is available in Github and can be accessed via this link `github.com/elisabethRMH/U-PASS`.

CONFLICT OF INTEREST

All authors declare no financial or non-financial competing interests.

References

Arnal, P. J., Thorey, V., Debellemaniere, E., Ballard, M. E., Bou Hernandez, A., Guillot, A., ... Sauvet, F. (2020, nov). The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep*, *43*(11), 1–13. Retrieved from `https://academic.oup.com/sleep/article/43/11/zsaa097/5841249` doi: 10.1093/SLEEP/ZSAA097

Basunia, M., Fahmy, S. A., Schmidt, F., Agu, C., Bhattarai, B., Oke, V., ... Quist, J. (2016, sep). Relationship of symptoms with sleep-stage abnormalities in obstructive sleep apnea-hypopnea syndrome. *Journal of Community Hospital Internal Medicine Perspectives*, *6*(4). Retrieved from `/pmc/articles/PMC5016742//pmc/articles/PMC5016742/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5016742/` doi: 10.3402/JCHIMP.V6.32170

Baumert, M., Hartmann, S., & Phan, H. (2023). Automatic sleep staging for the young and the old – evaluating age bias in deep learning. *Sleep Medicine*, *107*, 18-25. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1389945723001260` doi: https://doi.org/10.1016/j.sleep.2023.04.002

Berry, R., Quan, S., & Abreu, A. (2020). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6*. Darien: American Academy of Sleep Medicine.

Bianchi, M. T., Cash, S. S., Mietus, J., Peng, C. K., & Thomas, R. (2010). Obstructive Sleep Apnea Alters Sleep Stage Transition Dynamics. *PLoS ONE*, *5*(6). Retrieved from `/pmc/articles/PMC2893208//pmc/articles/PMC2893208/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2893208/` doi: 10.1371/JOURNAL.PONE.0011356

Bubu, O. M., Andrade, A. G., Umasabor-Bubu, O. Q., Hogan, M. M., Turner, A. D., de Leon, M. J., ... Osorio, R. S. (2020). Obstructive sleep apnea, cognition and alzheimer's disease: A systematic review integrating three decades of multidisciplinary research. *Sleep Medicine Reviews*, *50*, 101250. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1087079219302187` doi: https://doi.org/10.1016/j.smrv.2019.101250

D'Rozario, A. L., Chapman, J. L., Phillips, C. L., Palmer, J. R., Hoyos, C. M., Mowszowski, L., ... Naismith, S. L. (2020). Objective measurement of sleep in mild cognitive impairment: A systematic review and meta-analysis. *Sleep Medicine Reviews*, *52*, 101308. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1087079220300514` doi: https://doi.org/10.1016/j.smrv.2020.101308

Finan, P. H., Richards, J. M., Gamaldo, C. E., Han, D., Leoutsakos, J. M., Salas, R., ... Smith, M. T. (2016). Validation of a Wireless, Self-Application, Ambulatory Electroencephalographic Sleep Monitoring Device in Healthy Volunteers. *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*, *12*(11), 1443–1451. Retrieved from `https://pubmed.ncbi.nlm.nih.gov/27707438/` doi: 10.5664/JCSM.6262

Gal, Y., & Ghahramani, Z. (2015, jun). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *33rd International Conference on Machine Learning, ICML 2016*, *3*, 1651–1660. Retrieved from `https://arxiv.org/abs/1506.02142v6`

Guillot, A., & Thorey, V. (2021). RobustSleepNet: Transfer Learning for Automated Sleep Staging at Scale. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 1441–1451. doi: 10.1109/TNSRE.2021.3098968

Heremans, E. R. M., Phan, H., Borzée, P., Buyse, B., Testelmans, D., & De Vos, M. (2022, jun). From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging. *Journal of Neural Engineering*, *19*(3), 036044. Retrieved from `https://iopscience.iop.org/article/10.1088/1741-2552/ac6ca8https://iopscience.iop.org/article/10.1088/1741-2552/ac6ca8/meta` doi: 10.1088/1741-2552/AC6CA8

