Sleep Stage Estimation by Introduction of Sleep Domain Knowledge to AI: Towards Personalized Sleep Counseling System with GenAI

Iko Nakari and Keiki Takadama

The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, Japan 182-8585
iko0528@cas.lab.uec.ac.jp, keiki@inf.uec.ac.jp

Abstract

As a first step towards realizing an AI sleep counselor capable of generating personalized advice, this paper proposes a method for monitoring daily sleep conditions with a mattress sensor. To improve the accuracy of sleep stage estimation and to get accurate sleep structure, this paper introduced sleep domain knowledge to machine learning for improving the accuracy of sleep stage estimation. Concretely, the proposed method estimates ultradian rhythm based on the body movement density, updates prediction probabilities of each sleep stage by ML model and applies WAKE/NR3 detection based on the large/small body movement. Through the human subject experiment, the following implications have been revealed: (1) the proposed method improved the percentage of Accuracy by 65.0% from 61.5% and the QWK score by 0.196 from 0.297 by the conventional machine learning method; (2) the proposed method prevents over-NR12 estimating and is useful for understanding sleep structure by estimating REM sleep and NR3 sleep correctly. (3) the correct estimation of ultradian rhythms significantly improved the sleep stage estimation, with an Accuracy of 77.6% and a QWK score of 0.52 when all subjects' ultradian rhythms were estimated correctly.

Introduction

Sleep is the most important activity in life and has a significant impact on human physical health and mental health (Scott et al. 2021). According to the Ministry of Health, Labour and Welfare in Japan, one out of every five Japanese respondents states that they do not feel rested from sleep or have some kind of insomnia symptoms. Particularly among the elderly, an increasing number of people, even traditionally healthy people, have trouble falling asleep, sleeping more lightly, and waking up more frequently in the middle of the day. If sleep problems are left untreated, physical and mental health deteriorate, and the quality of sleep deteriorates further, creating a vicious cycle. For this problem, the number of clinics and counselors specializing in sleep problems has increased in Japan. In addition to this trend, online services for diagnosis are also becoming more widespread. These facts indicate that the demand for sleep counseling is increasing in Japan. However, the number of counselors has not kept pace with the increase in the number of suffers, as the number of people with advanced knowledge about sleep is limited.

In recent years, the field of natural language processing (NLP) has witnessed significant advancements, one of which is the development of ChatGPT by OpenAI. ChatGPT is a state-of-the-art language model based on the GPT (Generative Pre-trained Transformer) architecture. This innovative technology leverages deep learning techniques to understand and generate human-like text, offering remarkable capabilities in a wide range of applications, from conversational agents and automated writing assistance to complex problem-solving tasks. In the realm of public health, the advent of generative AI models like ChatGPT represents a promising frontier for enhancing healthcare delivery and patient support (Baclic et al. 2020). Particularly noteworthy is the application of ChatGPT in mental health support, where it can serve as an accessible first line of assistance. ChatGPT, with its ability to provide instant, empathetic, and informed responses, could act as a supplement or even an alternative to human counselors in addressing common sleep concerns. This approach not only makes mental health support more accessible but also reduces the stigma associated with seeking help, encouraging more individuals to address their sleep problems proactively.

To leverage generative AI models like ChatGPT as sleep counselors effectively, it is crucial to pre-train these models with sleep-related data. Clinics accumulate such data that responses to past sleep problems which could potentially be utilized to develop high-accuracy AI sleep counselors. By utilizing models trained on comprehensive sleep-related response data, it becomes possible to address concerns such as "I can't sleep at night, what should I do?" Such models can list a range of potential causes, including irregular sleep habits, stress or anxiety, changes in the living environment, and the consumption of caffeine or alcohol. They are then capable of suggesting practical solutions, like seeking exposure to sunlight at the start of the day or engaging in moderate exercise, to help establish a healthy sleep rhythm. However, sleep concerns vary greatly from one individual to another, necessitating more detailed information for generating specific and personalized advice. In particular, daily sleep patterns and conditions are important to generate a response to sleep problems.

