dHealth 2024
D. Hayn et al. (Eds.)

© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI240041

Neural Network-Based Prediction of Perceived Sleep Quality Through Wearable Device Data

Martin BAUMGARTNER[a,b,1], Manuel GRÖSSL[c], Raphaela HAUMER[c], Katharina POIMER[c], Flora PRANTL[c], Katharina WEICK[c], Markus FALGENHAUER[a], Stefan BEYER[a], Andreas ZIEGL[d], Aaron LAUSCHENSKY[a], Fabian WIESMÜLLER[a,b,e], Karl KREINER[a], Dieter HAYN[a,c] and Günter SCHREIER[a,b]

[a] *AIT Austrian Institute of Technology, Graz & Vienna, Austria*
[b] *Institute of Neural Engineering, Graz University of Technology, Graz, Austria*
[c] *University of Applied Sciences Wiener Neustadt, Wiener Neustadt, Austria*
[d] *telbiomed Medizintechnik und IT Service GmbH, Graz, Austria*
[e] *Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*

Abstract. Background: This study focuses on the development of a neural network model to predict perceived sleep quality using data from wearable devices. We collected various physiological metrics from 18 participants over four weeks, including heart rate, physical activity, and both device-measured and self-reported sleep quality. Objectives: The primary objective was to correlate wearable device data with subjective sleep quality perceptions. Methods: Our approach used data processing, feature engineering, and optimizing a Multi-Layer Perceptron classifier. Results: Despite comprehensive data analysis and model experimentation, the predictive accuracy for perceived sleep quality was moderate (59%), highlighting the complexities in accurately quantifying subjective sleep experiences through wearable data. Applying a tolerance of 1 grade (on a scale from 1-5), increased accuracy to 92%. Discussion: More in-depth analysis is required to fully comprehend how wearables and artificial intelligence might assist in understanding sleep behavior.

Keywords. sleep quality, wearable electronic devices, neural network, multi-layer perceptrons, electronic data processing

1. Introduction

Wearable devices represent a significant advancement in personal health monitoring, offering a convenient way to track various physiological parameters. These devices, which range from smartwatches to fitness trackers, are designed to measure and record health-related data such as heart rate, steps taken, calories burned, and even sleep patterns. In particular, the ability of these devices to assess sleep quality (SQ) is promising,

[1] Corresponding Author: Martin Baumgartner, AIT Austrian Institute of Technology, Vienna, Austria, E-Mail: martin.baumgartner@ait.ac.at

considering the vital importance of sleep in maintaining overall health and well-being [1] .

SQ, an essential determinant of physical and mental health, is influenced by various factors, including duration, continuity, and the composition of sleep stages [2]. Sleep is typically divided into three stages: light sleep, deep sleep, and rapid eye movement (REM) sleep. Each stage plays a unique role in the sleep cycle, contributing to the restorative processes of the body and mind [3]. The direct relationship between physiological measures and SQ is challenging to investigate since no definitive markers for objective SQ have been formulated. SQ is determined by a number of objective (e.g. duration) and subjective (e.g. restfulness) criteria [4].

Assessing SQ currently relies on polysomnography, a comprehensive sleep study conducted in specialized clinics with application of a multitude of measurement devices. This gold standard method includes monitoring brain waves via electroencephalogram, blood oxygen levels, heart rate, breathing, and eye and leg movements [2]. The application of this extensive use of equipment is often disturbing the SQ [5]. Further, sleeping in unfamiliar environments (i.e. sleep clinic) is associated with disturbed sleep especially in the first night (e.g. first-night effect [6]). However, wearable devices offer a more accessible and less intrusive way to track sleep patterns in comfortable environment, though they may not capture data with the same level of detail as polysomnography [7]. A study from 2019 demonstrated that – compared to polysomnography – wearable devices (Fitbit Alta HR and Philips Respironics Actiwatch 2) underestimated total sleep time by up to 47 minutes (Fitbit) and 38 minutes (AW2) and overestimated wake after sleep onset by up to 42 minutes. The sensitivity for detecting sleep was over 90% for both devices. Despite certain limitations in measurement accuracy, wearable devices have shown to be a viable alternative to polysomnography [8].

This leads to the research question, which vital parameters are ultimately needed to assess SQ. The purpose of this study was to explore the relationship between various factors affecting SQ, particularly focusing on the correlation between physiological measures captured by wearable devices and the subjective experience of SQ. By exploring these dynamics, this paper aspires to meaningfully contribute to the growing field of personal health monitoring, focused on its impact on sleep-related health results.

2. Methods

The approach encompassed data collection from a diverse group of participants, data processing procedures, and the early development of a neural network model to predict perceived SQ.

