Predicting Sleep Quality via Unsupervised Learning of Cardiac Activity

Max Moebus*, Julien Wolfensberger*, and Christian Holz
Department of Computer Science, ETH Zürich, Switzerland
firstname.lastname@inf.ethz.ch

***Abstract*—While highly important for a person's mood, productivity, and physical performance, perceived sleep quality is challenging to model and, thus, predict with passive means such as physiological and behavioral signals alone. In this paper, we propose a novel method that diverges from traditional feature-based modeling of sleep quality. Instead, our method is unsupervised and derives states of cardiac activity from polysomnography (PSG) recordings of more than 6,800 participants. We then demonstrate that the proportion of time spent in these states strongly correlates with perceived sleep quality using a longitudinal study of 16 participants over one month. Our method classifies participants' perceived sleep quality with a balanced accuracy of 68%, significantly exceeding prior methods and feature-based approaches that incorporate established metrics of cardiac activity. Interestingly, we find that the states of cardiac activity our method derives oppose traditional sleep stages—even though the states seem easily explainable based on simple metrics of cardiac activity. Thus, we provide evidence that there are still little-understood processes during sleep that need further investigation, potentially even a rethinking of sleep analysis, especially for perceived sleep quality.**

***Index Terms*—sleep quality, cardiac activity, unsupervised learning, wearable sensor**

I. Introduction

Sleep quality is crucial for overall well-being, impacting productivity, mood, and physical strength [1], [2]. Its significance is even more pronounced in individuals with chronic conditions like multiple sclerosis or Parkinson's disease, influencing recovery, pain management, and fatigue [3]. In the general population, sleep quality is a key indicator of various sleep disorders and medical conditions [4], [5]. Unaddressed sleep disorders greatly heighten the risk of medical and psychiatric illnesses, affecting potentially over 40% of the population [6]. Consequently, understanding sleep quality and detecting sleep disorders early is increasingly vital.

Polysomnography (PSG) is the gold standard for sleep study measurements, typically conducted in specialized labs [7]. PSG involves overnight stays and comprehensive signal recording, analyzed by medical professionals for reliable sleep analysis [8]. However, due to its high costs, researchers have explored alternatives such as wearables for broader and more cost-effective sleep quality assessments [8], [9]. While they provide inferior signal quality and modalities, they can be conducted passively, outside lab environments, and afford more longitudinal studies [7], [10].

* M.M. and J.W. contributed equally

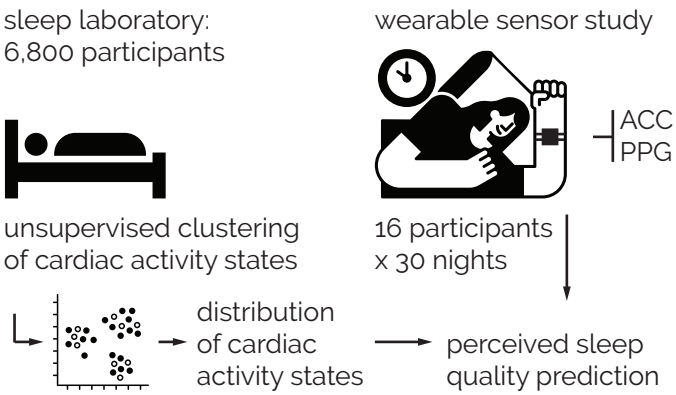

Fig. 1. We make use of clusters of cardiac activity learned in an unsupervised manner from PSG recordings to predict perceived sleep quality during an intensive longitudinal study with a wrist-worn wearable sensor.

Using a wrist-worn accelerometer, actigraphy devices have emerged as reliable sleep-wake classifiers [11], [12]. Classifying sleep stages on such devices is difficult and previous work has often used multimodal sensor designs (e.g., photoplethysmography and accelerometer [13]). Though equally desirable, few wearable efforts have been capable of predicting perceived sleep quality [4]. A recent study on skin temperature and electrodermal activity on the wrist in addition to wrist movement to predict binary perceived sleep quality [4] achieved an accuracy of 57%—too low for practical purposes.

In this paper, we present a novel method for perceived sleep quality estimation. We derive states of cardiac activity from a large PSG study with more than 6,800 participants [14] using unsupervised learning. During an intensive longitudinal study including a wearable sensor [4], we used the derived states to predict daily perceived sleep quality for 16 participants across one month. Our findings highlight that some of the processes while sleeping are still little-understood and not captured by commonly used metrics of cardiac activity—or sleep stages. Using the time spent in the calculated states of cardiac activity, we achieve a balanced accuracy of 68%—significantly outperforming existing approaches [4] and classical feature-based classification.