As a first step towards realizing an AI sleep counselor capable of generating personalized advice, this paper proposes a method for monitoring daily sleep conditions with a mattress sensor. Many non-contact devices, wearable devices and smartphone applications that estimate the sleep stages by machine learning (ML) have been developed. De et. al. summarized the accuracy of the sleep stage estimation with wearable devices such as smartwatches and most methods employ ML to estimate sleep stage (De Zambotti et al. 2019), and they conducted in the paper that "in using wearable sleep technology, a careful path forward needs to be navigated, understanding the benefits and pitfalls of this technology as applied in sleep research and clinical sleep medicine.," this means that they have not enough accuracy of sleep stage estimation which are available in recent years (e.g., smartwatches, smart rings and so on). When sleep stages are estimated by AI, the proportion of NREM2 sleep, one of the sleep stages, accounts for about 60% of the total sleep, resulting in over-estimating of NREM2 sleep and difficulty in understanding sleep structure. To tackle the problems, this paper introduces sleep domain knowledge to machine learning (ML) for improving the accuracy of sleep stage estimation. Concretely, the proposed method employs ultradian rhythm and updates prediction probabilities of each sleep stage by ML model and applies WAKE/NR3 detection based on the large/small body movement.

Sleep Stage

The sleep stage represents the depth of sleep defined by the Rechtschaffen and Kales method (Rechtschaffen and Kales 1968) method. The depth of sleep is classified into the five stages in each epoch (30 seconds), i.e., WAKE, REM, Non-REM1 (NR1), NR2 and NR3 (NR4 is often grouped with NR3). In order to determine the sleep stage, the biological data of electroencephalography (EEG), electrooculogram (EOG), and electromyogram (EMG) are needed and they are acquired in the polysomnography (PSG) test. Fig. 1 shows an example of the sleep stage of a healthy person, where the vertical and horizontal axes indicate the sleep stage and the time, respectively. As shown in this figure, the deep sleep (i.e., the NR3 sleep) and the shallow sleep (i.e., the REM sleep) repeat alternately, and this sleep cycle follows the ultradian rhythm, which is found three to five times during sleep.

Related Works

Methods Based on Physiological Knowledge

Watanabe et. al. tried to extract the relation between the change in the heart rate and sleep stages through the frequency band containing the multiple biological rhythms of a human to build a foundation of sleep stage estimation from heart rate variability (HRV) (Watanabe and Watanabe 2001). They focused on two biological rhythms, the ultradian rhythm and the circadian rhythm, which is an approximate 25 hours cycle. From their study, the relations between the frequency of HRV and sleep stage have been revealed, and they built a sleep stage estimation method based on the heart rate data acquired from the air mattress sensor (Watanabe and Watanabe 2004).

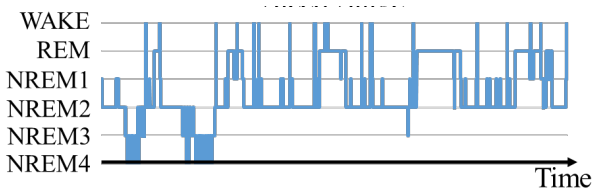


Figure 1: Overnight sleep stage.

Methods Based on Machine Learning

Tal et. al. used the mattress sensor to estimate sleep stage, compared the result with the PSG test for 63 subjects (85 nights), and the 4-stage accuracy was 65% (Tal et al. 2017).

Liu et. al. studied the sleep-wake (i.e., the two sleep stages) estimation model by the hidden Markov models (HMM) (Liu et al. 2020), and found the significant differences in the estimated parameters of HMM among subjects (i.e., the inter-individual difference) and that between weekday and weekend of the same subjects (i.e., the intra-individual difference). Such inter/intra-individual differences make it difficult for ML to correctly learn the sleep stage by the conventional ML model.

Proposed Method

We previously proposed sleep stage estimation by ML that updates prediction probabilities with estimated ultradian rhythm (Nakari and Takadama 2023). The proposed method in this paper changes the ultradian rhythm estimation method and adds WAKE/NR3 detection. Figure 2 shows the overview of the proposed framework, which is executed as follows:

1. **Feature selection:** The effective features for classifying the sleep stages are selected from Python package tsfresh (Time Series FeatuRe Extraction on the basis of Scalable Hypothesis tests) (Christ et al. 2018) using the sensor values (i.e., the biological vibration data composed of the vibration of heartbeats, respirations and body movements).
2. **ML:** The ML model is trained for classifying the sleep stages in each epoch with the features selected in 1. And, the prediction probabilities of the sleep stages in each epoch are calculated from the ML model.
3. **Ultradian rhythm estimation:** Since we cannot acquire the correct ultradian rhythm by non-contact sensors, this paper estimates the row ultradian rhythm by body movement density (Dement and Kleitman 1957).
4. **Probability update:** The predicted probabilities of all sleep stages are updated to be close to the estimated ultradian rhythm.
5. **WAKE/NR3 Detection:** Since large body movements tend to occur during WAKE, it can be detected from the relationship between the magnitude of body movements overnight. And, since NR3 sleep is stable sleep with no

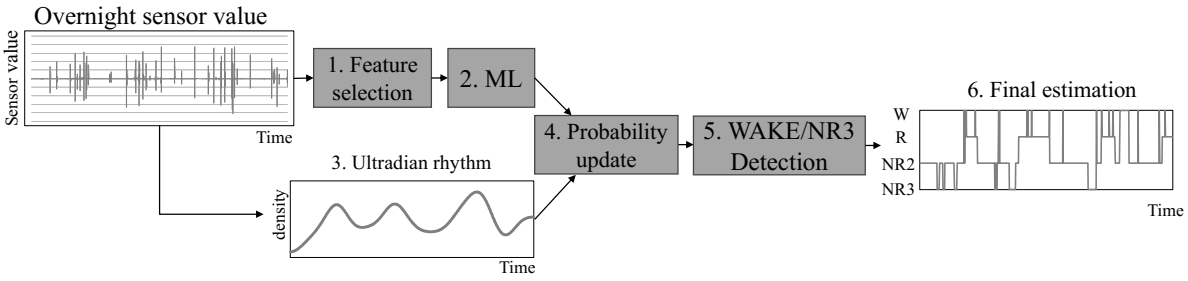


Figure 2: Overview of the proposed framework.

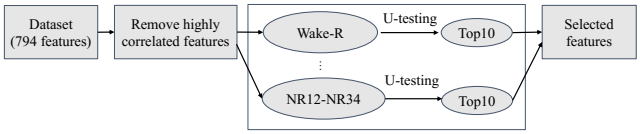


Figure 3: Flow of feature selection.

body movement, the period where body movement does not occur (i.e., the epoch where the estimated UR value is low) is detected as NR3.

6. **Sleep stage adjustment:** The estimated sleep stage is adjusted in each epoch according to the updated probabilities of the sleep stage and WAKE/NR3 detection in 4.

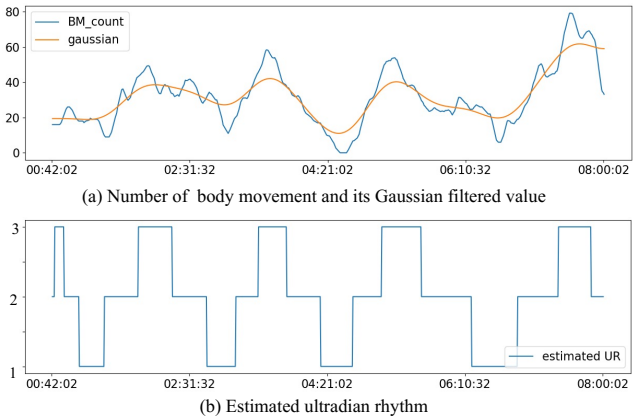
Feature Selection

Figure 3 shows the flow of the feature selection. First, the 794 time series features based on the 63 time series characterization method in the Python package tsfresh are calculated in all epochs, and they are standardized for each subject. To avoid multicollinearity, the highly correlated features (i.e., correlation coefficient greater than 0.7) are removed. After that, to find the effective features for classifying the sleep stages, the Mann-Whitney U-test is applied for all combinations of the two sleep stages (e.g., WAKE-REM, WAKE-N12,...,N12-N34) which total is ${}_4C_2$ and the top 10 features with the smallest p-value are selected in each combination. If the selected top 10 features are overlapped among the combination of the two sleep stages, the unique features are selected for training a ML model.