Heremans, E. R. M., Seedat, N., Buyse, B., Testelmans, D., van der Schaar, M., & De Vos, M. (2024, mar). U-PASS: An uncertainty-guided deep learning pipeline for automated sleep staging. *Computers in Biology and Medicine*, *171*, 108205. Retrieved from `https://linkinghub.elsevier.com/retrieve/pii/S0010482524002890` doi: 10.1016/J.COMPBIOMED.2024.108205

Irwin, M. R., & Vitiello, M. V. (2019). Implications of sleep disturbance and inflammation for alzheimer's disease dementia. *The Lancet Neurology*, *18*(3), 296-306. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1474442218304502` doi: https://doi.org/10.1016/S1474-4422(18)30450-2

Jordan, A. S., McSharry, D. G., & Malhotra, A. (2014). Adult obstructive sleep apnoea. *The Lancet*, *383*(9918), 736-747. Retrieved from https://www.sciencedirect.com/science/article/pii/S0140673613607345 doi: https://doi.org/10.1016/S0140-6736(13)60734-5

Kang, D. Y., DeYoung, P. N., Tantiongloc, J., Coleman, T. P., & Owens, R. L. (2021, sep). Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine. *npj Digital Medicine 2021 4:1*, *4*(1), 1–9. Retrieved from https://www.nature.com/articles/s41746-021-00515-3 doi: 10.1038/s41746-021-00515-3

Lee, Y. J., Lee, J. Y., Cho, J. H., & Choi, J. H. (2022, jan). Interrater reliability of sleep stage scoring: a meta-analysis. *Journal of Clinical Sleep Medicine*, *18*(1), 193–202. Retrieved from https://jcsm.aasm.org/doi/10.5664/jcsm.9538https://jcsm.aasm.org/ doi: 10.5664/JCSM.9538

Leng, Y., McEvoy, C. T., Allen, I. E., & Yaffe, K. (2017, 10). Association of Sleep-Disordered Breathing With Cognitive Function and Risk of Cognitive Impairment: A Systematic Review and Meta-analysis . *JAMA Neurology*, *74*(10), 1237-1245. Retrieved from https://doi.org/10.1001/jamaneurol.2017.2180 doi: 10.1001/jamaneurol.2017.2180

Lucey, B. P., Wisch, J., Boerwinkle, A. H., Landsness, E. C., Toedebusch, C. D., McLeland, J. S., ... Holtzman, D. M. (2021, 10). Sleep and longitudinal cognitive performance in preclinical and early symptomatic Alzheimer's disease. *Brain*, *144*(9), 2852-2862. Retrieved from https://doi.org/10.1093/brain/awab272 doi: 10.1093/brain/awab272

Mander, B. A., Winer, J. R., & Walker, M. P. (2017). Sleep and human aging. *Neuron*, *94*(1), 19-36. Retrieved from https://www.sciencedirect.com/science/article/pii/S0896627317300880 doi: https://doi.org/10.1016/j.neuron.2017.02.004

Martin, J. L., & Hakim, A. D. (2011, jun). Wrist Actigraphy. *Chest*, *139*(6), 1514–1527. doi: 10.1378/CHEST.10-1872

Mikkelsen, K., & De Vos, M. (2018, jan). Personalizing deep learning models for automatic sleep staging. *pre-print*. Retrieved from http://arxiv.org/abs/1801.02645

Mikkelsen, K. B., Ebajemito, J. K., Bonmati-Carrion, M. A., Santhi, N., Revell, V. L., Atzori, G., ... De Vos, M. (2019, apr). Machine-learning-derived sleep–wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy. *Journal of Sleep Research*, *28*(2), e12786. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.12786 doi: 10.1111/jsr.12786

Mikkelsen, K. B., Phan, H., Rank, M. L., Hemmsen, M. C., de Vos, M., & Kidmose, P. (2017). *Light-weight sleep monitoring: electrode distance matters more than placement for automatic scoring* .