2.1. Subject recruitment & data collection

A total of 18 participants were recruited for the study, evenly distributed between the sexes (nine men and nine women) and ranging in age from 22 to 41 years. These participants consented to wear a variety of smartwatches, including models from Polar, FitBit, and Apple Watch, specifically for sleep tracking. The data collection phase spanned four weeks during the month of October 2023.

Participants were granted access to a dedicated online platform for data management. This platform enabled participants to record, visualize and review their sleep-related data and observe trends. Notably, in addition to manual data input at this platform, data from the smartwatches could be transmitted automatically through a dedicated mobile application that synchronized various accounts (e.g., FitBit, Polar). This flexibility allowed participants to record data with minimal effort but allowed for retrospective corrections.

Each participant recorded six data metrics daily: *Resting Heart Rate*, *Step Count*, *Physical Activity*, *Well-being*, *Device-measured SQ*, and *Perceived SQ*. Further granularity was provided for the *Physical Activity* metric by including details regarding the type of activity, a self-reported estimate of its intensity, and its duration in minutes. Well-being was evaluated through self-reported assessments, allowing participants to categorize their daily state of general well-being as "*good"*, "*average"* or "*poor". Device-measured SQ* was recorded as the devices' manufacturers provided it (typically a sleep score ranging between 0 and 100). Additionally, the *Perceived SQ* was collected using a grading system where 1 represents "*excellent sleep"* and 5 represents "*poor sleep"* according to subjective feeling of sleep restfulness.

2.2. Data processing

After the data collection phase, the acquired data was exported from the platform, imported into a Python programming environment, and processed based on data type. Preliminary statistical analyses were conducted to gain insights into the data and guided the subsequent steps of choice of model type and training strategy. The following steps were taken to develop a predictive model to estimate SQ:

Data preparation and exploration: We loaded the dataset from a CSV file and performed initial data cleaning by removing non-relevant or non-numeric columns. Descriptive statistics were computed to gain insights into the data's characteristics. Additionally, we visualized feature distributions and the distribution of the target variable (*Perceived SQ)*. The dataset occasionally included duplicated or multiple values (e.g. due to input or transmission errors). We applied different deduplication strategies according to data type. If multiple entries for heart rate or well-being data existed, the mean value was used. In case of multiple step counts, the sum was calculated if the counts were different (duplicated, identical counts of steps per day were considered erroneous and thus removed). If multiple physical activities were recorded, durations were summed up, the type was categorized as "*mixed*", and the highest reported intensity was chosen. If multiple recordings of SQ were present, the last one was selected (as it indicated a corrective input). If no value was available for a specific parameter, it was set to zero. All data was summarized on a daily level (referred to as "daily reports" from here on). These daily reports were included only if *Device-measured SQ* was non-zero and *Perceived SQ* fell within the range of 1 to 5.

Correlation analysis: A correlation matrix was computed to assess the relationships between different features in the dataset. This analysis helped identify feature dependencies and potential multicollinearity.

Data splitting: The data was split into training and testing sets, allocating 20% of the data for testing to ensure a robust evaluation of the model's performance. Standard scaling was applied to each feature to ensure consistent feature scaling.

Model experiments: To optimize our model, we conducted a series of experiments using a Multi-Layer Perceptron (MLP) classifier. We explored various aspects, including

different architectures, activation functions, learning rates, maximum iterations, and a range of alpha values (0.0001, 0.001, 0.01) to fine-tune regularization. Additionally, we experimented with feature transformation, data augmentation ("*digital twins*" with small random variation within 1 standard deviation of feature value), and ensemble methods to enhance model performance.

Feature engineering: In our feature engineering process, we introduced specific types of new features. We developed combined features like "*steps_to_hr_ratio*", where we merged steps and heart rate to understand the interplay between physical activity and cardiovascular response. Additionally, we employed aggregated features such as rolling averages over periods of 3 or 7 days to discern long-term trends from daily fluctuations. Furthermore, we explored feature interactions, exemplified by "*activity_product*", which multiplies activity duration and intensity to provide a holistic view of physical activity.

3. Results

3.1. Descriptive statistics and visualizations

Comprehensive descriptive statistics and visual analyses of data on measured and perceived SQ were conducted. Distribution of features and target variable (*Perceived SQ*) was depicted using histograms and graphical representations, followed by a detailed correlation matrix to visualize relationships among the different variables. As the original ascending scale of both *Perceived SQ* and *Well-being* naturally leads to negative correlation values with *Device-measured SQ*, they were inverted (e.g. 1 equals "*poor sleep*") to show positive correlation, which is more intuitive.