II. Datasets and Methods

A. Datasets

We make use of the 'Multi-Ethnic Study of Atherosclerosis' (MESA) and the M2Sleep dataset [4], [14]. In the MESA

dataset, sleep stages and cardiac activity were recorded for 6,814 participants during a one-night stay in a sleep laboratory. In the M2Sleep dataset, 16 participants (5 female, 11 male), ages 19–35 years, wore a wearable sensor (Empatica E4) for one month and recorded their perceived sleep quality each morning. Participants also provided their rough sleep and wake times. We use both datasets to derive states of cardiac activity in an unsupervised manner to predict perceived sleep quality on the M2Sleep dataset.

Ultimately, we model perceived sleep quality based on various combinations of computed states of cardiac activity, well-established heart rate variability (HRV) metrics, actigraphy data, and participants' sleep duration. On the M2Sleep dataset, we use the output of the Empatica E4 to derive the above features. Worn at the wrist, the Empatica E4 supplies signals of blood volume pulse (BVP) at 64 Hz, and inertial motion at 32 Hz. Even though the MESA dataset also supplies electrocardiogram (ECG) signals to measure cardiac activity, we used a blood volume pulse signal recorded at the finger to match the setting of the M2Sleep dataset as closely as possible.

B. Signal aggregation

To assign accurate moments of sleep to the M2Sleep dataset, we fused the estimations of widely verified methods (Sadeh et al. [11], Cole-Kripke [12]) on inertial motion signals over 2-hour intervals centered on the sleep and wake times reported by participants. Similar to [4], we derived activity counts to quantify the amount of wrist movement while asleep [13].

We obtain inter-beat intervals from the BVP signals in both, the MESA and M2Sleep dataset in non-overlapping windows of 30 seconds. For the calculation of HRV metrics, we examine the 10-minute intervals centered on each 30-second window. In particular, we first apply a Chebychev Type 2 4^{th} order bandpass filter with cutoff frequencies of 0.5 and 10 Hz to each of the 10 minute intervals (as recommended for short BVP signals [15]). We then obtain heartbeats from HeartPy's analysis [16]. Based on the inter-beat intervals across the 10 minutes, we calculate participants' heart rates (HR) at 1 Hz and various HRV metrics: SD1, SD2, SDNN, LF, HF, LF/HF, VLF, mean NNI, NNI_{20}, NNI_{50}, pNNI_{20}, and pNNI_{50} [17]. During the initial 30-second window, we calculate minimum, mean, and maximum HR and the standard deviation of HR. To each 30-second window, we thus attribute summary statistics of HR during the 30 seconds, a 10-minute HR signal, and various HRV metrics (based on 10 minutes). We normalize all HRV metrics and the summary statistics of the HR per participant by subtracting each participant's average value per metric and dividing by the respective standard deviation [18].

C. Clustering algorithms for cardiac activity

We compare three unsupervised clustering techniques that each assign 30-second windows of data to one of C clusters. Based on each window's average HR and its associated HRV metrics normalized per participant, we use 'K-Means' and a 'Gaussian Mixture Model' (GMM) [19]. For each 30-second window, we further use 'GEMINI' [20] based on the attributed 10-minute HR signal. We integrate a 1D convolutional neural network (LeNet-5 [21]) into the GEMINI framework to extract features from the 10-minute HR signal. Input to all clustering algorithms is C—the number of intended clusters. Each 30-second window is thus assigned to one of the C clusters by each of the three clustering algorithms. Based on the MESA and the M2Sleep dataset, we derive C clusters of cardiac activity while participants were asleep, which we will refer to as *states of cardiac activity*. From M2Sleep, we also derive C clusters of cardiac activity while participants were awake.

TABLE I
CLASSIFICATION PERFORMANCE FOR PERCEIVED SLEEP QUALITY.