Miner, B., & Kryger, M. H. (2017). Sleep in the aging population. *Sleep Medicine Clinics*, *12*(1), 31-38. Retrieved from https://www.sciencedirect.com/science/article/pii/S1556407X16301023 Social and Economic Dimensions of Sleep Disorders. doi: https://doi.org/10.1016/j.jsmc.2016.10.008

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017, 05). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236-1246. Retrieved from https://doi.org/10.1093/bib/bbx044 doi: 10.1093/bib/bbx044

Pallayova, M., Donic, V., Gresova, S., Peregrim, I., & Tomori, Z. (2010). Do differences in sleep architecture exist between persons with type 2 diabetes and nondiabetic controls? *Journal of diabetes science and technology*, *4*(2), 344–352. Retrieved from https://pubmed.ncbi.nlm.nih.gov/20307395/ doi: 10.1177/193229681000400215

Pase, M. P., Harrison, S., Misialek, J. R., Kline, C. E., Cavuoto, M., Baril, A.-A., ... Himali, J. J. (2023, 07). Sleep Architecture, Obstructive Sleep Apnea, and Cognitive Function in Adults. *JAMA Network Open*, *6*(7), e2325152-e2325152. Retrieved from https://doi.org/10.1001/jamanetworkopen.2023.25152 doi: 10.1001/jamanetworkopen.2023.25152

Patanaik, A., Ong, J. L., Gooley, J. J., Ancoli-Israel, S., & Chee, M. W. L. (2018, 03). An end-to-end framework for real-time automatic sleep stage classification. *Sleep*, *41*(5), zsy041. Retrieved from https://doi.org/10.1093/sleep/zsy041 doi: 10.1093/sleep/zsy041

Perez-Pozuelo, I., Zhai, B., Palotti, J., Mall, R., Aupetit, M., Garcia-Gomez, J. M., ... Fernandez-Luque, L. (n.d.). The future of sleep health: a data-driven revolution in sleep science and medicine. Retrieved from https://doi.org/10.1038/s41746-020-0244-4 doi: 10.1038/s41746-020-0244-4

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., & Igel, C. (2021, dec). U-Sleep: resilient high-frequency sleep staging. *npj Digital Medicine*, *4*(1), 72. Retrieved from http://www.nature.com/articles/s41746-021-00440-5 doi: 10.1038/s41746-021-00440-5

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., & De Vos, M. (2018, sep). SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(3), 400–410. Retrieved from http://dx.doi.org/10.1109/TNSRE.2019.2896659 doi: 10.1109/TNSRE.2019.2896659

Phan, H., Chén, O. Y., Koch, P., Lu, Z., Mcloughlin, I., Mertins, A., & De Vos, M. (2021). Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning. *IEEE Transactions on Biomedical Engineering*, *68*(6), 1787–1798. doi: 10.1109/TBME.2020.3020381

Phan, H., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., & De Vos, M. (2020, jul). XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Retrieved from http://arxiv.org/abs/2007.05492http://dx.doi.org/10.1109/TPAMI.2021.3070057 doi: 10.1109/TPAMI.2021.3070057

Phan, H., & Mikkelsen, K. (2022, apr). Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement*, *43*(4), 04TR01. Retrieved from https://iopscience.iop.org/article/10.1088/1361-6579/ac6049 doi: 10.1088/1361-6579/ac6049

Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A., & De Vos, M. (2021, may). SleepTransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification. *IEEE Transactions on Biomedical Engineering*, *69*(8), 2456–2467. Retrieved from http://arxiv.org/abs/2105.11043http://dx.doi.org/10.1109/TBME.2022.3147187 doi: 10.1109/TBME.2022.3147187

Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, *14*, 156-180. doi: 10.1109/RBME.2020.3013489

Qi, Y., Soh, C. B., Gunawan, E., Low, K. S., & Thomas, R. (2014, aug). An Ultrasonic Contactless Sensor for Breathing Monitoring. *Sensors (Basel, Switzerland)*, *14*(8), 15371. Retrieved from /pmc/articles/PMC4179033//pmc/articles/PMC4179033/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4179033/ doi: 10.3390/S140815371

Radha, M., Fonseca, P., Moreau, A., Ross, M., Cerny, A., Anderer, P., ... Aarts, R. M. (n.d.). A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. Retrieved from https://doi.org/10.1038/s41746-021-00510-8 doi: 10.1038/s41746-021-00510-8

