

SleepNetZero: Zero-Burden Zero-Shot Reliable Sleep Staging With Neural Networks Based on Ballistocardiograms

SHUZHEN LI*, Tsinghua University, China

YUXIN CHEN*, Tsinghua University, China

XUESONG CHEN, Beijing Wuji Medical Technology Co., Ltd., China

RUIYANG GAO, Beijing Wuji Medical Technology Co., Ltd., China

YUPENG ZHANG, Beijing Wuji Medical Technology Co., Ltd., China

CHAO YU, Tsinghua University, China

YUNFEI LI, Tsinghua University, China

ZIYI YE, Tsinghua University, China

WEIJUN HUANG, Shanghai Jiao Tong University School of Medicine, China

HONGLIANG YI, Shanghai Jiao Tong University School of Medicine, China

YUE LENG, University of California, San Francisco, United States

YI WU, Tsinghua University, China

Sleep monitoring plays a crucial role in maintaining good health, with sleep staging serving as an essential metric in the monitoring process. Traditional methods, utilizing medical sensors like EEG and ECG, can be effective but often present challenges such as unnatural user experience, complex deployment, and high costs. Ballistocardiography (BCG), a type of piezoelectric sensor signal, offers a non-invasive, user-friendly, and easily deployable alternative for long-term home monitoring. However, reliable BCG-based sleep staging is challenging due to the limited sleep monitoring data available for BCG. A restricted training dataset prevents the model from generalization across populations. Additionally, transferring to BCG faces difficulty ensuring model robustness when migrating from other data sources. To address these issues, we introduce SleepNetZero, a zero-shot learning based approach for sleep staging. To tackle the generalization challenge, we propose a series of BCG feature extraction methods that align BCG components with corresponding respiratory, cardiac, and movement channels in PSG. This allows models to be trained on large-scale PSG datasets that are diverse in population. For the migration challenge, we employ data augmentation techniques, significantly enhancing generalizability. We conducted

*equal contribution

Authors' addresses: Shuzhen Li, Tsinghua University, 30 Shuangqing Rd, Beijing, China, thulisz21@gmail.com; Yuxin Chen, Tsinghua University, 30 Shuangqing Rd, Beijing, China, yuxin-ch21@mails.tsinghua.edu.cn; Xuesong Chen, Beijing Wuji Medical Technology Co., Ltd., 1 Zhongguancun Rd (E), Beijing, China, chenxuesong1128@163.com; Ruiyang Gao, Beijing Wuji Medical Technology Co., Ltd., 1 Zhongguancun Rd (E), Beijing, China, railgun@gaoruiyang.cn; Yupeng Zhang, Beijing Wuji Medical Technology Co., Ltd., 1 Zhongguancun Rd (E), Beijing, China, bookshu1265@gmail.com; Chao Yu, Tsinghua University, 30 Shuangqing Rd, Beijing, China, zoeyuchao@gmail.com; Yunfei Li, Tsinghua University, 30 Shuangqing Rd, Beijing, China, yunfeili.cloud@gmail.com; Ziyi Ye, Tsinghua University, 30 Shuangqing Rd, Beijing, China, yeziyi1998@gmail.com; Weijun Huang, Shanghai Jiao Tong University School of Medicine, 227 Chongqing Rd (S), Shanghai, China, hellohuangwj@126.com; Hongliang Yi, Shanghai Jiao Tong University School of Medicine, 227 Chongqing Rd (S), Shanghai, China, yihongl@126.com; Yue Leng, University of California, San Francisco, 675 18th St, San Francisco, California, United States, 94107, yue.leng@ucsf.edu; Yi Wu, Tsinghua University, 30 Shuangqing Rd, Beijing, China, jxwuyi@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, or post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2474-9567/2024/12-ART185

<https://doi.org/10.1145/3699743>

extensive training and testing on large datasets (12393 records from 9637 different subjects), achieving an accuracy of 0.803 and a Cohen’s Kappa of 0.718. ZeroSleepNet was also deployed in real prototype (monitoring pads) and tested in actual hospital settings (265 users), demonstrating an accuracy of 0.697 and a Cohen’s Kappa of 0.589. To the best of our knowledge, this work represents the first known reliable BCG-based sleep staging effort and marks a significant step towards in-home health monitoring.

CCS Concepts: • **Computing methodologies** → **Neural networks; Supervised learning by classification**; • **Applied computing** → **Consumer health**; • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Sleep Staging, Ballistocardiography, Health Monitoring

ACM Reference Format:

Shuzhen Li, Yuxin Chen, Xuesong Chen, Ruiyang Gao, Yupeng Zhang, Chao Yu, Yunfei Li, Ziyi Ye, Weijun Huang, Hongliang Yi, Yue Leng, and Yi Wu. 2024. SleepNetZero: Zero-Burden Zero-Shot Reliable Sleep Staging With Neural Networks Based on Ballistocardiograms. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 185 (December 2024), 25 pages. <https://doi.org/10.1145/3699743>

1 INTRODUCTION

Sleep is a vital physiological process for humans to maintain health and homeostasis [62]. The consequences of sleep disorders span from daytime sleepiness to increased cardiovascular disease and stroke risk [51]. The current gold standard of sleep disorder diagnosis is the *polysomnogram* (PSG) [51]. PSGs are collections of various physiological signals, including *electroencephalograms* (EEGs), *electrocardiograms* (ECGs), *electrooculograms* (EOGs), *electromyograms* (EMGs), *photoplethysmography* (PPG), and respiratory signals [2].

As a vital aspect of sleep analysis, sleep staging, also known as sleep stage classification, is pivotal for sleep disorder diagnosis [9, 55]. Namely, based on EEG changes, overnight sleep is divided into wake (W), rapid eye movement (REM, R) sleep related to dreaming, and three non-REM sleep stages (N1 – N3 from light to deep), according to the AASM standards [5]. Sleep staging helps detect potential health problems, including sleep apnea, stroke, diabetes, brain injury, Parkinson’s disease, depression, and Alzheimer’s disease [6, 37, 42, 50, 56, 59, 74]. Traditionally, sleep staging is performed by experienced technicians reading EEGs, assisted by EOGs and EMGs [5]. As the process is labor-intensive, automated sleep staging with machine learning methods has been a research focus.

While state-of-the-art sleep staging methods have reached human-level performance [34, 44], they rely on EEGs, whose acquisition requires expensive specialized sensing equipment and is limited in hospital or laboratory environments. This brings the following drawbacks. First, the extra burden of sleep diagnosis could impede people with sleep disorders from getting diagnosis and treatment, as evidenced by Kapur et al. [27], not to mention that longitudinal monitoring is impractical. Furthermore, the subject has to stay in an unfamiliar hospital or laboratory setting, with uncomfortable electrodes stuck to the skin. In such environments, the “first-night effect” alters natural sleep patterns, thereby impacting the reliability of the results [14, 21]. Finally, setting up the monitoring environment is burdensome for physicians, patients, researchers, and study participants. Though some EEG-based devices like Dreem [13] and MUSE [29] can monitor sleep stage changes overnight outside a laboratory setting, they also burden users to sleep as the electrodes have to be attached tightly to their heads.

In contrast, recent studies have utilized household or wearable sensors to circumvent environmental constraints, such as *photoplethysmograms* (PPGs) [28, 30, 47] and *ballistocardiograms* (BCGs) [39, 40, 48, 69, 71]. In particular, BCGs, which “capture the ballistic forces of the heart caused by the sudden ejection of blood into the great vessels with each heartbeat, breathing, and body movement”, can be acquired with piezoelectric sensors [54], which can be deployed easily at home and cause no burden for consumers.

While current BCG-based sleep staging methods show high accuracy, their reliability faces the following challenges. First, annotated BCG data are hard to acquire, as taking BCGs is uncommon in sleep monitoring.

Although there have been large-scale open-source sleep datasets such as SHHS [46, 75], MESA [11, 75], and HSP [68] that contain thousands of subjects, all of them do not contain BCG data. Therefore, existing methods are trained and tested upon restricted datasets (≤ 25 subjects), bringing concerns about overfitting. Second, existing works have revealed that sleep patterns significantly shift across ages [36], and can be affected by diseases [1, 22, 59]. Therefore, evaluating the methods on a diverse population is crucial for reliability. Finally, while Wu et al. [69] transferred existing models to BCGs in a zero-shot manner, the sleep staging performance evaluated on BCGs was unrevealed. Such methods may be problematic, as BCG sensors differ in principle from medical sensors. In a household scenario, the sensor may receive multiple disturbances, resulting in a lower-quality signal than PSGs.

To tackle the challenges above, this paper introduces SleepNetZero, a zero-shot framework designed for zero-burden signal generalization. Specifically, we leverage extensive publicly available PSG datasets to address the scarcity of BCG samples. Given the BCG encapsulates the collective effects of heartbeat, breathing, and body movement, we can discern and correlate these components with specific PSG channels. The wealth of PSG data at our disposal facilitates the application of advanced deep learning methodologies, thereby enhancing performance and reliability.

To mitigate the sensor gap, we introduce data augmentation methods that are novel to bio-signals. Our approach augments the input BCG components through amplification and speed perturbation, tailoring the model to accommodate the disturbed signals.