Number of clusters C	Features	BA	F1	Cohen's κ	Classifier
0	✓	58%	0.58	0.15	KNN
3	×	63%	0.63	0.25	KNN
3	✓	64%	0.65	0.28	KNN
5	×	58%	0.70	0.18	KNN
5	✓	62%	0.65	0.24	NN
7	×	62%	0.72	0.24	SVM
7	✓	64%	0.66	0.27	NN
10	×	**68%**	**0.76**	**0.36**	**SVM**
10	✓	67%	0.67	0.24	SVM
15	×	62%	0.67	0.24	NN
15	✓	63%	0.66	0.26	NN

D. Models for Perceived Sleep Quality

We model binary normalized perceived sleep quality on the M2Sleep dataset given the high intra- as well as interindividual variability of subjective phenomena such as sleep quality [18]. That is, we predict whether a perceived sleep quality label lies above or below each participant's mean sleep quality label, which we will simply refer to as perceived sleep quality. We compare 5 different classifiers: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Logistic Regression (LR), and a small Multi-Layer Perceptron (NN) [19]. We evaluate each classifier using leave-one-participant-out cross-validation. At each split, we tune the hyperparameters of each algorithm on all participants apart from the left-out subject based on 3-fold cross-validation.

Input to these models is either the distribution of time attributed to each of the C states of cardiac activity throughout the night, or hand-crafted features, or both. The hand-crafted features include participants' sleep duration, summary statistics about activity counts derived from the wrist's motion during sleep, summary statistics of the HR during sleep, and well-established HRV features derived while asleep. Similar to the perceived sleep quality labels, we normalize all model inputs per participant given the high variability of HR metrics between two individuals (cf. [18]).

III. RESULTS

We investigate the performance at binary normalized perceived sleep quality classification in Table I. We compare modeling perceived sleep quality only based on hand-crafted features (sleep duration, activity counts, and cardiac activity),

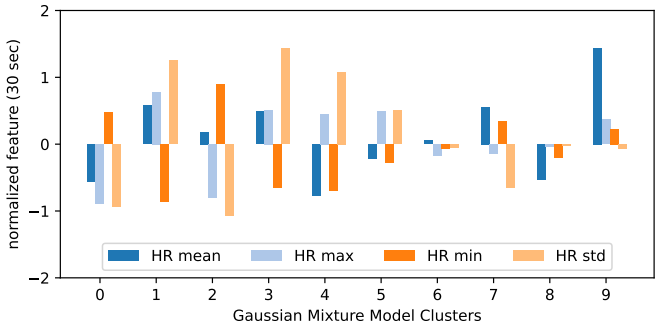

Fig. 2. State visualization using normalized heart rate (HR) features.

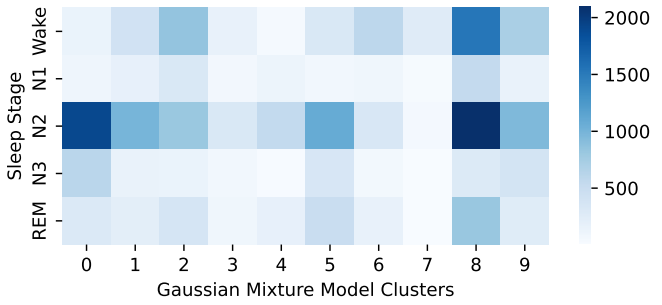

Fig. 3. Comparison of sleep stages and derived states of cardiac activity. The color corresponds to the number of 30-second windows (100 participants).

to constructing 3, 5, 7, 10, or 15 clusters (C) for participants' cardiac activity while asleep. For each value of C we assess whether including the hand-crafted features is of value and report the performance of the best-performing classifier (out of RF, LR, SVM, KNN, and NN) in the classifier column.

We find that dividing cardiac activity during the night into 10 states results in the highest model performance: SVM with a balanced accuracy (BA) of 68% (0.76 F1, 0.36 Cohen's κ) when excluding hand-crafted features. In terms of balanced accuracy, this is closely followed by deriving 10 states of cardiac activity and including the hand-crafted features as input to a SVM classifier. Modeling perceived sleep quality based on the time spent in 7 clusters of cardiac activity alone results in the next best F1 score of 0.72 (62% BA and 0.24 κ). The next highest κ is achieved when we predict perceived sleep quality based on 7 clusters of cardiac activity and hand-crafted features (0.27 κ, 64% BA, 0.66 F1). Modeling perceived sleep quality only based on the hand-crafted features leads to a balanced accuracy of 58% (0.58 F1, and 0.15 κ).

A. Cardiac activity states from the Gaussian mixture model

In the following, we describe the states of cardiac activity derived using the Gaussian mixture model (GMM) using simple HR summary statistics and highlight their relation to traditional sleep stages and perceived sleep quality. Representatively, we only investigate the states of cardiac activity derived using the GMM, as they show very similar characteristics to the states derived using GEMINI and K-Means.