Redline, S., Kirchner, H. L., Quan, S. F., Gottlieb, D. J., Kapur, V., & Newman, A. (2004, feb). The Effects of Age, Sex, Ethnicity, and Sleep-Disordered Breathing on Sleep Architecture. *Archives of Internal Medicine*, *164*(4), 406–418. Retrieved from https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/216692 doi: 10.1001/ARCHINTE.164.4.406

Rosenberg, R. S., & Hout, S. V. (2013). The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, *9*(1), 81. Retrieved from /pmc/articles/PMC3525994//pmc/articles/PMC3525994/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3525994/ doi: 10.5664/JCSM.2350

Seedat, N., Crabbé, J., Bica, I., & van der Schaar, M. (2022). Data-iq: Characterizing subgroups with heterogeneous outcomes in tabular data. *Advances in Neural Information Processing Systems*, *35*, 23660–23674.

Shahveisi, K., Jalali, A., Moloudi, M. R., Moradi, S., Maroufi, A., & Khazaie, H. (2018). Sleep Architecture in Patients With Primary Snoring and Obstructive Sleep Apnea. *Basic and Clinical Neuroscience*, *9*(2), 147–156. Retrieved from http://bcn.iums.ac.ir/article-1-895-en.html doi: 10.29252/NIRP.BCN.9.2.147

Shrivastava, D., Jung, S., Saadat, M., Sirohi, R., & Crewson, K. (2014, jan). How to interpret the results of a sleep study. *Journal of Community Hospital Internal Medicine Perspectives*, *4*(5), 24983. Retrieved from /pmc/articles/PMC4246141//pmc/articles/PMC4246141/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4246141/ doi: 10.3402/JCHIMP.V4.24983

Siengsukon, C., Al-Dughmi, M., Al-Sharman, A., & Stevens, S. (2015). Sleep Parameters, Functional Status, and Time Post-Stroke are Associated with Offline Motor Skill Learning in People with Chronic Stroke. *Frontiers in Neurology*, *6*(OCT). Retrieved from /pmc/articles/PMC4626644//pmc/articles/PMC4626644/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4626644/ doi: 10.3389/FNEUR.2015.00225

Sridhar, N., Shoeb, A., Stephens, P., Kharbouch, A., Shimol, D. B., Burkart, J., ... Myers, L. (2020, aug). Deep learning for automated sleep staging using instantaneous heart rate. *npj Digital Medicine 2020 3:1*, *3*(1), 1–10. Retrieved from https://www.nature.com/articles/s41746-020-0291-x doi: 10.1038/s41746-020-0291-x

Swinnen, L., Chatzichristos, C., Jansen, K., Lagae, L., Depondt, C., Seynaeve, L., ... Van Paesschen, W. (2021, nov). Accurate detection of typical absence seizures in adults and children using a two-channel electroencephalographic wearable behind the ears. *Epilepsia*, *62*(11), 2741–2752. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/epi.17061https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.17061https://onlinelibrary.wiley.com/doi/10.1111/epi.17061 doi: 10.1111/EPI.17061

Tsuno, N., Besset, A., & Ritchie, K. (2005, oct). Sleep and Depression. *The Journal of Clinical Psychiatry*, *66*(10), 19685. Retrieved from https://www.psychiatrist.com/jcp/neurologic/neurology/sleep-depressionhttps://www.psychiatrist.com/jcp/neurologic/neurology/sleep-depression/

van Gorp, H., M Huijben, I. A., Fonseca, P., G van Sloun, R. J., Overeem, S., van Gilst, M. M., & author Hans van Gorp, C. (2022, jun). Certainty about uncertainty in sleep staging: a theoretical framework. *Sleep*, *2022*, 1–8. Retrieved from https://academic.oup.com/sleep/advance-article/doi/10.1093/sleep/zsac134/6604464 doi: 10.1093/SLEEP/ZSAC134

Winer, J. R., Deters, K. D., Kennedy, G., Jin, M., Goldstein-Piekarski, A., Poston, K. L., & Mormino, E. C. (2021, 10). Association of Short and Long Sleep Duration With Amyloid-β Burden and Cognition in Aging. *JAMA Neurology*, *78*(10), 1187-1196. Retrieved from https://doi.org/10.1001/jamaneurol.2021.2876 doi: 10.1001/jamaneurol.2021.2876