We have conducted comprehensive experiments to demonstrate the superiority and reliability of SleepNetZero. Our model has been thoroughly validated using extensive datasets with 12393 records from 9637 different subjects. Furthermore, our model has been integrated into a prototype – a non-intrusive monitoring pad equipped with a BCG sensor – which has seen widespread adoption in domestic and hospital settings. SleepNetZero has delivered outstanding results on parallel datasets, including PSG and BCG data, collected from 265 users in a hospital setting. These results substantiate the efficacy of our methodology. To our knowledge, this represents the first application of a product utilizing BCG signals for high-precision sleep staging.

To summarize, we make the following major contributions:

- **BCG Component Extraction:** We propose to extract three BCG components to leverage vast PSG datasets, scaling the dataset by two orders of magnitude. We adopt sophisticated inter-beat interval (IBI) extraction pipelines, indicating the heartbeat component. We find the respiration component through certain frequency bands. We design a novel indicator based on the proposed dynamic energy algorithm, finding the body movement component aligned between EEGs and BCGs. This approach addresses the challenge of limited BCG sample availability.
- **Generalization Enhancement:** To improve the generalizability of SleepNetZero across diverse real-life disturbances, we introduce data augmentation techniques. Specifically, we applied amplification and speed perturbation to the input BCG components to accommodate signals with varying characteristics.
- **Neural Network Framework:** We formulate the problem by a sequence labeling task and devise a deep learning model upon the extracted components. The model comprises a ResNet-based feature extractor, a Transformer encoder, and a linear classifier. Finally, a neural network framework named SleepNetZero is constructed, which offers theoretical innovation and demonstrates superior performance in empirical testing.
- **Experiments and Prototype Validation:** We conduct comprehensive experiments on the proposed method, achieving a Kappa of 0.718 and an Accuracy of 0.803, which showcases the reliability of SleepNetZero. We also verify the SleepNetZero through a real deployed sleep monitoring prototype, which is zero-burden for the subject. The promising results achieving a Kappa of 0.697 and an Accuracy of 0.589, enhance the confidence in the practicability of SleepNetZero.

The rest of our paper is organized as follows. Section 2 summarizes the related works. Section 3 formulates the task. Section 4 deeply analyzes the challenges of the BCG-based sleep staging task and introduces the basic ideas. Section 5 presents the details of SleepNetZero framework design. Section 6 describes the experimental setup, and Section 7 displays the results. We discuss the limitations and future works in Section 8, and finally, conclude this paper in Section 9.

2 RELATED WORK

Sleep staging can be conducted using various types of devices, each bringing its own set of advantages and limitations. Specifically, sleep staging can be categorized based on the device type into three main groups: medical sensors, wearable devices, and zero-burden devices. Each category offers distinct methodologies for collecting data and levels of intrusion into the user's life, thus influencing both the accuracy of the sleep stage assessment and its practicality for continuous monitoring.

2.1 Sleep Staging with Medical Sensors

With advancing deep learning techniques, EEG-based sleep staging based has progressed rapidly [15, 16, 26, 44, 45, 64, 77]. Phan et al. [44] proposed a hierarchical Transformer [66] architecture based on single-channel EEGs, resulting in state-of-the-art performance as shown by the experimental results over the SHHS dataset [46, 75] containing thousands of participants, which has achieved the human level [34].

While EEG-based methods have seen great success, EEGs are not recorded in common medical settings, which limits the usage of these methods. Alternative signals are explored for sleep staging, including ECGs [63, 76], heart rate [17, 19, 32, 33, 41, 57, 60, 73], respiratory effort [60], respiration rate [19, 32, 41], body movement [17, 33, 41, 73], and pulse oximetry [10]. While these signals are increasingly available for in-home sleep monitoring, household devices usually differ in principle from medical sensors, and hence signals taken by household devices may not have the high precision as specialized medical sensors. Therefore, the applicability to household scenarios of the mentioned methods remains unexplored.

2.2 Sleep Staging with Wearable Devices

With the prevalence of wearable devices, especially smartwatches, recent works have focused on in-home sleep monitoring through wearable sensors, especially wrist-worn ones. Typical wrist-worn sensors include *photoplethysmography* (PPG) and actigraphy (also known as accelerometry). Methods based on these sensors have been making rapid progress [4, 28, 30, 47, 72]. However, tightly worn sensors during sleep may cause discomfort to the skin, while loosely worn sensors may be affected by large artifacts. This impedes the practicality and reliability of the methods based on wearable devices.

2.3 Sleep Staging with Zero-Burden Devices

Unlike wearable devices, zero-burden devices are placed in the bedroom, contactless with the subject, further improving the comfort for consumers. Sleep staging through a variety of zero-burden devices has been studied. Zhao et al. [78] leveraged radio frequency signals and explored an adversarial architecture of neural networks. Hong et al. [23] leveraged nocturnal sounds with a neural network trained on the cleaner PSG audio dataset and tested on the noisier smartphone audio dataset. van Meulen et al. [65] utilized cameras above the bed for heartbeat extract, which the authors call "remote PPGs".

However, the above methods are remote, whose signals can be easily disturbed by the environment, hindering the robustness. Conversely, physical connections are more reliable. The ballistic force of vessels can be conducted through the pillow and captured by the piezoelectric materials below, validating the superiority of BCGs as it is physical yet zero-burden.

There have been studies on BCG-based sleep staging [39, 40, 48, 71]. Due to the lack of data, Migliorini et al. [39], Mitsukura et al. [40], Yi et al. [71] utilized traditional machine learning methods. Rao et al. [48] proposed a deep learning method that addressed the data shortage by pre-training with a mixture of raw ECGs, PPGs, and BCGs. All these works conducted experiments on no more than 25 subjects.

While pre-training could potentially leverage the PSG dataset, the mechanism is quite opaque. Moreover, transfer learning splits the BCG dataset for training and validation, reducing the test dataset. In contrast, our proposed method leverages the vast PSG dataset by component extraction, which requires no transfer learning. As a result, the precious BCG dataset could be held out entirely for testing, significantly improving the reliability of the experimental results.

Resembling our method, Wu et al. [69] did propose a method that constructs ECGs from BCG-based features, but the experiments on sleep staging were not conducted upon annotated BCG datasets. Therefore, as far as we know, we are the first to perform the component extraction method on BCG-based sleep staging.

3 PRELIMINARIES OF SLEEP STAGING

Sleep staging, as defined by the American Academy of Sleep Medicine (AASM) standards [5], is a critical process in sleep analysis that involves segmenting the entire duration of overnight sleep into distinct intervals. These intervals are then classified into one of five sleep stages by trained technicians. This categorization is fundamental for understanding sleep patterns and diagnosing sleep disorders.

3.1 Segmentation and Labeling

The sleep record for an individual night is divided into T successive segments, where each segment corresponds to a 30-second interval. This 30-second duration is chosen based on AASM guidelines, which recommend it as the standard epoch length for clinical sleep stage scoring due to its effectiveness in capturing sufficient cycles of sleep activity while maintaining manageable data granularity.

For every segment t , where $t = 1, 2, \dots, T$, it represents the time from $30(t - 1)$ seconds to $30t$ seconds of the overnight recording. The corresponding sleep stage for each of these segments is denoted by y_t , which can take one of the five possible categorical values: W, N1, N2, N3, and R. Their meanings are described in Appendix A.

3.2 Data Handling for Short Segments

In cases where the remaining portion of the sleep record does not neatly fit into a 30-second segment — typically the final few seconds of the recording — this remainder is excluded from the analysis. This practice ensures that every segment analyzed maintains uniform length and data integrity.

4 CHALLENGES AND BASIC IDEA OF SLEEPNETZERO

In this section, we highlight the aforementioned challenges in Section. 1 and present basic ideas to address the challenges.

4.1 Challenges

To the end of achieving accurate sleep staging, several challenges need to be solved as follows.

- (1) **Diversity of Population:** Although recent BCG-based sleep staging methods [39, 40, 48, 71] have reported high accuracy, their main defects are the reliability across populations. Some works trained and tested the model on data from the same subject [71], which leads to dependence on individual features. Others used the leave-one-subject-out method [39, 40, 48], where the test data is directly used for hyperparameter tuning. Possible flaws of such methods include overfitting the test population. Therefore, their performance relies on restricted datasets (no more than 25 subjects). As sleep pattern changes across individual features

like age or health status, the previous methods may suffer from performance drops when facing unseen data.

- (2) **Generalization Across Sensors:** In principle, each PSG channel is taken by a specialized medical sensor, while all BCG components share a household sensor. The sensor gap implies that the components extracted from BCGs may have a lower quality than those from PSGs. As a result, a model trained with high-quality PSG-derived data may fail to generalize to low-quality BCG-derived data. Therefore, the second difficulty arises: we must handle the quality gap between PSGs and BCGs, which is unprecedented.
- (3) **Multimodal Bio-Signal Modeling:** Different BCG components are in different modalities. Heartbeats are sparse yet uniformly spaced, breath is continuous, and body movement is a sequence of 0/1 indicators. The modeling and fusion of such modalities, especially body movement, lack reference.