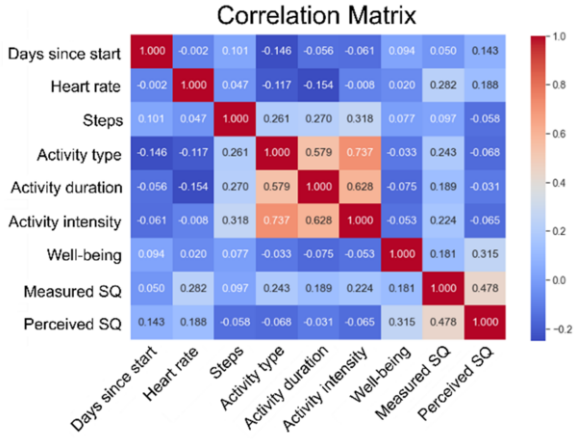

Figure 1. This matrix represents the Spearman correlations between nine variables. Perceived SQ and well-being values were inverted to display more intuitive positive correlation values.

Figure 1 shows the correlations between nine variables, of which the first six served as features for the model: days since start, heart rate, steps, activity type, activity duration, activity intensity, well-being, device-measured SQ and perceived SQ. In our case, the *Device-measured SQ* and *Perceived SQ* did not show strong correlations ($r > 0.7$) with any of the other values. *Device-measured* and *Perceived SQ* showed moderate correlation at 0.478. The *Device-measured SQ* exhibited a maximum positive correlation

with *Heart Rate* at 0.282. The *Perceived SQ* showed the highest positive correlation with *Well-being* at 0.315.

3.2. Modeling approach

Our modeling approaches encompassed a variety of experimental configurations and methods. We tested various architectures of MLPs, with an architecture of three layers (100, 50, and 25 neurons in respective layers) – achieving the highest accuracy (60%), while others like (150, 100, 50) and (200, 150, 100) reached 51% and 54%, respectively. Various activation functions (e.g., ReLU, tanh, logistic) were tested, with the logistic function, performing the best with an accuracy of 59%.

Alpha values and learning rates were tuned to control overfitting and enhance learning speed. Maximum iteration variations showcased sensitivity, where 500, 1000, and 1500 iterations achieved accuracies of 47%, 47%, and 44%, respectively. Feature transformation through PolynomialFeatures (degree 2) led to a reduced accuracy of 40%. Data augmentation with noise addition resulted in an accuracy of 46%. Ensemble methods, including VotingClassifier, achieved an accuracy of 46%. Hyperparameter tuning, model stacking, and ensemble approaches with RandomForest, GradientBoosting, SVM, and MLP were explored, with the VotingClassifier ensemble reaching the highest accuracy of 53%. Comprehensive feature engineering, encompassing additional, combined, and interaction features, along with RFECV for feature selection, contributed to a thorough evaluation. Visualization through bar charts provided a comprehensive overview of experiment and feature test accuracies, offering valuable insights into the model and feature dynamics. Furthermore, ensemble methods, including Stacking and Voting Classifiers, were employed to enhance prediction accuracy. An ensemble approach, combining MLP and RandomForest achieved an accuracy of 53%. For a more precise model configuration, hyperparameter tuning using scikit-learn modules GridSearchCV and RandomizedSearchCV was performed, integrating the best MLP model into ensemble methods.

As an additional experiment, we evaluated the best performing model with a tolerance of ±1 grade of perceived SQ. This increased accuracy significantly from 59% to 92%, which is visualized in Figure 2.

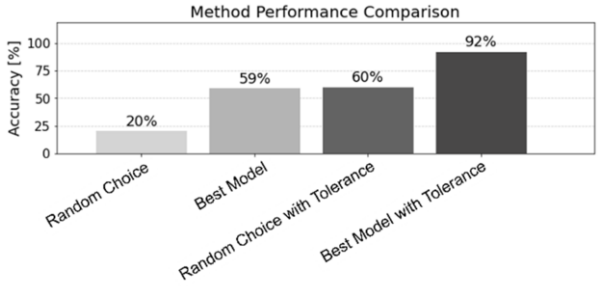

Figure 2. Comparison of different classification methods. Allowing a tolerance of ±1 grade increases accuracy significantly. Random choice is a simulated base line performance (1 out of 5 random selection).

