Feature Extraction for Machine Learning-based Sleep Stage Classification Using PPG-Derived Parameters and Skin Temperature

Suphachok Buaruk*, Sikawat Thanaviratananich†, Peerasit Treesuthacheep‡ and Somrudee Deepaisarn§
* Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
E-mail: d6522300067@g.siit.tu.ac.th, s.buaruk@gmail.com
† Division of Neurology, Department of Medicine, Bumrungrad International Hospital, Bangkok, Thailand
Sleep Disorders Center, Cleveland Clinic Neurological Institute, Cleveland, Ohio, USA
E-mail: thanavir@gmail.com
‡ Chulalongkorn Comprehensive Epilepsy Center of Excellence, King Chulalongkorn Memorial Hospital, Bangkok, Thailand
E-mail: ptpeerasit@gmail.com
§ Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
Research Unit in Sustainable Electrochemical Intelligent, Thammasat University, Pathum Thani, Thailand
E-mail: somrudee@siit.tu.ac.th (Corresponding Author)

***Abstract*—This study explores feature extraction and machine learning methods for automated sleep stage classification using physiological signals, focusing on PPG-derived parameters and skin temperature. We experimented with feature extraction techniques that required different engineering steps, i.e., with or without applying the first derivative and deviation from the mean of raw signals. The statistical values, including mean, standard deviation, maximum, minimum, and median of each 30-second segment of signals, are then extracted and used as input features for our selected classification model, the Extra Trees. The classification performance is evaluated under both subject-dependent and subject-independent scenarios using data from 93 subjects. The data include labeled stages of wake, non-rapid eye movement (NREM) sleep stages N1, N2, and N3, as well as rapid eye movement (REM) sleep. Optimal features were derived from skin temperature measurement along with the normalized timestamp, which is associated with sleep duration. Results show that the subject-dependent experiment yields the highest performance with an accuracy of 0.87 for three-stage classification (wake, NREM:N1/N2/N3, REM) and 0.77 for five-stage classification (wake, N1, N2, N3, REM). This work highlights the potential of artificial intelligence to enable the automatic labeling of sleep stages using non-constrained sensors. The important features related to sleep staging can be investigated, which could contribute to future advances in clinical diagnostics and healthcare applications.**

I. Introduction

Sleep is vital for health and well-being. It helps the body recover, enhances brain function, and maintains emotional stability. Good sleep is crucial for memory, a strong immune system, and proper metabolism. Not getting enough sleep or having poor-quality sleep can lead to health conditions, such as heart disease, diabetes, and cognitive difficulties [1]. Sleep staging is the process of categorizing different phases of sleep. It is crucial for understanding sleep patterns, diagnosing sleep disorders, and developing effective treatment strategies. The gold standard for sleep staging is Polysomnography (PSG), a recording of the bio-physiological changes that occur during sleep [2]. PSG involves the use of electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) signals to monitor brain activity [3]. These signals are essential for distinguishing sleep stages [4]. However, PSG is complex and requires a high level of technical expertise to interpret. According to its complexity and the need for technical expertise to interpret, a study by the American Academy of Sleep Medicine (AASM) found that the sleep staging agreement among 2500 scorers using PSG is around 83% efficient [5]. Another study examined the overall agreement for sleep staging between seven scorers, finding 82% agreement rate [6]. Both studies found the highest agreement on REM sleep, while the lowest was on N1 sleep. The procedure is typically conducted in a clinical setting with many sensors and electrodes, which can be uncomfortable for the patient. There is a growing interest in finding less invasive and more accessible methods for sleep staging [7]. Recent research has shown that skin body temperature holds the potential for sleep staging. Biologically, skin temperature is affected by the thermoregulatory processes, which vary during different sleep stages [8]–[10]. For instance, skin temperature tends to increase during the transition to sleep and remains relatively higher during REM sleep [11]. However, there is a lack of quantitative studies exploring the direct use of skin temperature for sleep staging. In addition, the use of photoplethysmography (PPG) on sleep staging can be used to detect sleep and wake stages. This detection relied on the combined features between heart rate variability (HRV) and ECG-derived respiration (EDR) signals, measured from wearable sensors [12]. To the best of our knowledge, no studies investigate the combination of skin temperature and PPG-derived parameters as analyzing features for sleep staging.