We present our basic idea to the aforementioned challenges in the following subsections.

4.2 Basic Idea

4.2.1 Leveraging PSG Datasets. To address the first challenge, we noticed the large-scale open-source PSG datasets, such as SHHS [46, 75], which contains about 5.8k subjects. The reliability inherently comes from the patients' diversity of ages and diseases.

However, formerly people tend to consider PSG and BCG as separate signals. There are very few works using PSG data to assist in BCG-based sleep staging. Although Wu et al. [69] proposed a pipeline leveraging ECGs from PSG datasets, its performance on annotated BCG data remains unexamined. We are the first to utilize PSG data for BCG-based sleep staging with convincing results.

As mentioned in Sadek et al. [54], BCGs represent heartbeat, breath, and body movement. Therefore, we aim to extract these components from BCGs and PSGs, bridging both modalities. For the heartbeat, as suggested by the recent practice [19, 33, 41], we extract inter-beat intervals (IBIs) and then interpolate them into a continuously varying signal. The interpolated IBI signal represents the heart rate variability, and the identification pipelines are sophisticated [8, 35, 38]. We propose new methods for the breath and movement components.

4.2.2 Generalization Across Sensors. The basic idea to handle the difficulty is to train the model with lower-quality data. Although the variation from PSGs to the BCG components is unstudied, we may hypothesize some typical perturbation patterns. As bio-signals are 1-dimensional real-valued signals, we incorporate random amplification and speed perturbation referring to the audio practice.

4.2.3 Multimodal Bio-Signal Modeling. The key insight of modeling the multimodal bio-signals comes from the fact that the body movement indicator has only 0/1 values and plays the role of mask. Therefore, we will first introduce the framework of modeling the heartbeat and breath waves, and then adhere the body movement to them.

The common practice to model a single waveform is a hierarchical solution [3, 24]. First, the high-resolution input is encoded by a convolutional feature extractor, resulting in a local representation for each window. Consequently, the representations are sent into a Transformer encoder [66], capturing contextual information.

Next, we fuse the two modalities. Besides the individual representation of each modality, their coupling should also be captured. Therefore, we use three distinct feature extractors. The first encodes the heartbeat, the second encodes the breath, and the third encodes both.

Finally, we add the body movement to each feature extractor. Namely, the first feature extractor encodes the heartbeat and body movement, the second encodes the breath and body movement, and the third encodes the three components.

The extracted features are concatenated to form the local representation. The local representations are then sent into the Transformer encoder, giving the contextual information. Finally, the contextual information is sent into a multi-layer perceptron (MLP), whose output is the probability over the classes.

In conclusion, we use a neural network to model the problem. The neural network comprises three feature extractors, a Transformer encoder [66], and a MLP classifier.

5 SLEEPNETZERO FRAMEWORK DESIGN

This section presents the framework of SleepNetZero. First, an overview of the framework of SleepNetZero is given. Then, each module of the framework is described in the following subsections.

5.1 Overview of SleepNetZero

As shown in Figure 1, the proposed framework consists of three main stages: component extraction, generalization, and the neural network. In the component extraction stage, the framework processes signals from BCG signals to extract meaningful features related to heartbeats, breathing, and body movements. Each type of signal is treated with specialized algorithms to ensure the accuracy and relevance of the data for sleep staging. Following extraction, the generalization stage prepares the data for deep learning applications with two innovative data augmentation techniques. The final stage employs a neural network, specifically ResNet architectures and Transformer encoders, to classify sleep stages based on the processed signals. This multi-layered approach leverages deep learning techniques to interpret the intricate patterns in the data, ultimately classifying sleep stages with high precision.

The specific settings and methodologies applied in the model training process are discussed in subsequent sections, detailing the technical strategies that enhance the framework's performance.

5.2 Component Extraction Module

We extract three components representing heartbeat, breath, and body movement. The algorithms are listed below.

Heartbeat Component. Following the recent practice, we represent heartbeats by inter-beat intervals (IBIs). The IBI representation enables a lower sampling frequency and offers a convenient artifact removal method. For ECGs, the heartbeats are identified as R peaks, and the IBI is known as the RR interval (RRI). For BCGs, the heartbeats are identified as J peaks, and the IBI is known as the JJ interval (JJI).

- Let n denote the number of heartbeat intervals.
- Let P_i denote the end time of the i^{th} heartbeat interval.
- Let T_i denote the length of the i^{th} heartbeat interval.

From BCGs, we adopt the method proposed in [8] to extract (P_i, T_i) . From PSGs, we clean the ECG and find the R peaks P_0, P_1, \dots, P_n using NeuroKit 2 [35], and then assign $T_i = P_i - P_{i-1}$.

To prevent outliers, while preserving the end timestamps P_i , we replace the intervals T_i with normal-to-normal intervals using the `get_nn_intervals` function in the package HRVAnalysis [38] with default parameters, i.e., $T_i \leftarrow \text{get_nn_intervals}(T_i)$. The function removes physiologically implausible or ectopic T_i 's. The removed values are linearly interpolated between the remaining ones. Removed T_i 's at the beginning or the end of the sequence are dropped, and the corresponding P_i 's are dropped, too. Note that P_i does not necessarily equal $P_{i-1} + T_i$ after the correction.

The heartbeat signal fed into the model is linearly interpolated at a sampling frequency of 4Hz between the anchor points (P_i, T_i) . The signal before P_1 is padded by the constant T_1 , and the signal after P_n is padded by the constant T_n .

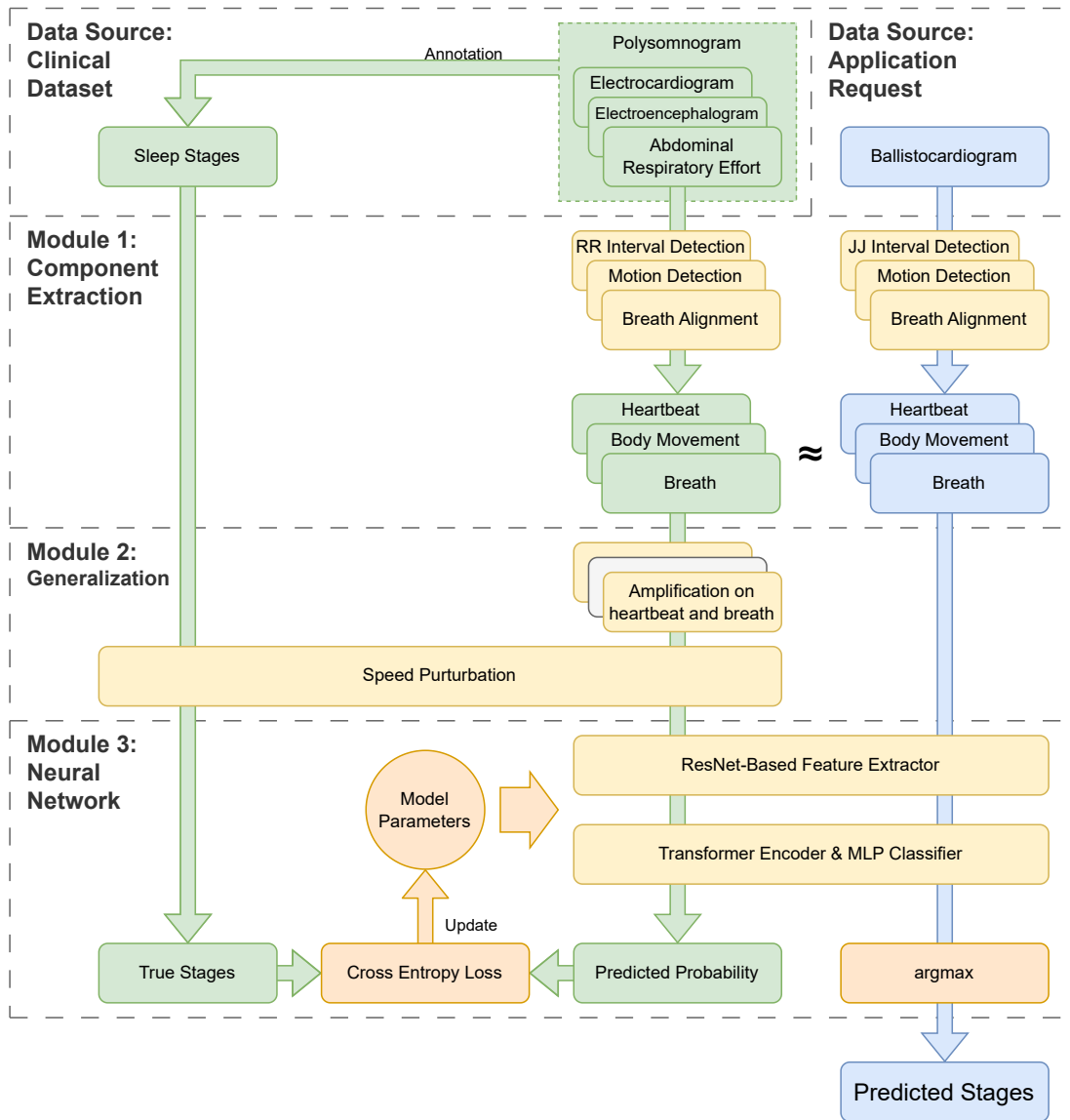


Fig. 1. The overview of the proposed framework. Training data from clinical datasets and test data from application requests are aligned through the component extraction module, which comprises three extractors. For training data, the extracted components are augmented by the generalization module, which consists of amplification and speed perturbation. Finally, these components are fed into a neural network for training and inference.