1) Description of cardiac activity states: Figure 2 visualizes the 10 clusters derived using a GMM in terms of easily interpretable HR features. The 10 clusters show a broad range of patterns in terms of the per-participant normalized standard deviation of the HR and the maximum, mean, and minimum HR during a 30-second window. While 30-second windows that fall into cluster 9 have on average a high mean HR and around average minimum, and maximum HR, cluster 4 is characteristic for 30-second windows with very low average HR but very high variation. Cluster 1 and 3 are both dominated by 30-second windows with highly fluctuating HR, but differ in terms of average HR.

2) Cardiac activity states and traditional sleep stages: Figure 3 shows how the clusters derived by the GMM relate to professionally scored sleep stages for 100 randomly chosen participants of the MESA dataset. With 10 states ($C = 10$) we use the best-performing model of Table I. We find that the derived states align only partially with specific sleep stages. State 3 and 7, for instance, seem dominated by participants being awake. State 0 and 5, on the other hand, are mainly comprised of N2 and N3 (deep sleep) with some 30-second windows also falling in the REM phase. Most states do not show a clear trend. Similarly, no sleep stage is only attributed to a single state of cardiac activity.

3) Cardiac activity states and perceived sleep quality: Figure 4 shows how the average time spent in the cardiac states derived using a GMM relates to binary normalized perceived sleep quality on the M2Sleep dataset. We find that participants who report high perceived sleep quality tend to have 5% more epochs attributed to cardiac state 0 than participants who report low perceived sleep quality: the largest difference of any state. As shown in Figure 2, cardiac state 0 comprises 30-second windows with relatively low HR that fluctuate only little. Mostly, these windows correspond to the N2 sleep stage but also to N3 (deep sleep) and REM sleep.

IV. Discussion

Our results highlight that states of cardiac activity derived using unsupervised clustering techniques predict perceived sleep quality significantly better than traditional hand-crafted

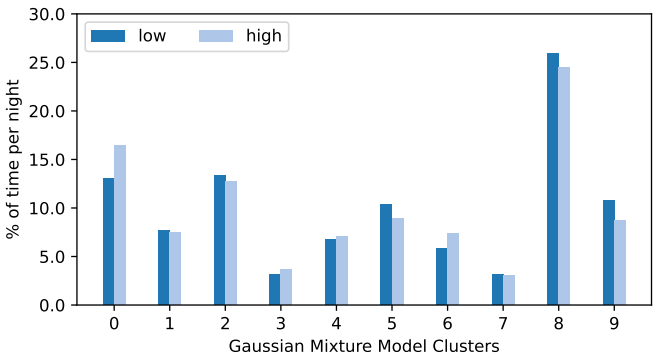

Fig. 4. Comparison of the % of time spent in each state of cardiac activity for low & high binary normalized perceived sleep quality.

features of cardiac activity. The states of cardiac activity we derived do not link well to traditional sleep stages, however, are easily interpretable and distinguishable even using simple HR metrics. Our work thus clearly highlights that some of the processes influencing (perceived) sleep quality are still little understood. This poses a new avenue for sleep research.

Despite no clear connection between traditional sleep stages and the states of cardiac activity we derived, there is evidence of an interplay. State 0 in Figure 2 showed the greatest difference in distribution between high- and low-quality nights. In Figure 3, we see that it comprises mostly N2 sleep, but also N3 (deep sleep) and REM sleep—hinting that the interplay of the three sleep stages might be of importance. Cycling through the different sleep stages is—per definition—crucial for objective sleep quality [22]. Sufficient REM and N3 (deep sleep) have also been linked to subjective sleep quality, however [22]. This interplay of the different sleep stages captured through unsupervised clustering of cardiac activity offers a potential intuition as to why the derived states are more explanatory of perceived sleep quality than hand-craft features. Generally, this highlights that some of the dynamics that are visible in cardiac activity while asleep and influence perceived sleep quality extend past the traditional definition of sleep stages and our current means of sleep analysis.

Our proposed methodology merges the advantages of large in-lab PSG studies with smaller in-the-wild wearable sensor studies to identify hidden states of cardiac activity. However, while the MESA dataset is a very large and representative in-lab PSG study, the universality of our findings and the success of our methodology for a broader population remain unclear given the relatively small size of the M2Sleep dataset. Larger wearable sensor datasets that incorporate information about perceived sleep quality are needed to confirm the suitability of our approach.

V. Conclusion

Our findings in this paper demonstrate that there are states of cardiac activity during sleep that predict perceived sleep quality better than traditional features of cardiac activity (i.e., HR and HRV) often used in related analyses. While these states are easily distinguishable using only simple heart rate summary statistics, they do not relate well to the traditional sleep stages. However, we demonstrate that different states of cardiac activity might capture relationships between different sleep stages (N2, N3, and REM) that were found important for sleep quality by related works and might explain some of our findings. Overall, we highlight that we need to incorporate new methods of analysis into sleep research to improve our understanding of (perceived) sleep quality.