Advancements in machine learning (ML) and deep learning (DL), combined with the biological signals from PSG, have opened new techniques for sleep staging. This approach could be applied to offer more efficiency in the aspect of manpower

compared to the traditional techniques [13]. In the context of sleep staging, the ML technique can process signals from PSG devices to classify different sleep stages. As a subset of ML, DL takes this a step further by using neural networks with multiple layers to model complex and non-linear relationships in the data. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been particularly effective in handling time-series data and recognizing intricate patterns associated with sleep stages. Many studies have adopted CNNs and RNNs for sleep staging with signals retrieved from PSG. Li. F. *et al.* developed an end-to-end automatic sleep staging method by proposing the stacking model architecture with a collection of consecutive convolutional micro-networks (CCNs) and squeeze-excitation (SE) block [14]. The proposed architecture was applied to use different single-channel EEGs. Research has constructed sleep staging models based on the random forest algorithm to distinguish four sleep stages based on core body temperature rhythm, achieving accuracies of 0.70 for males and 0.77 for females [15]. PPG signals, measured by finger pulse oximeters in a high prevalence of obstructive sleep apnea (OSA) patients, were employed with combined CNNs and RNNs to classify sleep stages into three (wake/NREM/REM), four (wake/N1+N2/N3/REM), and five (wake/N1/N2/N3/REM). The performance results showed that the three-stage model achieved 80.1% accuracy, while the four- and five-stage models achieved accuracies of 68.5% and 64.1%, respectively [16].

In this study, skin temperature and a set of parameters derived from PPG sensors are used to classify sleep stages. The statistical features, including the mean and standard deviation of these signals, are utilized as inputs for classical ML approaches. Additionally, the first derivatives of each prepared signal are included to capture the rate of change and further enhance the model's ability to detect complex variations in the data. Various experimental scenarios are explored, including different numbers of sleep stage labels, subject dependency, and groups of features. This comprehensive approach aims to understand the effectiveness and robustness of the proposed methods in classifying sleep stages. By using easy-access tools, it offers a more accessible and less complex alternative to traditional PSG-based techniques, making sleep stage classification more feasible for a broader range of users and applications.

The subsequent sections of this paper are organized as follows. Section II describes the methodologies employed in this study, including an overview of the dataset, data preparation techniques, and classification experiments. In section III, the results and discussion regarding the classification performance of each experimental scenario are presented. Finally, section IV provides a summary of the study and explores potential directions for future research.

II. Method

A. Dataset

This work utilizes a public dataset acquired from 100 unique participants by the Duke University Health System Sleep Disorder Lab [17], [18]. This sleep study aimed to detect and monitor apnea events during sleep using the PSG gold-standard approach to sleep testing. PSG data collection began around 11 PM and continued until the participant naturally awoke around 6 AM, resulting in a 7-hour sleep test on each participant. In addition to PSG components, the Empatica E4 wristband (Empatica Inc., Milano, Italy) was placed on the left wrist of each participant and was deactivated upon awakening. The E4 device collected six raw signals, including blood volume pulse (BVP) from the PPG sensor, accelerometry in three axes, electrodermal activity (EDA), and skin temperature. In the dataset, heart rate (HR) and interbeat interval (IBI) were derived from the BVP. Additionally, technician-annotated sleep stages derived from the PSG data were recorded every 30 seconds. All data were resampled to a frequency of 64 Hz and synchronized using timestamps, which were time-shifted to start from zero. An example of raw signals used in this study, including skin temperature, BVP, HR, and IBI, is presented in Fig. 1.