Breath Component. The breath signal resides in the BCG and the abdominal respiratory effort (ABD effort) in the PSG. The model input is extracted through our proposed four-stage pipeline: filtering, resampling, integration, and normalization.

First, a $\frac{1}{10} \sim \frac{1}{3}$ Hz band-pass third-order Bessel filter [61] is applied to the raw BCG signal and the ABD effort. The output is then resampled to a sample frequency of 4Hz. Next, we integrate the BCG-derived breath signal over time. The integration is not applied to the PSG-derived breath. Finally, the final signal is normalized to z-scores.

The integration remedies the phase gap between BCG signals and ABD effort, since the BCG signal represents the breathing flux, while the ABD effort reflects the volume of the breathed air. The normalization removes the magnitude difference, which is because the two types of sensors differ in principle.

Body Movement Component. Body movements appear as significant disturbances to EEGs [5] and BCGs. As BCG and EEG data is sampled at a rate of 500Hz, given the frequent changes in the BCG signal during body movements, a 2-second window is appropriate to capture these rapid variations. Within this 2-second window, we use the peak-to-peak amplitude as a feature to characterize the signal changes.

For each window, we establish a 30-second dynamic baseline of statistical metrics, including the mean (μ) and standard deviation (σ), calculated from the windowed data. The baseline period is divided into 15 2-second windows. Each of these 2-second windows provides a feature value, forming a sequence of 15 values. We then compute the mean and variance of this sequence. This baseline helps in adapting to the intrinsic variability of the EEG or BCG signals over time. A threshold is then applied to detect significant deviations from this baseline, indicative of potential body movements. Specifically, any window where the peak-to-peak amplitude exceeds $\mu + 5\sigma$ is flagged as containing body movement. This threshold is chosen based on the distribution characteristics of the signal and is intended to ensure that only substantial deviations, which are likely due to movements rather than physiological fluctuations, are considered. The use of a multiplier (5) of the standard deviation is guided by statistical norms where such a level typically signifies a departure from normal variance, assuming a Gaussian distribution.

5.3 Generalization Module

The Generalization Module is designed to enhance the robustness of our model by addressing discrepancies caused by sensor gaps. We introduce two innovative data augmentation techniques specifically tailored for bio-signals. These techniques are random amplification and speed perturbation, each targeting different aspects of signal variation. Namely, the random amplification stretches the signal in the magnitude direction, and the speed perturbation stretches the signal in the time direction.

5.3.1 Random Amplification. The first technique, random amplification, aims to simulate variations in signal intensity that might occur due to sensor sensitivity or user differences. This technique adjusts the amplitude of the heartbeat and breath signals independently, making the model more resilient to fluctuations in signal strength. The heartbeat signal and the breath signal are amplified independently. For the signal \mathbf{x} , we pick a scalar $\alpha \sim \text{Uniform}(0.9, 1.1)$, and the model input is replaced by $\alpha\mathbf{x}$. This process ensures that the model can handle slight variations in amplitude without compromising the accuracy of stage classification.

5.3.2 Speed Perturbation. The second technique, speed perturbation, addresses variations in the temporal domain, mimicking scenarios where the signal sampling speed might vary due to technical issues or user differences. This technique uniformly stretches or compresses the three input components of the model along with their corresponding ground truth stages.

The speed perturbation affects the three components and the ground truth stages by a common factor $\beta \sim \text{Uniform}(0.75, 1.25)$. The three input components are perturbed adopting the implementation from torchaudio [25,

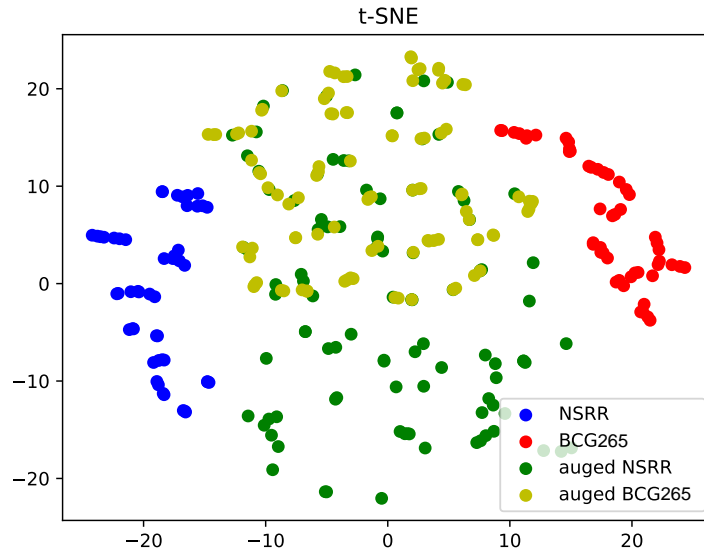


Fig. 2. The t-SNE visualization of the proposed data augmentation technique in the spectrum space. The original distributions of NSRR (blue) and BCG265 (red) data are distinct and separate. They cluster closer to each other (green, yellow) after data augmentation, illustrating that our data augmentation effectively eliminates the disparity between the two datasets.

70]. Once the signal is adjusted, suppose the signal length is L' seconds, then the ground truth should have a length of $T' = \lfloor L'/30 \rfloor$, where 30s is the annotation segment length. The stages y' are transformed by a linear spacing sampling:

$$t(i) = \left\lfloor \frac{T(i-0.5)}{T'} \right\rfloor + 1, \quad y'_i = y_{t(i)}, \quad i = 1, 2, \dots, T'.$$

Intuitively, for each 30-second segment, we find the center's original position and sample the original stage. After that, the signals are truncated to a length of $30T'$ seconds, maintaining a standardized input size for the model.

5.3.3 Visualization of the Module. We demonstrate the effectiveness of the proposed data augmentation technique with Figure 2. In the demonstration, we randomly pick 30 records from NSRR and BCG265 respectively. The second hour of the record is cropped. Each crop is augmented with three different random seeds. For each crop, we concatenate the Welch power spectrum densities (PSDs) of the heartbeat and breath components to form a feature vector. Then, we visualize the 2D projection with t-SNE of the vectors from various datasets. The data from NSRR and BCG265, originally distributed in distinct and separate areas, cluster closer to each other after data augmentation. This provides a piece of evidence that our data augmentation effectively eliminates the disparity between the two datasets.

5.4 Neural Network Module

In the proposed model, we address the task of sleep staging as a sequence labeling problem, utilizing the rich temporal dynamics of physiological signals.

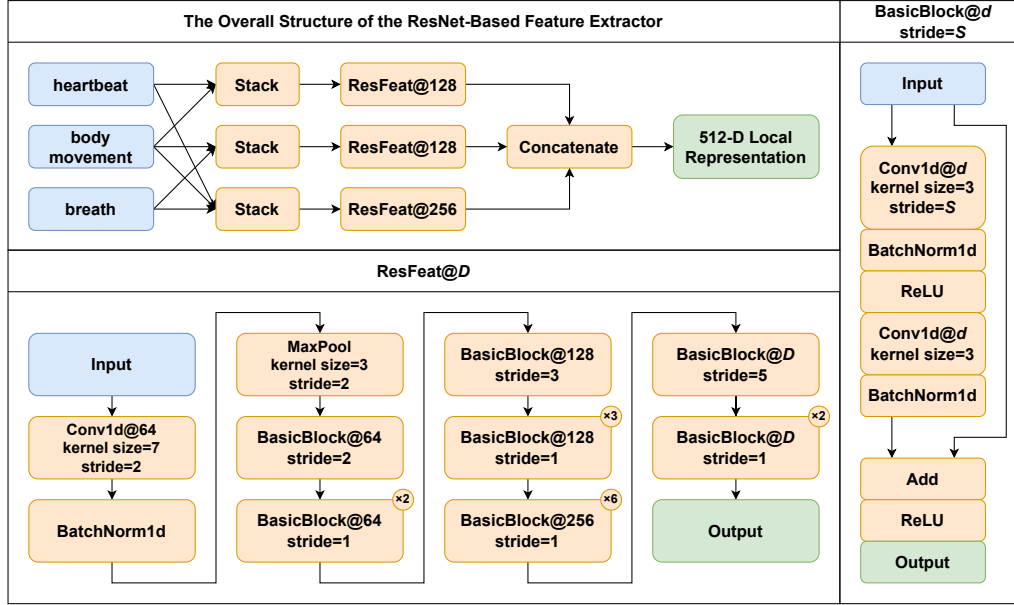


Fig. 3. This is the architecture of the proposed three feature extractors, key components that extract high-dimensional features and fuse different modalities. The upper left part shows the overall architecture. The three components – heartbeat, body movement, and breath – are taken as inputs. Three ResNet-based feature extractors, denoted by ResFeat in the figure, produce high-dimensional representations for the three distinct groupings of components. The representations provided by the three extractors are concatenated together. The lower left part shows the detailed architecture of the ResFeat module, which is the main body of the feature extractor, denoted f_1, f_2, f_3 in the main text. The right part shows the detailed architecture of the BasicBlock module, which is part of the ResFeat module.