VI. Data availability

The M2Sleep dataset is publicly available [4]. The MESA dataset is available upon approval from the 'National Sleep Research Resource' (NSRR) [23].

References

[1] K. Klier, S. Dörr, and A. Schmidt, "High sleep quality can increase the performance of crossfit® athletes in highly technical-and cognitive-demanding categories," *BMC Sports Sci Med Rehabil*, 2021.

[2] D. M. van Dijk, W. van Rhenen, J. M. Murre, and E. Verwijk, "Cognitive functioning, sleep quality, and work performance in non-clinical burnout: the role of working memory," *PloS one*, 2020.

[3] S. Brass, P. Duquette, J. Proulx-Therrien, and S. Auerbach, "Sleep disorders in patients with multiple sclerosis," *Sleep Med Rev*, 2010.

[4] S. Gashi, L. Alecci, E. Di Lascio, M. E. Debus, F. Gasparini, and S. Santini, "The role of model personalization for sleep stage and sleep quality recognition using wearables," *IEEE Pervasive Computing*, 2022.

[5] H. Urponen, M. Partinen, I. Vuori, and J. Hasan, "Sleep quality and health: Description of the sleep quality index," *Sleep and health risk*, 1991.

[6] Y. Jung, M. R. Junna, J. N. Mandrekar, and T. I. Morgenthaler, "The national healthy sleep awareness project sleep health surveillance questionnaire as an obstructive sleep apnea surveillance tool," *Journal of Clinical Sleep Medicine*, 2017.

[7] V. Ibáñez, J. Silva, E. Navarro, and O. Cauli, "Sleep assessment devices: types, market analysis, and a critical view on accuracy and validation," *Expert review of medical devices*, 2019.

[8] A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel, and G. Clifford, "A review of signals used in sleep analysis," *Physiological measurement*, 2013.

[9] R. D. Kim, V. K. Kapur, J. Redline-Bruch, M. Rueschman, D. H. Auckley, R. M. Benca, N. R. Foldvary-Schafer, C. Iber, P. C. Zee, C. L. Rosen *et al.*, "An economic evaluation of home versus laboratory-based diagnosis of obstructive sleep apnea," *Sleep*, 2015.

[10] A. Carek and C. Holz, "Naptics: convenient and continuous blood pressure monitoring during sleep," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018.

[11] A. Sadeh, M. Sharkey, and M. A. Carskadon, "Activity-based sleep-wake identification: an empirical test of methodological issues," *Sleep*, 1994.

[12] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, 1992.

[13] O. Walch, Y. Huang, D. Forger, and C. Goldstein, "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device," *Sleep*, 2019.

[14] X. Chen, R. Wang, P. Zee, P. Lutsey, S. Javaheri, C. Alcántara, C. Jackson, M. Williams, and S. Redline, "Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis," *Sleep*, 2015.

[15] Y. Liang, M. Elgendi, Z. Chen, and R. Ward, "An optimal filter for short photoplethysmogram signals," *Scientific data*, 2018.

[16] P. van Gent, H. Farah, N. van Nes, and B. van Arem, "Analysing noisy driver physiology real-time using off-the-shelf sensors: Heart rate analysis software from the taking the fast lane project," *Journal of Open Research Software*, 2019.

[17] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, 2017.

[18] M. Moebus, S. Gashi, M. Hilty, P. Oldrati, and C. Holz, "Meaningful digital biomarkers derived from wearable sensor to predict daily fatigue in multiple sclerosis patients and healthy controls," *iScience*, 2024.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 2011.

[20] L. Ohl, P.-A. Mattei, C. Bouveyron, W. Harchaoui, M. Leclercq, A. Droit, and F. Precioso, "Generalised mutual information for discriminative clustering," *Advances in Neural Information Processing Systems*, 2022.

[21] L. Wan, Y. Chen, H. Li, and C. Li, "Rolling-element bearing fault diagnosis using improved lenet-5 network," *Sensors*, 2020.

[22] S. J. McCarter, P. T. Hagen, E. K. S. Louis, T. M. Rieck, C. R. Haider, D. R. Holmes, and T. I. Morgenthaler, "Physiological markers of sleep quality: a scoping review," *Sleep Med Rev*, 2022.

[23] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The national sleep research resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, 2018.