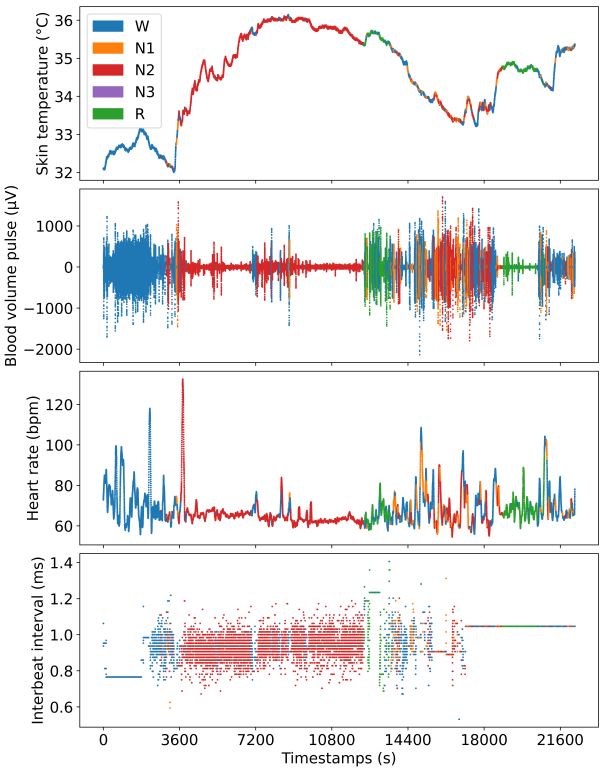

Fig. 1. Signal profiles of skin temperature, blood volume pulse, heart rate, and interbeat interval.

B. Feature Extraction

In this study, a total of 93 subjects were selected from the initial 100 subjects of the dataset. The 7 subjects were excluded from the analysis due to either having more than one sleep stage within certain 30-second intervals or having labels indicating a "preparing stage" (the stage before the PSG recording starts) between the five sleep stages. The preprocessing and

feature extraction steps involved the following:

1) **Feature Signals**: The feature signals used in this work include derived signals from a PPG sensor, specifically heart rate (HR), blood volume pulse (BVP), interbeat interval (IBI), and skin temperature.
2) **Resampling**: All signals were resampled from a 64 Hz sampling frequency to 1 Hz.
3) **Feature Engineering - Derivative Set**: The first derivative was applied to all resampled signals to create the derivative set. For a given signal x, the first derivative Δx_i is calculated as:

$$\Delta x_i = x_i - x_{i-1}, \tag{1}$$

Note that the first sample's derivative is initialized to zero: $\Delta x_1 = 0$. For the last sample, the derivative is computed as the difference between the last two samples to maintain consistency: $\Delta x_n = x_n - x_{n-1}$.

4) **Feature Engineering - Deviation Set**: The resampled signals were subtracted from their mean values to generate the deviation set. For a given signal x with n samples, the mean μ_x is calculated as:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

The deviation of each sample is then calculated as:

$$d_i = x_i - \mu_x \tag{3}$$

5) **Segmentation**: Both the derivative set and deviation set were segmented into 30-second epochs, with each segment labeled according to the corresponding sleep stage.
6) **Final Dataset Creation**: Each segment (or chunk) was processed to compute the mean and standard deviation for both the derivative set and the deviation set. These computed features formed the final dataset used for the classification task in sleep staging.

C. Classification Experiments

The classification experiment investigates various scenarios by varying factors involved in the training and evaluation of sleep staging classification. The parameters include:

1) **Number of stages**: The original dataset includes signals along with labels for five sleep stages: Wake (W), Non-Rapid Eye Movement (N1/N2/N3), and Rapid Eye Movement (R). This study examines three different classification schemes:
   - Three stages: Wake, combined N1/N2/N3, and REM
   - Four stages: Wake, combined N1/N2, N3, and REM
   - Five stages: Wake, N1, N2, N3, and REM
2) **Dependency**:
   - Subject-independence: Ensuring that subjects in the training and test sets are separate, with 74 subjects in the training set and 19 subjects in the test set. All experiments under subject-independent conditions will be evaluated using the same test set.
   - Subject-dependence: Splitting the data by a 70:30 ratio for training and testing, respectively. The splitting is performed by grouping ID and sleep stages, ensuring that all stages from each ID are split with the same ratio in both the training and test sets.
3) **Classification**: The classification training model is set up using the PyCaret framework [19], which provides robust data preprocessing steps. Since the imbalance in the sleep stages labels within the dataset, PyCaret addresses this issue by applying the SMOTE (Synthetic Minority Over-Sampling TEchnique). Additionally, the framework uses stratified K-fold cross-validation to ensure that all stages from each ID are separated in the same ratio in both the training and validation sets. In this work, the Extra Trees (ET) algorithm is employed as the classification algorithm with a 10-fold cross-validation. The performance of the classification model is monitored using several key metrics, including Accuracy, Recall, Precision, and F1-score, applied to the test dataset to compare the performance across different classification experiment scenarios.