5.4.1 Data Representation. The input to our model consists of three distinct physiological signal sequences, denoted as x_1, x_2, x_3 . x_1, x_2 are continuous signals: x_1 is the IBI wave representing the heartbeat, and x_2 is the breath wave. x_3 is the 0/1 sequence indicating body movements. To standardize the input, we select a unified sampling rate of 4 Hz for each sequence, resulting in sequence lengths of $120T$ for a total recording time of $30T$ seconds.

5.4.2 Feature Extraction. To capture the unique characteristics of each signal component effectively, we employ three independent feature extractors based on the ResNet architecture [20], as shown in Figure 3. Each extractor, denoted f_1, f_2, f_3 , is trained to transform its respective input sequence into a high-dimensional feature space. For each t^{th} classification segment, the features are aggregated as follows:

$$z_t = \text{concatenate} \left(f_1(\text{stack}(x_1^{(t)}, x_3^{(t)})), f_2(\text{stack}(x_2^{(t)}, x_3^{(t)})), f_3(\text{stack}(x_1^{(t)}, x_2^{(t)}, x_3^{(t)})) \right), \quad (1)$$

where $x_i^{(t)}$ denotes the perception field of the i^{th} sequence at time t . This design allows the network to learn from the local context of each signal, enhancing the model's ability to discern subtle patterns indicative of different sleep stages.

5.4.3 Context Encoding and Classification. Following feature extraction, the aggregated feature vectors \mathbf{z}_t are processed by a Transformer encoder [66], which incorporates a sinusoidal positional encoding [66] to maintain the temporal context of the sequence:

$$[\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_T] = \text{encoder}([\mathbf{z}_1 + \mathbf{e}_1, \mathbf{z}_2 + \mathbf{e}_2, \dots, \mathbf{z}_T + \mathbf{e}_T]; \theta), \quad (2)$$

where \mathbf{e}_t represents the sinusoidal positional encoding for the t^{th} segment. This step is crucial as it allows the model to leverage both the local features extracted by ResNet and the global dependencies between segments captured by the Transformer architecture.

The context vectors \mathbf{c}_t output from the Transformer encoder are then fed into a 2-layer MLP classifier, which computes the probability distribution over sleep stages for each segment using a softmax function:

$$\mathbf{p}_t = \text{softmax}(\mathbf{W}_2 \max(\mathbf{0}, \mathbf{W}_1 \mathbf{c}_t + \mathbf{b}_1) + \mathbf{b}_2), \quad (3)$$

where $\mathbf{p}_{t,c}$ indicates the predicted probability of the t^{th} segment being in the c^{th} sleep stage. The final prediction \hat{y}_t for each segment is determined by selecting the class with the highest probability:

$$\hat{y}_t = \arg \max_c \mathbf{p}_{t,c}. \quad (4)$$

The Transformer encoder used has 6 layers. Each layer is 512-dimensional, with a 2048-dimensional feedforward neural network. The hidden layer of the MLP classifier is 64-dimensional.

5.4.4 Training. During the training phase, we optimize the parameters of the model by minimizing the cross-entropy loss between the predicted probabilities and the true class labels:

$$L = - \sum_{t=1}^T \log \mathbf{p}_{t,y_t}, \quad (5)$$

where y_t denotes the true class label for the t^{th} segment. This loss function encourages the model to accurately predict the correct sleep stage, thus improving its overall performance on unseen data.

6 EXPERIMENTAL SETUP

This section outlines the specific configurations of our experiments, including the datasets used, the evaluation metrics applied, and the details concerning the training process.

6.1 Datasets

The NSRR Dataset. We train and evaluate our method using the combination of the following publicly available PSG datasets: SHHS [46, 75], MrOS [7, 75], MESA [11, 75], CCSHS [53, 75], CFS [49, 75], and HomePAP [52, 75]. We will refer to the combined dataset by NSRR. After cleaning, the NSRR dataset contains 12393 records from 9637 different subjects. We include its age and sex distributions in Appendix B.

Among all the subjects in the NSRR dataset, 10% are held out for testing, 10% are split for validation, and the remaining 80% are used for training the model. The training/validation/test subsets consist of the data from their corresponding subjects. Therefore, recordings from different splits are from distinct subjects. Across experiments, the training/validation/test split does not change.

The ground truth labels are imbalanced. In the NSRR dataset, the distribution of class labels is detailed in Table 1.

Stage	W	N1	N2	N3	R
Percentage	33%	3%	37%	14%	12%

Table 1. Stage distribution of the NSRR dataset.



Fig. 4. The prototype of the monitoring pad with the BCG sensor.



Fig. 5. The monitoring pad and the unencapsulated piezoelectric sensors. The sensors are already placed in their related encapsulation positions.

AHI Group	Criterion	Percentage
Normal	$AHI < 5$	18.4%
Mild Sleep Apnea	$5 \leq AHI < 15$	20.5%
Moderate Sleep Apnea	$15 \leq AHI < 30$	18.0%
Severe Sleep Apnea	$AHI \geq 30$	43.1%

Table 2. The apnea-hypopnea index (AHI) distribution of the BCG265 dataset. Most (81.6%) of the subjects suffer from sleep apnea, and 43.1% are severe.

Age Group	Percentage
0–20	1.4%
20–40	56.5%
40–60	33.2%
60+	8.8%

Table 3. The age distribution of the BCG265 dataset. Most (89.7%) subjects are 20 to 60 years old. Given that most subjects are sleep apnea patients, this may be a joint effect of prevalence and visiting rate.

Sex	Percentage
Male	65.7%
Female	34.3%

Table 4. The sex distribution of the BCG265 dataset. Males are more than females. This is because most subjects are sleep apnea patients. Male sleep apnea prevalence is higher than female.

The BCG265 Dataset. We also verify our method using a BCG dataset collected by our monitoring pad prototype. As shown in Fig. 4, piezoelectric BCG sensors are encapsulated in the pad and can be placed under the pillow for sleep monitoring. The piezoelectric sensors are encapsulated near the lower edge of the monitoring pad, as shown in Figure 5.

We provide further discussions of the BCG sensor in Appendix C.

We deployed the monitoring pads in a hospital environment, recording BCG signals in synchronization with the PSG. Technicians in the hospital annotate the sleep stages according to the PSG, enabling us to obtain the ground truth sleep stages for the BCG signal.

We collected 265 parallel PSG/BCG records from 265 distinct individuals in the hospital, with one PSG and one BCG record for each patient. The collection has passed the ethical review. We will refer to the dataset by BCG265.

BCG265 consists of individuals with obstructive sleep apnea (OSA)-related diseases, with the specific apnea-hypopnea index (AHI) distribution shown in Table 2. The age and sex distribution is shown in Tables 3 and 4, respectively. We provide the class label distribution of the BCG265 dataset in Appendix B.

The whole BCG265 dataset serves as the test dataset.

6.2 Evaluation Metrics

Sleep staging is a multi-class classification task. Therefore, we evaluate the proposed method by four main metrics, higher is better. Let y_1, y_2, \dots, y_T denote the ground truth sleep stages, and $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_T$ denote the predicted ones. The metrics are formulated as follows.

Accuracy: $\text{ACC} = \frac{1}{T} \sum_{t=1}^T \mathbf{1}_{y_t=\hat{y}_t}$. In the multi-class classification setting, the micro F1 score equals ACC.

Cohen's κ [12]: The coefficient takes the class imbalance into account:

$$\kappa = \frac{T \sum_{t=1}^T \mathbf{1}_{y_t=\hat{y}_t} - \sum_{i=1}^T \sum_{j=1}^T \mathbf{1}_{y_i=\hat{y}_j}}{T^2 - \sum_{i=1}^T \sum_{j=1}^T \mathbf{1}_{y_i=\hat{y}_j}}.$$

Basically, $\kappa = 1$ if y and \hat{y} are in perfect agreement, and $\kappa = 0$ if y and \hat{y} are independent.

Macro F1: $\text{MF1} = \frac{1}{C} \sum_{c=1}^C 2P_c R_c / (P_c + R_c)$, where P_c, R_c are the c^{th} class precision and recall, respectively:

$$P_c = \frac{\sum_{t=1}^T \mathbf{1}_{y_t=\hat{y}_t=c}}{\sum_{t=1}^T \mathbf{1}_{\hat{y}_t=c}}, R_c = \frac{\sum_{t=1}^T \mathbf{1}_{y_t=\hat{y}_t=c}}{\sum_{t=1}^T \mathbf{1}_{y_t=c}}$$

Weighted F1: $\text{WF1} = \sum_{c=1}^C w_c \cdot 2P_c R_c / (P_c + R_c)$, where P_c, R_c are defined above, and $w_c = \frac{1}{T} \sum_{t=1}^T \mathbf{1}_{y_t=c}$.