Ultradian Rhythm Estimation

Since body movements are more likely to occur in shallow sleep and less likely to occur in deep sleep, we estimate the ultradian rhythm by calculating the number of body movements that occur in a given period (i.e., 15 minutes before and after). Algorithm 1 shows the algorithm of calculation of the number of body movements, where the "sensorValues" indicates the list of the bio-vibration data (i.e., acquired from the mattress sensor), "bmCounts" indicates the number of body movements (i.e., it is indicated as BM_count in Figure 4), and $\mathrm{TH_{BM}}$ is calculated by Equation (1).

$$\mathrm{TH_{BM}} = \overline{\mathrm{sensorValues}} + 4\sigma_{\mathrm{sensorValues}} \quad (1)$$



(a) Number of body movement and its Gaussian filtered value

(b) Estimated ultradian rhythm

Figure 4: Ultradian rhythm estimation.

In Equation (1), $\overline{\mathrm{sensorValues}}$ indicates the average of overnight bio-vibration data except for the top 1% values, and $\sigma_{\mathrm{sensorValues}}$ indicates the standard deviation of that. Figure 4 (a) shows the number of body movements with the blue line (i.e., it is calculated by Algorithm 1) and its Gaussian filter value with the orange line. The vertical axis indicates the number of body movements and the horizontal axis indicates the time. As shown in Figure 4(a), the line of the number of body movements is jagged as it is, so a Gaussian filter is applied to make the curve smooth like the orange line. The orange line is discretized into 3 sleep depth states (3 (shallow), 2 (middle), and 1 (deep)) by finding the maximum, inflection point, and minimum of the orange line, with each point as the center. Figure 4 (b) shows the discretized orange line, which is the estimated ultradian rhythm in this paper, where the vertical axis indicates the sleep depth state and the horizontal axis indicates the time.

Probability Update

The proposed method updates the prediction probability of ML in each sleep stage based on Equation (2),

$$P_s = P_{MLs}(1 + W_{s,d}) \quad (2)$$

where P_s indicates the updated prediction probability of ML in sleep stage s, P_{MLs} indicates the original prediction probability of that, and $W_{s,d}$ indicates the weight for sleep stage s in sleep depth state d. The parameter settings of $W_{s,d}$ are

Algorithm 1: Calculation of the number of body movements

IN: sensorValues (list of bio-vibration data)
OUT: $bmCounts$ (list of BM count every epoch)
for $i = 0$ to sensorValues.length/30 **do**
    Count the number of body movements exceeding $\mathrm{TH}_{\mathrm{BM}}$ for the data of 15 minutes before and after every epoch, and add it to $tempBmCounts$.
end for
for $i = 0$ to tempBmCounts.values **do**
    Calculate the average of the count values (i.e., tempBmCounts) for the 5 epochs before and after, and add it to $bmCounts$.
end for

state	WAKE	REM	N2	N34
shallow	0.3	0.3	0.1	-0.1
middle	0.2	0.0	0.2	0.0
deep	-0.1	-0.1	0.1	0.3

Table 1: Parameter settings of $W_{s,d}$.

summarized in Table 1. Note that, the parameters are determined from the results of preliminary experiments (cross-validation). Since the sleep stages become shallow in a shallow sleep state, set a large positive value for $W_{s,d}$ to update the prediction probabilities of WAKE and REM sleep become higher. While in a deep sleep state, set a small negative value for $W_{s,d}$ to update the prediction probabilities of WAKE and REM sleep to become lower and set a large positive value for $W_{s,d}$ to update the prediction probabilities of NR3 sleep to become higher.

WAKE/NR3 Detection

WAKE Detection: Since large body movements are likely to occur in a WAKE stage, the proposed method detects WAKE based on the size of the body movement. Concretely, the proposed method detects WAKE if the size of the body movement is larger than $\mathrm{TH}_{\mathrm{WAKE}}$ as shown in Equation (3),

$$\mathrm{TH}_{\mathrm{WAKE}} = \overline{\mathrm{BM}} + \sigma_{\mathrm{BM}} \tag{3}$$

where $\overline{\mathrm{BM}}$ indicates the average size of overnight body movements and σ_{BM} indicates the standard deviation of that.