III. Results and Discussion

A. Data Exploration

Fig. 1 illustrates an example of measured signals from the Empatica E4 wristband device. The skin temperature measurements obtained from this device exhibit a pattern similar to the distal temperature reported in the study of C. Cajochen *et al.* [20]. This finding suggests that the distal-proximal temperature relationship should be considered when evaluating the performance of the easy-access tool used in this study.

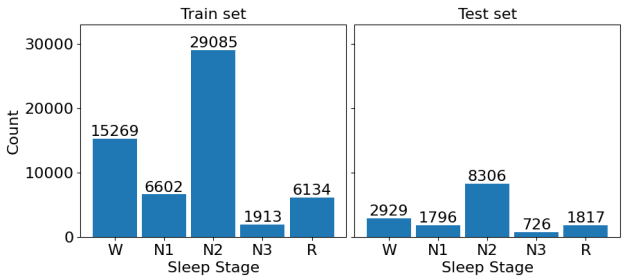

Fig. 2. Number of each sleep stage on the train (left) and test (right) datasets, split by subject-independent scenarios.

Fig. 2 illustrates the significant imbalance in the distribution of labels across the 74 subjects and 19 subjects in training and test sets, respectively, from the subject-independent strategy. The majority of the samples belong to the N2 and Wake stages, while the N3 stage has the fewest samples. This distribution could be influenced by the first night effect (FNE), which increases time in the light sleep stages and decreases time in deep sleep stages [21]. FNE also increases the transition probability from stage N2 to N1 and decreases the transition probability from stage N2 to N3 [22]. Additionally, patients with obstructive sleep apnea (OSA) who have short sleep times

probably experience a decrease in the N3 deep sleep stage and REM stage, along with an increase in N1/N2-stage time due to fragmented sleep architecture [23], [24].

B. Classification Performance

This subsection discusses the performance of each experimental scenario. The parameters include groups of features, including PPG-derived signals, skin temperature, and a combination of both. Additionally, the number of sleep stages used as labels are varied for three-, four-, and five-stages. The details of the subject-dependent and independent scenarios are described as follows:

In the result of subject-independence in Table I, the skin temperature profile shows better performance than using PPG-derived signals alone. However, the use of PPG-derived signals in this study yields lower performance than previous research reported by H. Korkalainen *et al* [16]. This difference might be due to the feature extraction methods in order to represent raw signals.

TABLE I
CLASSIFICATION PERFORMANCE OF SUBJECT-INDEPENDENCE USING STATISTICAL VALUES EXTRACTED FROM RAW SIGNALS, ALONG WITH NORMALIZED TIMESTAMPS.

Features	Stages	Accuracy	Recall	Precision	F1
PPG derived	3	0.37	0.37	0.62	0.39
	4	0.39	0.39	0.53	0.41
	5	0.29	0.29	0.45	0.31
Skin temperature	3	0.58	0.58	0.57	0.57
	4	0.55	0.55	0.51	0.53
	5	0.37	0.37	0.39	0.37
PPG-derived & Skin temperature	3	0.42	0.42	0.61	0.45
	4	0.42	0.42	0.52	0.44
	5	0.29	0.29	0.42	0.30

The significant impact of imbalanced labels also affects the ability to classify less occurred stages accurately. Models may become biased towards the more frequent stages. This imbalance obviously impacts classification performance, especially under the subject-independence. The performance gap between the five-stage classification scenario and the three- and four-stage scenarios is considerable. Separating into five stages requires considering N2 separately from N1 and N3. Since N1 has a significantly lower sample compared to N1 and N2, this results in less support for test classification performance. Consequently, the overall performance in the five-stage classification scenario is lower.

The classification performance improved by 10-16% across all experimental scenarios when using the statistical values of deviation and the first derivative of the raw signal, compared to using the statistical values of the raw signals, as described in Table II.

Table III presents results for the 70:30 train-test split ratios, respectively. The classification performance using skin temperature alone is 87%, 85%, and 77% for three-, four-, and five-stage classification, respectively. Combining both PPG-derived and skin temperature features decreases a bit performance by approximately 1%. across all experiment scenarios. This suggests that including the statistical values of the skin temperature profile as input features can enhance classification performance.