Zhang, Y., Ren, R., Yang, L., Zhang, H., Shi, Y., Okhravi, H. R., ... Tang, X. (2022, dec). Sleep in Alzheimer's disease: a systematic review and meta-analysis of polysomnographic findings. *Translational Psychiatry*, *12*(1). doi: 10.1038/s41398-022-01897-y

□

APPENDIX

A EXPERIMENTAL SETUP

A.1 Uncertainty at training and test time

All experiments were performed on the wearable data. The sleep staging model was trained on the full training dataset for 10 training epochs. Early stopping was applied using a validation dataset consisting of 1 to 2 recordings, by retaining the model with the best performance on the validation set. This experiment was repeated 23 times, with different recordings in the validation set for every repeat. Accuracies were computed on the separate test set consisting of 45 recordings. The reported test set accuracies were computed as means $\pm$ standard deviations over these 23 repeats.

A.2 Manual review

We used ten randomly selected test recordings for the manual review experiment. None of these were marked as difficult to score based on the recording-wise uncertainty detection step. The clinician reported that they were unable to interpret two of these recordings because of too high noise levels. As we did not get manual annotations for these recordings, they were left out of the subsequent analysis. For the remaining eight recordings, we compared one model (randomly selected from the 23 trained models) to the scorings of the clinician.

A.3 Estimating sleep parameters

The predicted sleep parameters were calculated for each subject, based on the predicted sleep stage labels. As we trained 23 models, the final sleep parameter for each subject was calculated as the mean over the 23 predicted sleep parameters. The means and standard deviations in Table 2 were calculated over all subjects, except the seven recordings that were flagged for high levels of uncertainty. Pearson's correlation test was performed with the subjects as samples. The correlation coefficients in Table D.1 (in Appendix D) were also calculated between the averaged sleep parameters per subject. The same procedure for estimating the sleep parameters was used in the Bland-Altman plots in Figure C.1 (Appendix C) and for the scatter plots in Figure 4.

B SOURCES OF UNCERTAINTY

The uncertainty in sleep staging emerges from two main sources: biological factors such as disease and age, and measurement technique-related factors like segmentation and interference (van Gorp et al., 2022). We can visualize some of the uncertainty-causing factors in the training data. Figure B.1 shows two clinical parameters, the apnea-hypopnea index (AHI) on the total sleep time, and the sleep efficiency (total sleep time over total time in bed) for all the patients in the training set. The AHI is a measure for the sleep apnea severity. Each patient is one dot, and the color of each dot indicates the average confidence over all the samples of that patient's recording. The confidence is the predicted probability for the correct class. It is inversely correlated with both sleep efficiency and the AHI, showing that these clinical factors impact the uncertainty of the model.

C BLAND-ALTMAN PLOTS

D SLEEP PARAMETERS' CORRELATIONS WITH AHI

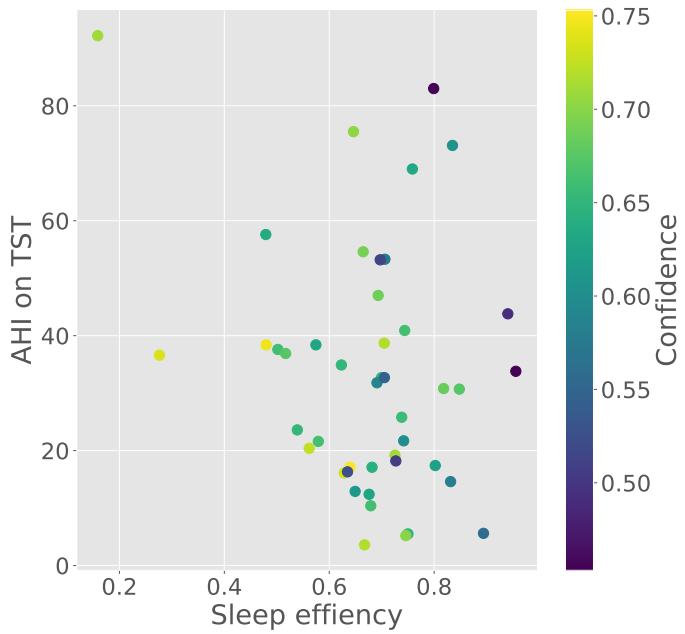


FIGURE B.1 Clinical parameters lie at the source of data uncertainty in the wearable training dataset. Mean confidence for every recording in the training set. The confidence is strongly correlated with sleep efficiency (Pearson's $\rho = -0.52, p < 0.05$). It has a weaker, non-significantly relation to the apnea-hypopnea index (Pearson's $\rho = -0.12, p = 0.45$).