4. Discussion

The purpose of this study was to find out 1) whether SQ is influenced by one or more specific vital parameters or activities and 2) whether the measured SQ corresponds to the perceived SQ. As the presented results clearly show, no strong correlation between measured data and SQ was identified. It is also open to question if more accurate recording devices would have resulted in a stronger correlation. Furthermore, for better comparability, limiting the subjects to a single device type would have stratified the measured SQ. Although the resulting sleep scores were given in already normalized 0-100% in all devices, the underlying algorithm for calculating this score likely differs between manufacturers and thus express different aspects of SQ. Evaluation of the engineered features did not significantly enhance prediction of SQ, they provided valuable insights into the complex nature of the data, contributing to our understanding and formulating future model refinements. These efforts aimed to capture additional information potentially relevant for predicting SQ. While accuracy scores between 40-60% percent were superficially underwhelming, allowing for a 1 grade tolerance in evaluation increased the accuracy dramatically to over 90% (see Figure 2), meaning that while the model did miss fine nuances of sleep, it also did not completely misjudge trends of perceived SQ.

Sleep is not only influenced by the factors already mentioned. Therefore, it is suggested to consider other parameters that influence SQ in future studies. For example, the timing of the measurement of SQ can have a significant impact on the results. Sleep patterns vary during different stages of sleep, making the assessment of SQ dependent on when it is measured. It therefore makes a difference whether one wakes up in a REM sleep phase or in a light sleep phase, for example. According to van Hasselt et al., waking up in the REM sleep phase might report to feel worse than subjects waking up in light sleep phases, although the *Device-measured SQ* is identical in both situations [9]. Furthermore, daylight hours are shorter during the winter which can affect the circadian rhythm and melatonin production. As a result, time of year during measurements has an impact on SQ [10]. This fact could be relevant for our analyses because the data was recorded in October and the nights in this month are longer than the days. Other parameters also vary over the course of the day. For example, during illnesses, perceived well-being can be lower at later hours of the day. The participants of this analysis recorded their well-being during earlier hours of the day and time of data collection was not controlled between subjects. This potentially led to suboptimal quality of physiological data like the resting heart rate that changes during the day. Illness can further affect the quality of sleep as pain, discomfort or medication can irritate sleep. For this reason, it is important to consider the health status of the participants when assessing SQ [11]. For future experiments, a more detailed self-assessed status of both physical as well as mental well-being could be helpful.

Sports watches are increasingly frequently used to measure SQ. However, their reliability and accuracy are debatable. Especially watches designed for wellness or sport applications are not certified as medical devices and thus their reliability is not guaranteed. Our findings of low correlation between measured and perceived SQ also support the assumption that their capabilities to accurately measure sleep are limited. Therefore, more research could be helpful to understand the capabilities of different technologies [12].

In conclusion, the analysis of SQ requires the consideration of various influencing factors over a longer period on more participants. In addition, the choice of equipment must carefully be considered to obtain accurate results.

References

[1] S. M. A. Iqbal et al., Advances in healthcare wearable devices, *NPJ Flexible Electronics* **5**(1) (2021), 9

[2] A. D. Krystal and J. D. Edinger, Measuring sleep quality, *Sleep Medicine* **9** (2008), S10–S17

[3] A. S. Tubbs et al., The basics of sleep physiology and behavior, in *Sleep and health*, Elsevier, (2019), 3–10.

[4] S. J. McCarter et al., Physiological markers of sleep quality: A scoping review, *Sleep Medicine Reviews* **64** (2022), 101657

[5] T. Blackwell et al., A novel approach using actigraphy to quantify the level of disruption of sleep by in-home polysomnography: the MrOS Sleep Study: Sleep disruption by polysomnography, *Sleep Medicine* **32** (2017), 97–104

[6] L. Ding et al., A meta-analysis of the first-night effect in healthy individuals for the full age spectrum, *Sleep Medicine* **89** (2022), 159–165

[7] A. J. Boe et al., Automating sleep stage classification using wireless, wearable sensors, *NPJ digital medicine* **2**(1) (2019), 131

[8] X. K. Lee et al., Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations, *Journal of Clinical Sleep Medicine* **15**(9) (2019), 1337–1346

[9] L. W. A. Hermans et al., Representations of temporal sleep dynamics: Review and synthesis of the literature, *Sleep Medicine Reviews* **63** (2022), 101611

[10] S. J. van Hasselt et al., Seasonal variation in sleep time: jackdaws sleep when it is dark, but do they really need it?, *Journal of Comparative Physiology B* (2023), 1–11

[11] M. Kemple et al., Sleep quality in patients with chronic illness, *Journal of clinical nursing* **25**(21–22) (2016), 3363–3372

[12] E. Guillodo et al., Clinical applications of mobile health wearable–based sleep monitoring: systematic review, *JMIR mHealth and uHealth* **8**(4) (2020), e10733