TABLE II
CLASSIFICATION PERFORMANCE OF SUBJECT-INDEPENDENCE USING STATISTICAL VALUES EXTRACTED FROM THE DEVIATION AND FIRST DERIVATIVE OF RAW SIGNALS, ALONG WITH NORMALIZED TIMESTAMPS.

Features	Stages	Accuracy	Recall	Precision	F1
PPG derived	3	0.53	0.53	0.63	0.56
	4	0.52	0.52	0.57	0.54
	5	0.41	0.41	0.46	0.43
Skin temperature	3	0.61	0.61	0.56	0.58
	4	0.56	0.56	0.50	0.52
	5	0.38	0.38	0.37	0.38
PPG-derived & Skin temperature	3	0.55	0.55	0.62	0.57
	4	0.52	0.52	0.55	0.53
	5	0.40	0.40	0.45	0.41

Similar to subject-independent scenarios, the deviation and the first derivative are computed for subject-dependent scenarios. However, as shown in Table IV, the classification performance decreases when using either PPG-derived features or skin temperature features alone. By incorporating a range of known labels for the individuals' sleep stages into the training set, these models can be fine-tuned to enhance their accuracy and reliability. This personalized approach highlights the importance of subject-dependent data in improving sleep stage classification performance.

TABLE III
CLASSIFICATION PERFORMANCE OF SUBJECT-DEPENDENCE USING STATISTICAL VALUES EXTRACTED FROM RAW SIGNALS, ALONG WITH NORMALIZED TIMESTAMPS.

Features	Stages	Accuracy	Recall	Precision	F1
PPG derived	3	0.81	0.81	0.82	0.81
	4	0.79	0.79	0.80	0.79
	5	0.72	0.72	0.71	0.71
Skin temperature	3	0.87	0.87	0.87	0.87
	4	0.85	0.85	0.85	0.85
	5	0.77	0.77	0.77	0.77
PPG-derived & Skin temperature	3	0.86	0.86	0.86	0.86
	4	0.84	0.84	0.85	0.84
	5	0.77	0.77	0.77	0.77

Fig. 3 illustrates the normalized confusion matrix of classification performance using statistical values from the skin temperature profile. This confusion matrix shows the highest performance for REM sleep and the lowest for N1 sleep, aligning with the agreement of sleep staging by scorers [5], [6]. N1 is often characterized as the transitional phase between wakefulness and deeper sleep stages. During this stage, phys-

TABLE IV
CLASSIFICATION PERFORMANCE OF SUBJECT-DEPENDENCE USING STATISTICAL VALUES EXTRACTED FROM THE DEVIATION AND FIRST DERIVATIVE OF RAW SIGNALS, ALONG WITH NORMALIZED TIMESTAMPS.

Features	Stages	Accuracy	Recall	Precision	F1
PPG derived	3	0.76	0.76	0.77	0.76
	4	0.73	0.73	0.75	0.74
	5	0.66	0.66	0.66	0.66
Skin temperature	3	0.85	0.85	0.84	0.84
	4	0.83	0.83	0.83	0.83
	5	0.75	0.75	0.75	0.75
PPG-derived & Skin temperature	3	0.80	0.80	0.80	0.80
	4	0.79	0.79	0.79	0.79
	5	0.71	0.71	0.70	0.70

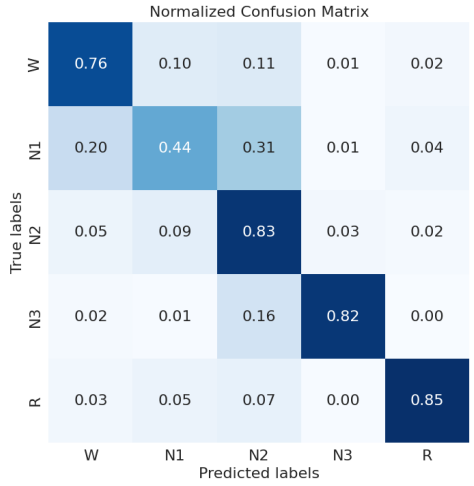

Fig. 3. Confusion matrix for 5 stages classification in subject-dependence, using skin temperature and normalized timestamps.

iological signals, i.e., skin temperature and heart rate, may not differ significantly from neighboring stages, leading the model to misclassify N1 as wake and N2, as presented in Fig. 3. Moreover, it can be seen that the staging performance using skin temperature-related features is better than that of using PPG-derived features. Medical studies have shown that skin temperature, especially near the wrist, increases at the sleep onset and changes predictably during different sleep stages. The distinction between distal and proximal temperature measurements is essential in the context of sleep staging. Distal temperature, measured at locations such as the wrist or ankle over the radial artery [25], can be easily captured using wearable devices.