6.3 Training Details

All experiments are repeated with 10 distinct seeds, implemented in PyTorch [43], and conducted on a single NVIDIA H100 80G GPU. Each model is trained by AdamW [31] for 50 epochs (300 steps each, 15k iterations in total), with a learning rate of 1.1×10^{-4} , a weight decay coefficient of 10^{-5} , and a batch size of 32. The dropout probability of each Transformer encoder layer is set to 0.05.

The model has 29.5M trainable parameters. The data augmentation ensures the model with sufficient input samples to learn effectively from.

7 EXPERIMENTAL RESULTS

This section presents the results of experiments conducted on four aspects: overall performance, generalization validation, baseline comparison, and ablation study.

7.1 Overall Performance

Table 5 presents the comprehensive evaluation results of our SleepNetZero model for sleep staging, assessed on both the NSRR test split and the BCG265 dataset. The results show several important aspects of the model's performance across different metrics and datasets:

Accuracy (ACC): The model achieves an accuracy of 0.803 ± 0.002 on the test split, indicating high overall correctness in its predictions. However, when evaluated on the BCG265 dataset, the accuracy drops to 0.697 ± 0.007 .

Table 5. The results of our SleepNetZero for sleep staging on the NSRR test split and the BCG265 dataset (mean \pm standard). The model is trained on NSRR. In the second column, the model is evaluated on the NSRR test split. In the third column, the model is evaluated on the BCG265 dataset. We observe a significant performance drop on BCG265, which can be considered a consequence of the sensor gap.

Metric	NSRR test split	BCG265
ACC	0.803 ± 0.002	0.699 ± 0.007
κ	0.718 ± 0.003	0.588 ± 0.013
MF1	0.665 ± 0.004	0.612 ± 0.009
WF1	0.792 ± 0.001	0.664 ± 0.007

This decline could be attributed to the sensor gap with the lower quality of the BCG265 dataset, which is unseen in the training data.

Cohen’s Kappa (κ): With a kappa score of 0.718 ± 0.001 on the test split and 0.589 ± 0.011 on the BCG265 dataset, the model demonstrates substantial agreement with the ground truth labels, albeit with a noticeable decrease on the BCG265 dataset. This decrease further confirms the challenges posed by the sensor gap introduced by the BCG265 dataset.

Macro F1 Score (MF1) and Weighted F1 Score (WF1): The MF1 and WF1 scores further elucidate the model’s capabilities in handling imbalanced data. Specifically, the MF1 scores of 0.664 ± 0.005 on the test split and 0.611 ± 0.010 on the BCG265 dataset, indicate room for improvement in achieving consistent performance across all classes.

The decline in performance on the BCG265 dataset can primarily be attributed to the sensor gap between the BCG and PSG signals. It is difficult for a model primarily trained on PSG-aligned data to perform equally effectively on BCG data, as the model’s feature detectors are optimized for signals of higher consistency and quality inherent to PSG. Despite that, our method has made significant strides in addressing the decline resulting from sensor gaps. However, the ongoing performance discrepancies between datasets highlight that matching PSG signal accuracy remains challenging.

Additionally, we report confusion matrices of our model on both the NSRR test split and BCG265. Table 6 shows the confusion matrix on the NSRR test split, while Table 7 shows the confusion matrix BCG265. Each confusion matrix is the average of ten random runs and is normalized along the row direction. Our method yields excellent results on the W, N2, and R stages, yet it struggles to accurately identify the N1 stages in the majority of instances. This issue can be attributed to two principal factors. First, the N1 stage exhibits the lowest annotation consistency among technicians, adding complexity to its identification [34]. Second, the extreme imbalance of classes in the dataset, with only about 3% containing N1 labels, further exacerbates the problem. Therefore, enhancing the model’s performance on unbalanced labels, especially with the challenge posed by the N1 stage, remains a significant task for future work.

7.2 Generalization of SleepNetZero

The generalization capability of SleepNetZero is thoroughly validated across diverse datasets encompassing variations in age, gender, race, and Apnea-Hypopnea Index (AHI). These factors are critical as they can significantly influence the physiological patterns associated with sleep, thereby impacting the accuracy and reliability of sleep stage classification. To show comprehensive results, we use the cross-validation method to obtain the performance on the whole dataset.

	W	N1	N2	N3	R
W	0.900	0.012	0.070	0.002	0.016
N1	0.268	0.122	0.516	0.002	0.093
N2	0.041	0.012	0.844	0.067	0.036
N3	0.008	0.000	0.459	0.527	0.005
R	0.026	0.007	0.093	0.002	0.872

Table 6. The confusion matrix of our SleepNetZero model on the NSRR test split. The model is trained on the NSRR train set. The row labels are the true stages, while the column labels are the predicted stages. The confusion matrix is normalized along the row direction.

	W	N1	N2	N3	R
W	0.920	0.013	0.052	0.003	0.013
N1	0.322	0.105	0.494	0.005	0.075
N2	0.067	0.024	0.820	0.036	0.053
N3	0.022	0.002	0.465	0.496	0.015
R	0.059	0.019	0.127	0.003	0.793

Table 7. The confusion matrix of our SleepNetZero model on the BCG265 dataset. The model is trained on the NSRR train set. The row labels are the true stages, while the column labels are the predicted stages. The confusion matrix is normalized along the row direction.

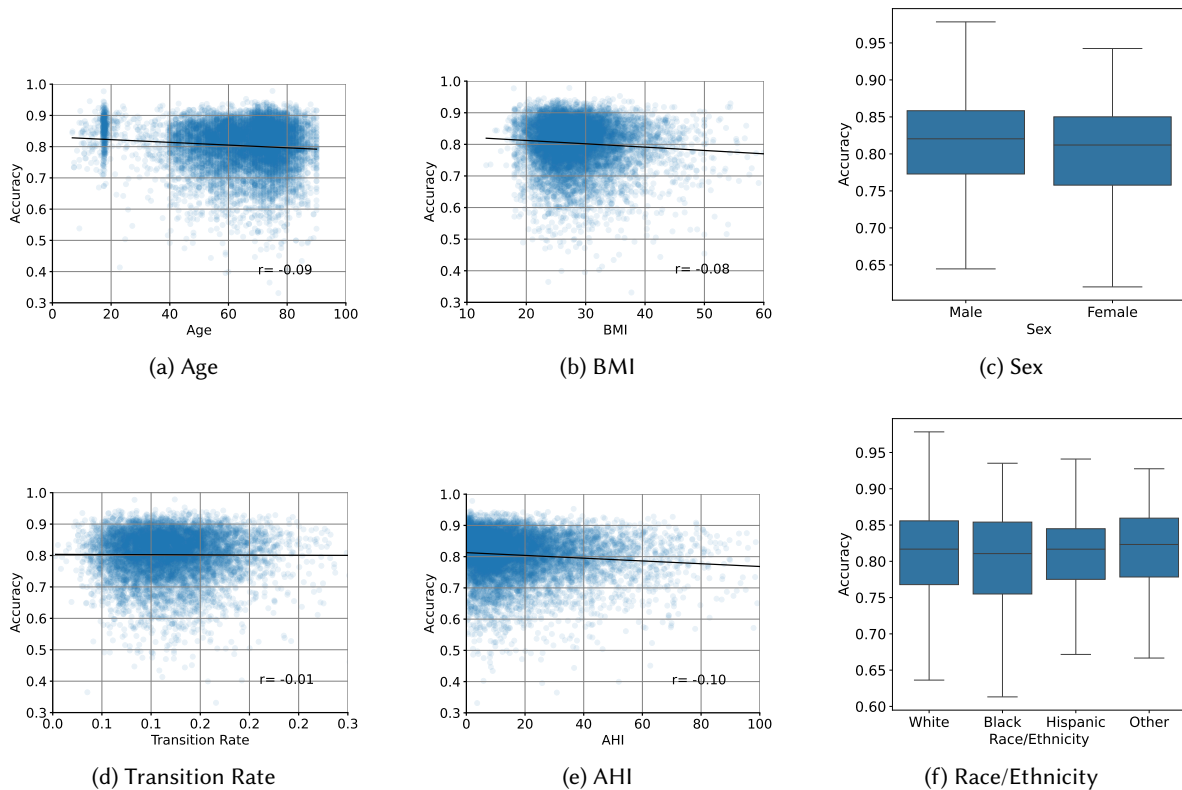


Fig. 6. Accuracy of the testing nights as a function of age, BMI, sex, transition rate, race, and AHI. The transition rate is defined as the ratio of transitions between different sleep stages to the total number of annotations. In sleep staging, an annotation is made every 30 seconds. If two consecutive annotations indicate different sleep stages, this is considered a transition. The transition rate is calculated by dividing the number of these transitions by the total number of 30-second annotations.

7.2.1 Generalization across Age. : The physiological signals associated with sleep stages can vary significantly with age. For instance, older adults may exhibit less pronounced sleep spindles, a feature crucial for identifying specific non-REM sleep stages. As shown in Figure 6a, age exhibits a slight negative but significant correlation with accuracy ($r=-0.09$, $p<0.001$). Although our dataset primarily consists of older adults, SleepNetZero demonstrates robust performance across all age groups, particularly in younger individuals.