NR3 Detection: Since NR3 sleep has fewer body movements, the proposed method detects NR3 sleep based on the number of body movements in a long time window. Concretely, the proposed method detects NR3 if the number of body movements (i.e., the values are equal to the number of body movements calculated in ultradian rhythm estimation) is smaller than $\mathrm{TH}_{\mathrm{NR3}}$ as shown in Equation (4),

$$\mathrm{TH}_{\mathrm{NR3}} = \overline{\mathrm{BM\_count}} - \sigma_{\mathrm{BM\_count}} \tag{4}$$

where $\overline{\mathrm{BM\_count}}$ indicates the average size of overnight of the number of body movements and $\sigma_{\mathrm{BM\_count}}$ indicates the standard deviation of that. This NR3 sleep detection sometimes might be over-detection. To prevent over-NR3 detection, the proposed method cancels the NR3 detection when the predicted probability of N12 is in the top 60% overnight.

Human Subject Experiment

Experimental Setup

To investigate the effectiveness of the proposed method, PUSS-UR, this paper conducted the human subject experiment which estimated the sleep stage from the sleep data of the 50 nights with the 30 healthy subjects, which is composed of the 30 nights of 20's subjects, the seven nights of the 30's subjects, the seven nights of the 40's subjects, the five nights of the 50's subjects, and the one night of the 60's subject. This experiment compares the results of the sleep stage estimation by ML and ML w/ the proposed method. For the ML, this paper employed Random Forests (Breiman 2001), and the parameter for the number of trees is 100 and the max depth is 10. The biological data of the subjects is measured in the PSG test and the bio-vibration data is acquired from the mattress sensor (TANITA sleep scan SL511 with a sampling rate is 16 Hz) placed under the mattress of the bed. The bio-vibration data includes the vibrations of heartbeats, respirations and body movement. For the human subject experiment, the ethics community of Ota General Hospital approved this study in the agreement with Helsinki's declaration, and all the subjects signed their consents.

Evaluation Criteria

The results of the sleep stage estimation were evaluated by the leave-one-out cross-validation. As evaluation criteria, the accuracy, the quadratic weighted kappa (QWK) coefficient (Cohen 1960), and the recall in the estimated sleep stage are employed.

Results

Figure 5 shows the box-and-whisker plots of sleep stage estimation results for all subjects, where the blue colors are the results of the RF and the orange colors are the results of the RF w/ the proposed method. The averaged 4-stage accuracy and QWK score, and the recalls for NR3, NR12 and WAKE of the results of the RF /w the proposed method are higher than the results of the RF. Concretely, in RF, the averaged 4-stage accuracy is 61.5%, the averaged 4-stage QWK score is 0.196, and in RF w/ the proposed method, the averaged 4-stage accuracy is 65.0%, 4-stage QWK is 0.297. Focusing on the 4-stage QWK score, the result of RF w/ the proposed method has a smaller variance value than the result of RF and is more stable in estimating sleep stages.

When comparing the results of the RF and RF w/ the proposed method, Focusing on the subjects whose ultradian rhythms were correctly estimated (n=29), Figure 6 indicates the 4-stage accuracy and QWK score, where the blue bar and orange bar indicate the result of RF and that of RF w/ the proposed method, respectively. The averaged 4-stage accuracy in RF and RF w/ the proposed method are 60.5% and 66.9%, and the averaged 4-stage QWK score in RF and RF w/ the proposed method are 0.20 and 0.36 respectively. The for-stage accuracy and QWK score have significant differences with the t-test between RF and RF w/ the proposed method, and the performance of RF w/ the proposed method is better than RF.

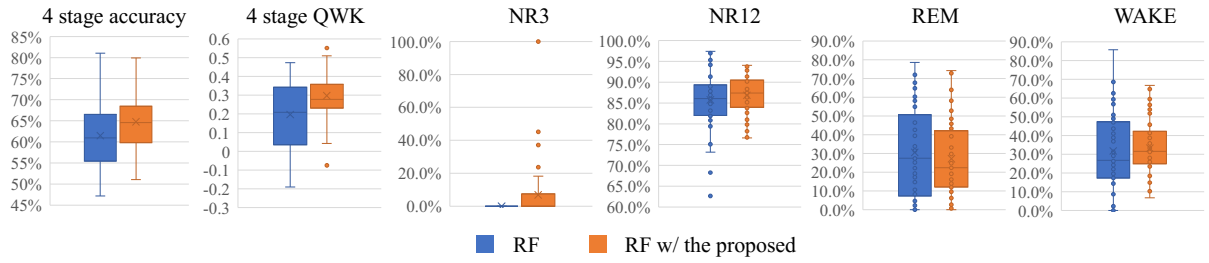


Figure 5: Results of 4-stage accuracy and QWK, and recall of each sleep stage (n=50).