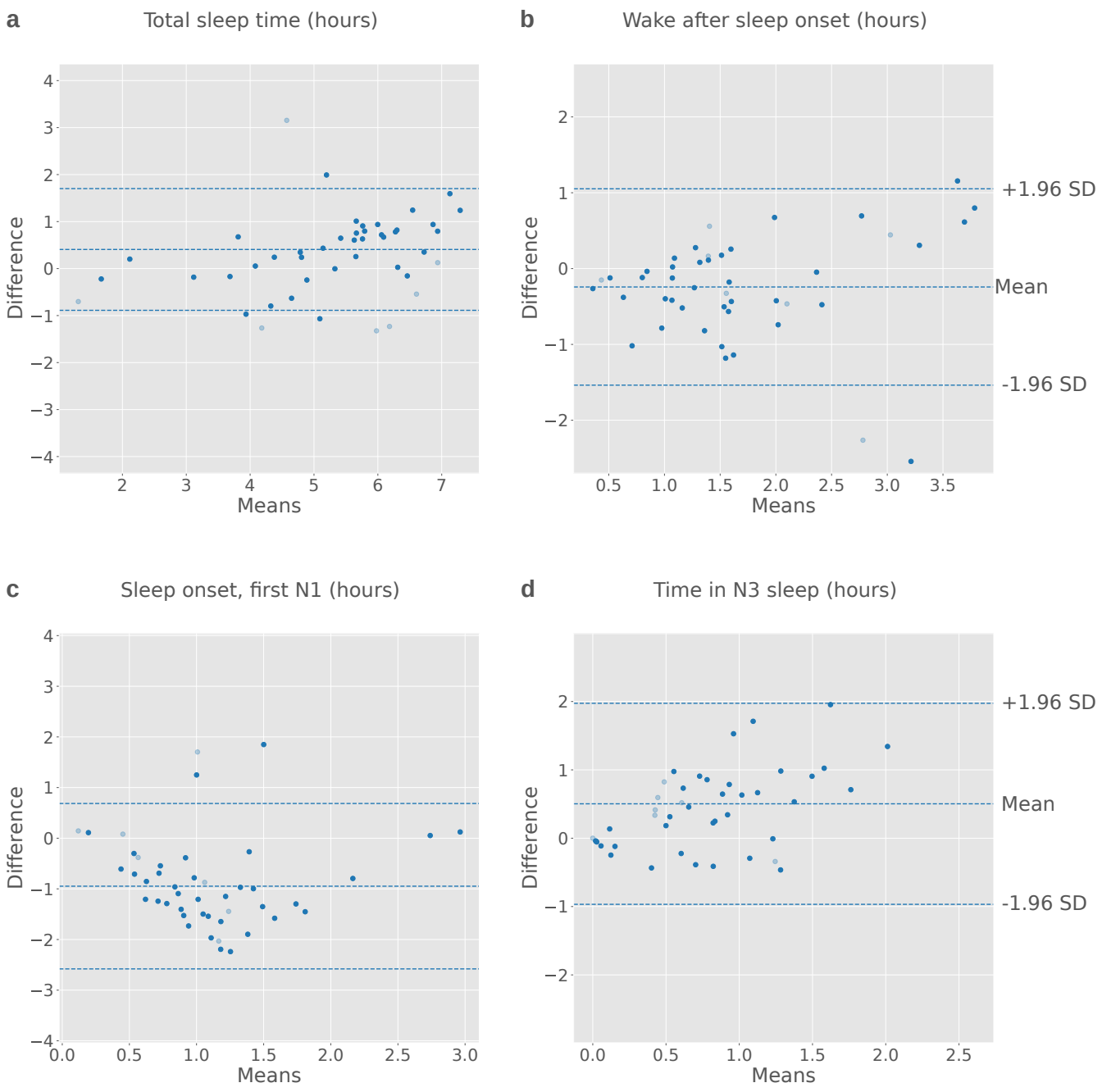


FIGURE C.1 Bland-Altman plots for a number of important sleep parameters, where each sample is one patient. Every sample shows an average over the 23 repeats. Each plot shows the agreement between the parameter calculated based on the PSG reference and model predictions. The difference between both is shown on the y-axis, and the mean of the two measures is plotted on the x-axis. The light-colored dots represent recordings that were discarded because of high levels of uncertainty. These Bland-Altman plots show an overall good agreement between the predicted and reference sleep parameters.

TABLE D.1 A total of 26 sleep parameters calculated based on the hypnograms obtained with our deep neural network and the reference hypnograms, as annotated by a clinician based on PSG. Pearson's correlation coefficient (ρ) is calculated between the sleep parameters and the apnea-hypopnea index on total sleep time (AHI on TST).

	Pearson's ρ[a]				Clinical papers[b]	
	Predicted		Reference		Sign.	Non-sign.
	ρ	p-value	ρ	p-value		
TST (hr)	-0.10	0.55	-0.16	0.33	Bianchi, Basunia: -	Shahveisi
Sleep efficiency (%)	**-0.49**	0.00	**-0.46**	0.00	Basunia, Redline: -	Bianchi, Shahveisi
W (hr)	**0.48**	0.00	**0.51**	0.00		
N1 (hr)	**0.40**	0.01	**0.44**	0.01		
N2 (hr)	0.05	0.75	**-0.32**	0.05		
N3 (hr)	**-0.34**	0.04	-0.14	0.41		
REM (hr)	-0.12	0.46	-0.21	0.20	Bianchi: -	
N1 rel (% TST)	**0.49**	0.00	**0.44**	0.01	Basunia, Shahveisi, Redline: +	Bianchi
N2 rel (% TST)	0.20	0.22	-0.27	0.10	Bianchi, Redline: +, Basunia: -	Shahveisi