In this work, the distal skin temperature was also measured using the Empatica E4 wristband device. Understanding the influence of distal temperature on sleep stages is important. During the transition to sleep and throughout NREM sleep, there is an increase in distal skin temperature. This increase is associated with vasodilation, promoting heat loss and helps lower the core body temperature [26]. In fact, distal temperature continues to increase during the N3 deep sleep stage [11]. Additionally, a warm pre-sleep environment before going to bed, known as the "Warm Bath Effect", increases NREM and decreases REM sleep [26]. Furthermore, environmental factors such as ambient temperature should be considered in the development of home-use measurement tools. Higher ambient temperature leads to increased skin temperature, potentially affecting the detection accuracy of sleep stages.

C. Effects of Skin Temperature and Sleep Duration on Sleep Stage Classification

The study on sleep stage classification explores the impact of incorporating normalized timestamps into the feature set. The results, as shown in Section III-B, demonstrate that taking into account the normalized timestamps significantly improves performance in subject-dependent scenarios. This improvement is evident when using either PPG-derived features or skin temperature features alone. However, the improvement is obviously presented with skin temperature features, achieving performance increase in the range of 19-23% depending on the number of sleep stage labels. These findings highlight the importance of temporal information in enhancing the accuracy of machine learning-based sleep staging classification. This suggests the possibility of developing personalized AI models for healthcare, particularly for sleep stage classification tasks.

IV. CONCLUSIONS AND FUTURE WORK

This work demonstrates that simple physiological parameters acquired from the wearable device, including PPG-derived parameters and skin temperature, can be effective for automatic sleep staging classification. The key process is feature extraction and engineering to select the suitable set of feature inputs to train the machine learning model. Our experimental design covers both subject-dependent and independent scenarios for the train-test splitting of the classification model with available sleep stage labels of 93 individual subjects. It was found that the information on skin temperature and sleep duration (i.e., normalized timestamps) led to a highly accurate model. The extra Tree model achieves up to 0.87 in accuracy in the subject-dependent scenario when labeling wake, NREM (N1/N2/N3), and REM. Our proposed model shows agreement with the sleep stage labels by human expert scorers both in terms of the overall classification performance and the bias in labeling stages arising from the ambiguity between staging behavior. The findings can pave the way for artificial intelligent solutions to analyze data from simple, non-constrained sensors to automate sleep stage labeling, which can be useful and practical for the healthcare industry. Environmental factors such as ambient temperature and light intensity can be taken into account in future research. An appropriate experimental design to collect primary data is necessary to improve scoring performance, reduce bias, and gain more impactful insights related to the sleep stage patterns of individuals.

ACKNOWLEDGMENT

The first author (S.B.) acknowledges the Graduate Scholarship Program for Excellent Thai Students, awarded by

Sirindhorn International Institute of Technology, Thammasat University, Thailand, for his doctoral study.

REFERENCES

[1] G. Medic, M. Wille, and M. E. Hemels, "Short-and long-term health consequences of sleep disruption," *Nature and science of sleep*, pp. 151–161, 2017.

[2] S. Keenan, M. Hirshkowitz, and H. Casseres, "Monitoring and staging human sleep," *Principles and practice of sleep medicine*, vol. 5, pp. 1602–1609, 2011.

[3] J. V. Rundo and R. Downey III, "Polysomnography," *Handbook of clinical neurology*, vol. 160, pp. 381–392, 2019.

[4] M. D. Oliver and S. Datta, "Electrophysiological correlates of the sleep–wake cycle," in *The Behavioral, Molecular, Pharmacological, and Clinical Basis of the Sleep-Wake Cycle*, Elsevier, 2019, pp. 17–26.

[5] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Sleep stage scoring," *Journal of clinical sleep medicine*, vol. 9, no. 1, pp. 81–87, 2013.