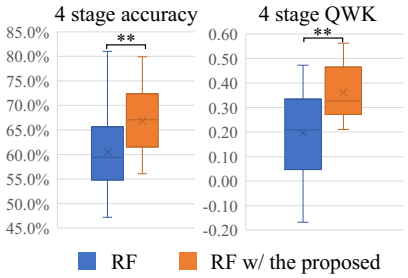


Figure 6: Results of 4-stage accuracy and QWK in subjects with correctly estimated ultradian rhythm (n=29).

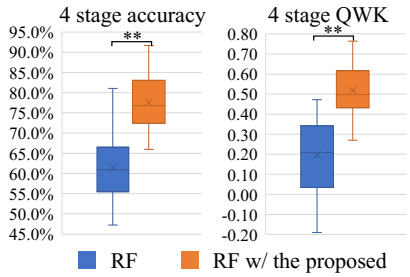


Figure 7: Results of 4-stage accuracy and QWK with correct ultradian rhythm calculated PSG test (n=50).

Discussions

Results with Correct Ultradian Rhythm

Ultradian rhythms estimated by the proposed method had inverted rhythms compared to the correct Ultradian rhythms calculated from the correct sleep stage acquired from the PSG test in 21 of the 50 subjects. Such misestimation of ultradian rhythms caused estimation accuracy to worsen or stagnate. Figure 7 shows the results of RF (blue bar) and RF w/ the proposed method with the correct ultradian rhythm (orange bar). The averaged 4-stage accuracy in RF and RF w/ the proposed method are 61.5% and 77.6%, and the averaged 4-stage QWK score in RF and RF w/ the proposed method are 0.196 and 0.52 respectively. The for-stage accuracy and QWK score have significant differences with the t-test between RF and RF w/ the proposed method, and the performance of RF w/ the proposed method is better than RF.

From the analysis, if the estimated ultradian rhythm deviates from the correct ultradian rhythm (especially in the inverted rhythm), the update of the prediction probability by the proposed method does not work and leads to worse estimation results, but if the ultradian rhythm can be estimated correctly, a significant improvement can be expected.

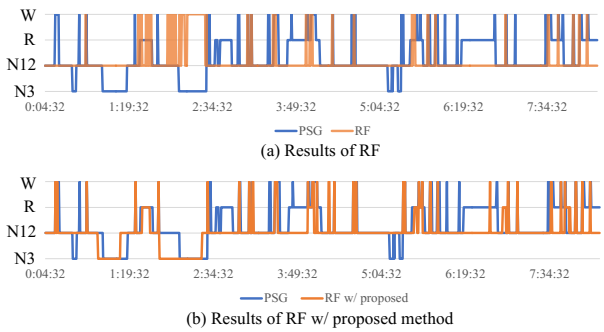
Example of Improved Sleep Stage Estimation

Figures 8 are the examples of the detailed sleep stage estimation result with (a) RF and (b) RF w/ the proposed method, where the vertical axis indicates the sleep stage, the horizontal axis indicates the time, the blue lines indicate the correct sleep stage acquired from PSG and the orange lines indicate the sleep stage estimation by RF or RF w/ the proposed method. As shown in Figure 8(a), the sleep stage estimation result with RF is over-estimating NR12 sleep and has no estimation for REM sleep and NR3 sleep. Furthermore, the wrong WAKE estimations are noticeable around 2 am. The sleep stage estimation result with RF w/ the proposed method, as shown in Figure 8(b), NR3 sleep and some REM sleeps are correctly estimated and the wrong WAKE estimations are reduced.

From the analysis, the proposed method may improve the estimation of NR3 sleep and REM sleep in cases where NR12 sleep is overestimated and the sleep structure is difficult to understand, and may facilitate the understanding of sleep structure. The derivation of a more accurate sleep structure than conventional RF using the proposed method is useful for understanding sleep quality and sleep cycles. By inputting this information, along with sleep concerns, to a generative AI, it is possible to output personalized advice.