N3 rel (% TST)	**-0.33**	0.04	-0.03	0.85	Bianchi, Basunia, Shahveisi, Redline: -	
REM (% TST)	-0.11	0.52	-0.18	0.27	Basunia, Redline: -	Shahveisi
Awakenings	**0.33**	0.04	**0.37**	0.02		
Sleep stage transitions	**0.32**	0.05	**0.37**	0.02		
Transitions to REM	0.00	1.00	0.01	0.93		
Transitions to N3	**-0.32**	0.05	-0.03	0.88		
Time between REM (hr)	0.12	0.51	**0.36**	0.04		
Time between N3 (hr)	-0.14	0.46	-0.07	0.72		
Sleep onset N1 (hr)	0.26	0.12	0.31	0.06		Basunia
Sleep onset N2 (hr)	0.29	0.08	**0.49**	0.00		Basunia
WASO (hr)	**0.41**	0.01	**0.35**	0.03		
Awakenings (/TST)	**0.42**	0.01	**0.43**	0.01	Basunia, Bianchi, Redline *	
Sleep stage transitions (/TST)	**0.44**	0.01	**0.44**	0.01		
Transitions to REM (/TST)	0.04	0.80	0.06	0.70		
Transitions to N3 (/TST)	-0.32	0.05	0.05	0.76		
WASO (/TST)	**0.43**	0.01	**0.34**	0.03		
REM latency (hr)	0.06	0.74	0.18	0.30	Bianchi: +	

[a]Significant correlations ($p \leq 0.05$) are marked in bold.

[b]Studies that found a significant correlation are reported in the 'sign.' column, with a '+' or '-' indicating the type of correlation, and papers that found the correlation was non-significant correlation are reported in the 'non-sign.' column.

References: (Bianchi et al., 2010; Basunia et al., 2016; Redline et al., 2004; Shahveisi et al., 2018)

Abbreviations: WASO: Wake after sleep onset, hr: hours, +: positive correlation, -: negative correlation, *: positive correlation with the arousal index, which is strongly related to but not equal to the amount of wakenings over TST.