7.2.2 Generalization across BMI. : BMI (Body Mass Index) characterizes a person's health in terms of weight and height. People with higher BMI tend to have a higher risk of sleep apnea and less percentage of deep sleep. As shown in Figure 6b, BMI exhibits a slight negative but significant correlation with accuracy ($r=-0.08$, $p<0.001$).

7.2.3 Generalization across Gender. Gender differences in sleep patterns, such as variations in REM sleep duration and sleep cycle architecture, have been documented in sleep research. As shown in Figure 6c, SleepNetZero has a better performance on Male subjects than Female ones (Welch's $T=7.98$, $p<0.001$ [67]).

7.2.4 Generalization across Transition Rate. The transition rate stands for the ratio of sleeping stages, as defined by the human scoring, that is different from the sleeping stage 30 seconds before. We use transition rate to represent the sleep staging stability of a sample. As shown in Figure 6d, we found that it is not a significant determinant of accuracy ($r=-0.01$, $p=0.581$).

7.2.5 Generalization across AHI. The Apnea-Hypopnea Index (AHI) quantifies the severity of sleep apnea based on the number of apneas and hypopneas per hour of sleep. AHI categorizes the severity of sleep apnea into four levels: healthy ($AHI < 5$), mild ($5 \leq AHI < 15$), moderate ($15 \leq AHI < 30$), and severe ($AHI \geq 30$). These levels profoundly influence sleep architecture and quality. As shown in Figure 6e, AHI was significantly correlated with accuracy ($r=-0.10$, $p<0.001$).

7.2.6 Generalization across Race. Racial and ethnic differences can influence sleep architecture and susceptibility to sleep disorders. We counted the accuracy rates on different racial groups including Caucasian (White), African American (Black), Hispanic, and other races (Other). As shown in Figure 6f, we found that race can influence the accuracy significantly (ANOVA [58], $F=6.24$, $p<0.001$).

7.3 Comparison to Related Works

We compare SleepNetZero with the related works in two aspects. First, we compare it with other BCG-based methods. Further, we compare it with the state-of-the-art sleep staging methods based on various signals.

BCG-based Sleep Staging. As shown in Table 8, all the existing BCG-based methods were evaluated on restricted datasets (no more than 25 subjects). Therefore, SleepNetZero is superior in reliability, as it is evaluated on ten times more subjects. Moreover, considering the difference in evaluation methods, SleepNetZero offers comparable performance to less reliable ones (-0.1 ACC compared to 5-class methods, -0.15 κ compared to 3-class methods).

Sleep Staging via Various Signals. As shown in Table 9, when tested on components extracted from PSGs, SleepNetZero offers comparable performance to previous works (-0.02 ACC/ κ relative to the 4-class SOTA PPG method). When tested on BCGs, SleepNetZero suffers from the sensor gap and exhibits a performance drop. However, the sensor gap is brought from the zero-shot experimental setting, and such a setting ensures the reliability of the result. Despite this, SleepNetZero performs the best among all reliable zero-burden methods. Moreover, it exhibits a comparable performance (-0.09 ACC/ -0.11 κ relative to 4-class radio frequency, -0.06 ACC/ κ relative to 4-class wrist PPGs) compared to the less reliable methods.

Table 8. The comparison of various BCG-based sleep staging methods. The results evaluated on less than 100 subjects are considered less reliable, while those evaluated on at least 100 subjects are considered reliable. Some methods adopted 3-class or 4-class evaluation metrics. The 4-class evaluation merges N1 with N2, and the 3-class evaluation merges N1 and N2 with N3, reducing the task difficulty. Therefore, 3-class or 4-class metrics are higher than the 5-class ones we adopt. The “†” marked method trained and tested the model on overlapping subjects, deteriorating the reliability. SleepNetZero is the only reliable BCG-based method. Further, it exhibits comparable performance to less reliable ones.

METHOD	EVALUATION METHOD	#TEST SUBJECTS	ACC	κ	RELIABILITY	ZERO-SHOT
QD-TVAM [39]	3-CLASS	17	0.77	0.55	× Low	× No
MITSUOKURA ET AL. [40]	5-CLASS	25	0.80	N/A	× Low	× No
DEEPSLEEP [48]	4-CLASS	25	0.74	N/A	× Low	× No
YI ET AL. [71]†	3-CLASS	5	0.85	0.74	× Low	× No
SLEEPNETZERO (OURS)	5-CLASS	265	0.70	0.59	✓ HIGH	✓ YES

Table 9. The comparison of sleep staging methods via various signals. The results evaluated on less than 100 subjects are considered less reliable (marked by “↓”), while those evaluated on at least 100 subjects are considered reliable (marked by “↑”). The methods marked by “*” adopted the 4-class evaluation metrics. The 4-class evaluation merges N1 with N2, reducing the task difficulty. Therefore, 4-class metrics are higher than the 5-class ones that we adopt. Our method offers comparable performance on the NSRR test dataset. On the BCG265 dataset, despite the BCG performance drop caused by the sensor gap, BCG sensors cause zero burden, and the zero-shot experimental setting ensures the reliability of the result. Yet, considering the reliability, SleepNetZero performs the best among zero-burden methods.

METHOD	TEST DATASET (#SUBJECTS)	SIGNAL	ACC	κ	BURDEN	ZERO-SHOT
SLEEPTRANSFORMER [44]	SHHS [46, 75] (1737 ↑)	EEG	0.88	0.83	× HIGH	× No
SLEEPPPG-NET [28]*	CFS [49, 75] (320 ↑)	FINGER PPG	0.82	0.74	× HIGH	× No
SLEEPNETZERO (OURS)	NSRR (963 ↑)	PSG	0.80	0.72	× HIGH	× No
WU ET AL. [69]*	MIT-BIH (18 ↓)	ECG	0.76	N/A	× HIGH	× No
RADHA ET AL. [47]*	EINDHOVEN [18] (60 ↓)	WRIST PPG	0.76	0.65	○ FAIR	× No
HONG ET AL. [23]*	PRIVATE (219 ↑)	SOUND	0.70	0.53	✓ ZERO	× No
ZHAO ET AL. [78]*	RF-SLEEP (25 ↓)	RADIO FREQUENCY	0.79	0.70	✓ ZERO	× No
VAN MEULEN ET AL. [65]*	PRIVATE (46 ↓)	REMOTE PPG	0.68	0.49	✓ ZERO	✓ YES
SLEEPNETZERO (OURS)	BCG265 (265 ↑)	BCG	0.70	0.59	✓ ZERO	✓ YES

7.4 Ablation Study

To show the effectiveness of the component extraction module and generalization module, we conduct the ablation study.

7.4.1 Ablation Study of Component Extraction Module. We conduct the ablation study of each component by replacing the component with uniform random noise. As shown in Table 10, SleepNetZero performs the best, with each of the three extracted components playing a crucial role. SleepNetZero without body movement has a more significant drop on BCG265 than on the test dataset, highlighting the critical role of body movement in sleep staging, especially for generalization on BCG. As BCG and PSG sensors handle body movements differently, successful modeling of body movements can greatly help generalization from PSG data to BCG data. SleepNetZero without heartbeat shows a moderate decline on both the test dataset and BCG265, underscoring the effectiveness of heartbeat in the sleep staging task. SleepNetZero without breath exhibits the greatest decline on both the test dataset and BCG265, confirming the vital importance of this channel for sleep staging. Combining the above

Table 10. Ablation study: Kappa and Accuracy on the NSRR test split and BCG265. The models are trained with the NSRR train split. Columns 2–3 are evaluated on the NSRR test split, while columns 4–5 are evaluated on BCG265. The first row evaluates the full SleepNetZero, while the other rows evaluate SleepNetZero without one of the components. We can see that each component here is crucial for the good performance of SleepNetZero, especially for zero-shot generalization on BCG265.

METHOD	NSRR KAPPA	NSRR ACC	BCG265 KAPPA	BCG265 ACC
SLEEPNETZERO	0.718 ± 0.003	0.803 ± 0.002	0.588 ± 0.013	0.699 ± 0.007
– W/O BODY MOVEMENT	0.704 ± 0.002	0.795 ± 0.001	0.540 ± 0.012	0.663 ± 0.010
– W/O HEARTBEAT	0.689 ± 0.004	0.786 ± 0.002	0.542 ± 0.015	0.666 ± 0.009
– W/O BREATH	0.668 ± 0.002	0.771 ± 0.001	0.496 ± 0.006	0.632 ± 0.006
– W/O RANDOM AMPLIFICATION	0.711 ± 0.001	0.799 ± 0.001	0.554 ± 0.008	0.671 ± 0.003
– W/O SPEED PERTURBATION	0.700 ± 0.001	0.791 ± 0.001	0.533 ± 0.010	0.657 ± 0.005

results, we can get the results that all three channels are intrinsically linked, and our innovative application of body movement signaling is well suited to help improve generalizability.

7.4.2 Ablation Study of Generalization Module. As shown in Table 10, removing the random amplification or speed perturbation leads to a significant drop in performance compared to the full model, especially on BCG265. This indicates that both elements are crucial for sleep staging and cross-modality generalization. Due to differences in signal sources, PSG data and BCG data have different distributions, and our data augmentation methods successfully aid SleepNetZero in generalizing across modalities.