Conclusion

To improve the accuracy of sleep stage estimation and to get accurate sleep structure, this paper introduced sleep domain knowledge to machine learning for improving the accuracy of sleep stage estimation. Concretely, the proposed method estimates ultradian rhythm based on the body movement density, updates prediction probabilities of each sleep stage by ML model and applies WAKE/NR3 detection based on the large/small body movement. To investigate the effectiveness of the proposed method, this paper conducted the human subject experiment, and revealed the following implications: (1) the proposed method improved the percent-



(a) Results of RF

(b) Results of RF w/ proposed method

Figure 8: Example of the sleep stage estimation results.

age of Accuracy by 65.0% from 61.5% and the QWK score by 0.196 from 0.297 by the conventional machine learning method; (2) the proposed method prevents over-NR12 estimating and is useful for understanding sleep structure by estimating REM sleep and NR3 sleep correctly. (3) the correct estimation of ultradian rhythms significantly improved the sleep stage estimation, with an Accuracy of 77.6% and a QWK score of 0.52 when all subjects' ultradian rhythms were estimated correctly.

Our future goal is to develop sleep counseling AI which can provide personal advice for better sleep by inputting daily sleep conditions in addition to users' sleep concerns. This marks a significant advancement in using generative AI to improve health care and public health by offering personalized help for sleep improvement. For the future goal, the following are the future works: (1) Integrating real-time sleep data monitoring with wearable devices or smart home systems to allow for more dynamic and immediate feedback to users based on their current sleep conditions. (3) Conducting extensive validation studies with diverse populations to ensure the generative AI's recommendations are effective across different age groups, health conditions, and cultural backgrounds. (4) Implementing machine learning techniques that can adapt and evolve with individual user feedback, enabling the system to become more accurate and personalized over time. (5) Developing a user-friendly interface that encourages regular interaction and makes it easy for users to understand and apply the AI's advice for improving their sleep quality.

Acknowledgments

This work was supported Grant-in-Aid for JSPS Research Fellow, Grant Number 22KJ1367.

References

Baclic, O.; Tunis, M.; Young, K.; Doan, C.; Swerdfeger, H.; and Schonfeld, J. 2020. Artificial intelligence in public health: Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*, 46(6): 161.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1): 5–32.

Christ, M.; Braun, N.; Neuffer, J.; and Kempa-Liehr, A. W. 2018. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307: 72–77.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

De Zambotti, M.; Cellini, N.; Goldstone, A.; Colrain, I. M.; and Baker, F. C. 2019. Wearable sleep technology in clinical and research settings. *Medicine and science in sports and exercise*, 51(7): 1538.

Dement, W.; and Kleitman, N. 1957. Cyclic variations in EEG during sleep and their relation to eye movements, body motility, and dreaming. *Electroencephalography and clinical neurophysiology*, 9(4): 673–690.

Liu, J.; Zhao, Y.; Lai, B.; Wang, H.; Tsui, K. L.; et al. 2020. Wearable device heart rate and activity data in an unsupervised approach to personalized sleep monitoring: algorithm validation. *JMIR mHealth and uHealth*, 8(8): e18370.

Nakari, I.; and Takadama, K. 2023. Personalized Noncontact Sleep Stage Estimation with Weighted Probability Estimation by Ultradian Rhythm. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1–4. IEEE.

Rechtschaffen, A.; and Kales, A. 1968. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington DC.

Scott, A. J.; Webb, T. L.; Martyn-St James, M.; Rowse, G.; and Weich, S. 2021. Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials. *Sleep medicine reviews*, 60: 101556.

Tal, A.; Shinar, Z.; Shaki, D.; Codish, S.; and Goldbart, A. 2017. Validation of contact-free sleep monitoring device with comparison to polysomnography. *Journal of Clinical Sleep Medicine*, 13(3): 517–522.

Watanabe, T.; and Watanabe, K. 2001. Estimation of the sleep stages by the non-restrictive air mattress sensor relation between the change in the heart rate and sleep stages. *Transactions of the Society of Instrument and Control Engineers*, 37(9): 821–828.

Watanabe, T.; and Watanabe, K. 2004. Noncontact method for sleep stage estimation. *IEEE Transactions on biomedical engineering*, 51(10): 1735–1748.