[6] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer, *et al.*, "Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard," *Journal of sleep research*, vol. 18, no. 1, pp. 74–84, 2009.

[7] T. Lauteslager, S. Kampakis, A. J. Williams, M. Maslik, and F. Siddiqui, "Performance evaluation of the circadia contactless breathing monitor and sleep analysis algorithm for sleep stage classification," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020, pp. 5150–5153.

[8] B. H. Te Lindert and E. J. Van Someren, "Skin temperature, sleep, and vigilance," *Handbook of clinical neurology*, vol. 156, pp. 353–365, 2018.

[9] R. Szymusiak, "Body temperature and sleep," *Handbook of clinical neurology*, vol. 156, pp. 341–351, 2018.

[10] X. Xu, J. Zhu, C. Chen, X. Zhang, Z. Lian, and Z. Hou, "Application potential of skin temperature for sleep-wake classification," *Energy and Buildings*, vol. 266, p. 112 137, 2022.

[11] S. Herberger, T. Penzel, I. Fietze, *et al.*, "Enhanced conductive body heat loss during sleep increases slow-wave sleep and calms the heart," *Scientific Reports*, vol. 14, no. 1, p. 4669, 2024.

[12] R. V. Sharan, "Ecg-derived respiration for sleep-wake stage classification," in *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, IEEE, 2021, pp. 1–4.

[13] D. Sarkar, D. Guha, P. Tarafdar, S. Sarkar, A. Ghosh, and D. Dey, "A comprehensive evaluation of contemporary methods used for automatic sleep staging," *Biomedical Signal Processing and Control*, vol. 77, p. 103 819, 2022.

[14] F. Li, R. Yan, R. Mahini, *et al.*, "End-to-end sleep staging using convolutional neural network in raw single-channel eeg," *Biomedical Signal Processing and Control*, vol. 63, p. 102 203, 2021.

[15] X. Xu and Z. Lian, "A sleep staging model based on core body temperature rhythm," *Energy and Buildings*, vol. 310, p. 114 099, 2024.

[16] H. Korkalainen, J. Aakko, B. Duce, *et al.*, "Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea," *Sleep*, vol. 43, no. 11, zsaa098, 2020.

[17] K. Wang, J. Yang, A. Shetty, and J. Dunn, "Dreamt: Dataset for real-time sleep stage estimation using multisensor wearable technology,"

[18] A. L. Goldberger, L. A. Amaral, L. Glass, *et al.*, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, e215–e220, 2000.

[19] M. Ali, *Pycaret: An open source, low-code machine learning library in python*, PyCaret version 1.0, Apr. 2020. [Online]. Available: https://www.pycaret.org.

[20] C. Cajochen, K. Kräuchi, and A. Wirz-Justice, "Role of melatonin in the regulation of human circadian rhythms and sleep," *Journal of neuroendocrinology*, vol. 15, no. 4, pp. 432–437, 2003.

[21] A. Mayeli, S. Janssen, K. Sharma, and F. Ferrarelli, *Examining first night effect on sleep parameters with hd-eeg in healthy individuals. brain sci. 2022, 12, 233*, 2022.

[22] A. Shirota, M. Kamimura, A. Kishi, H. Adachi, M. Taniike, and T. Kato, "Discrepancies in the time course of sleep stage dynamics, electroencephalographic activity and heart rate variability over sleep cycles in the adaptation night in healthy young adults," *Frontiers in Physiology*, vol. 12, p. 623 401, 2021.

[23] S. Nozawa, K. Urushihata, R. Machida, and M. Hanaoka, "Sleep architecture of short sleep time in patients with obstructive sleep apnea: A retrospective single-facility study," *Sleep and Breathing*, pp. 1–8, 2021.

[24] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset.," *Sleep*, vol. 23, no. 7, pp. 901–908, 2000.

[25] F. M. Acosta, B. Martinez-Tellez, D. P. Blondin, *et al.*, "Relationship between the daily rhythm of distal skin temperature and brown adipose tissue 18f-fdg uptake in young sedentary adults," *Journal of Biological Rhythms*, vol. 34, no. 5, pp. 533–550, 2019.

[26] E. C. Harding, N. P. Franks, and W. Wisden, "The temperature dependence of sleep," *Frontiers in neuroscience*, vol. 13, p. 336, 2019.