8 DISCUSSION

8.1 Superiority of SleepNetZero

This work’s strengths lie in its novel SleepNetZero framework, which, for the first time, proposes aligning BCG signals with PSG and achieving better performance than all other BCG-based methods for sleep staging. With this groundbreaking design, we address the scarcity of BCG samples and enhance the reliability of BCG-based sleep staging methods. Former BCG-based methods are heavily limited by the small size of the notated BCG dataset. In contrast, our method reaches a competitive result by leveraging large-scale publicly available PSG datasets.

Novel generalization techniques, such as data augmentation, demonstrate improved performance across sensors, ages, BMIs, sexes, health conditions, and races, as shown in 7. Some research has shown that there exists a gap between populations. With these generalization techniques we successfully get across the gap and reach a good performance across populations.

Extensive experiments validate the framework’s superiority and reliability, further supported by integration into a non-intrusive monitoring pad. By now, our algorithm has been running in several hospitals for some time and the response has been excellent. This affirms the real-world applicability of the SleepNetZero framework compared to previous studies [39, 40, 48, 69, 71].

8.2 Limitations

The main limitations of our work are identified as two problems.

First, methods based on component extraction and alignment face a twofold ceiling. The extraction pipelines still rely on rule-based methods. While they circumvent the scarcity of annotated data, they are not perfectly accurate, limiting the model performance. Moreover, the complicated physiological meanings residing in BCGs

may not be well-exploited. In such cases, utilizing the raw BCG signal may be preferable, as evidenced by the recent practice in the PPG field [28].

Second, the leveraged datasets may introduce biases in the population. While large-scale PSG datasets offer diverse populations, the subjects are recruited under certain conditions. For example, SHHS [46, 75] (5.8k subjects), MrOS [7, 75] (2.9k subjects), and MESA [11, 75] (2k subjects) contains only 40 years or older people. SHHS oversamples snorers. Therefore, patients and the elderly account for a large population. Removing the model's preference towards pathological or elder sleep patterns remains a future topic.

8.3 Future Work

Future studies may focus on two directions.

Learning on sparsely annotated BCG datasets: Despite the scarcity of annotated BCG data, non-annotated data are convenient to collect with the monitoring pad. The data can be recorded every night without extra effort through monitoring pads distributed to the research participants' homes. Large-scale BCG datasets may facilitate self-supervised learning methods, which learn a representation of BCGs, encoding morphology and contextual information. Fine-tuning a downstream task model based on the pre-trained encoding network requires much less annotation.

Sleep disorder diagnosis: The significance of sleep staging resides in sleep disorder diagnosis. The implication of sleep stages still requires experienced doctors to reveal, which causes burdens to patients. In contrast, we wish to pave the way for comprehensive in-home sleep analysis, discovering potential disorders, providing the consumer with early warnings, and preventing worse consequences.

9 CONCLUSION

This work presents a novel neural network framework named SleepNetZero, which achieve high-precision sleep staging by extracting certain physiological components to leverage cross-modality dataset, leading an effective paradigm for BCG-based sleep analysis. The powerful deep-learning model improves performance and generalizability, as the dataset grows in orders of magnitude. The work also incorporates novel data augmentation techniques, which help generalization across signals of different quality. Our work contributes to reliable in-home automated sleep analysis systems, which may unlock the potential of early diagnosis of sleep disorders, and enable longitudinal monitoring that helps evaluate the therapy effects.

ACKNOWLEDGMENTS

This work is supported by the Ministry of Science and Technology of China STI2030-Major Projects (No. 2021ZD0201900, 2021ZD0201902).

REFERENCES

- [1] Anna Abbasi, Sushilkumar Satish Gupta, Nitin Sabharwal, Vineet Meghrajani, Shaurya Sharma, Stephan Kamholz, and Yizhak Kupfer. 2021. A comprehensive review of obstructive sleep apnea. *Sleep Science* 14, 2 (2021), 142.
- [2] Emina Alickovic and Abdulhamit Subasi. 2018. Ensemble SVM method for automatic sleep stage classification. *IEEE Transactions on Instrumentation and Measurement* 67, 6 (2018), 1258–1265.
- [3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.
- [4] Zachary Beattie, Yang Oyang, A Statan, Atiyeh Ghoreyshi, Alexandros Pantelopoulos, Andrew Russell, and CJPM Heneghan. 2017. Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiological measurement* 38, 11 (2017), 1968.
- [5] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, Carole Marcus, Bradley V Vaughn, et al. 2012. The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* 176, 2012 (2012), 7.

- [6] Matt T Bianchi, Sydney S Cash, Joseph Mietus, Chung-Kang Peng, and Robert Thomas. 2010. Obstructive sleep apnea alters sleep stage transition dynamics. *PLoS One* 5, 6 (2010), e11356.
- [7] Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E Ensrud, Marcia L Stefanick, Alison Laffan, Katie L Stone, and Osteoporotic Fractures in Men Study Group. 2011. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *Journal of the American Geriatrics Society* 59, 12 (2011), 2217–2225.
- [8] C Brüser, Stefan Winter, and Steffen Leonhardt. 2013. Robust inter-beat interval estimation in cardiac vibration signals. *Physiological measurement* 34, 2 (2013), 123.
- [9] Mary A Carskadon and Allan Rechtschaffen. 2011. Monitoring and staging human sleep. *Principles and practice of sleep medicine* 5 (2011), 16–26.
- [10] Ramiro Casal, Leandro E Di Persia, and Gastón Schlotthauer. 2021. Automatic sleep staging from pulse oximeter using RNN. In *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 965–969.
- [11] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. 2015. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep* 38, 6 (2015), 877–888.
- [12] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
- [13] dreemhealth. 2024. *dreemhealth*. <https://dreemhealth.com/>
- [14] Jack D Edinger, Ana I Fins, Robert J Sullivan Jr, Gail R Marsh, Dorothy S Dailey, T Victor Hope, Margaret Young, Edmund Shaw, Donna Carlson, and Diane Vasilas. 1997. Sleep in the laboratory and sleep at home: comparisons of older insomniacs and normal sleepers. *Sleep* 20, 12 (1997), 1119–1126.
- [15] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), 809–818.
- [16] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2022. ADAST: Attentive cross-domain EEG-based sleep staging framework with iterative self-training. *IEEE Transactions on Emerging Topics in Computational Intelligence* 7, 1 (2022), 210–221.
- [17] Pedro Fonseca, Merel M van Gilst, Mustafa Radha, Marco Ross, Arnaud Moreau, Andreas Cerny, Peter Anderer, Xi Long, Johannes P van Dijk, and Sebastiaan Overeem. 2020. Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population. *Sleep* 43, 9 (2020), zsaa048.
- [18] Pedro Fonseca, Tim Weysen, Maaïke S Goelema, Els IS Møst, Mustafa Radha, Charlotte Lunsingh Scheurleer, Leonie van den Heuvel, and Ronald M Aarts. 2017. Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. *Sleep* 40, 7 (2017), zsx097.
- [19] Miriam Goldammer, Sebastian Zaunseder, Moritz D Brandt, Hagen Malberg, and Felix Gräßer. 2022. Investigation of automated sleep staging from cardiorespiratory signals regarding clinical applicability and robustness. *Biomedical Signal Processing and Control* 71 (2022), 103047.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [21] Ellen Herbst, Thomas J Metzler, Maryann Lenoci, Shannon E McCaslin, Sabra Inslicht, Charles R Marmar, and Thomas C Neylan. 2010. Adaptation effects to sleep studies in participants with and without chronic posttraumatic stress disorder. *Psychophysiology* 47, 6 (2010), 1127–1133.
- [22] R Holmes, S Tluk, V Metta, P Patel, R Rao, A Williams, and KR Chaudhuri. 2007. Nature and variants of idiopathic restless legs syndrome: observations from 152 patients referred to secondary care in the UK. *Journal of neural transmission* 114 (2007), 929–934.
- [23] Joonki Hong, Hai Hong Tran, Jinhwan Jung, Hyeryung Jang, Dongheon Lee, In-Young Yoon, Jung Kyung Hong, and Jeong-Whun Kim. 2022. End-to-end sleep staging using nocturnal sounds from microphone chips for mobile devices. *Nature and Science of Sleep* (2022), 1187–1201.
- [24] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.
- [25] Jeff Hwang, Moto Hira, Caroline Chen, Xiaohui Zhang, Zhaoheng Ni, Guangzhi Sun, Pingchuan Ma, Ruizhe Huang, Vineel Pratap, Yuekai Zhang, Anurag Kumar, Chin-Yun Yu, Chuang Zhu, Chunxi Liu, Jacob Kahn, Mirco Ravanelli, Peng Sun, Shinji Watanabe, Yangyang Shi, Yumeng Tao, Robin Scheibler, Samuele Cornell, Sean Kim, and Stavros Petridis. 2023. TorchAudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for PyTorch. arXiv:2310.17864 [eess.AS]
- [26] Ziyu Jia, Xiyang Cai, Gaoxing Zheng, Jing Wang, and Youfang Lin. 2020. SleepPrintNet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging. *IEEE Transactions on Artificial Intelligence* 1, 3 (2020), 248–257.

- [27] Vishesh Kapur, Kingman P Strohl, Susan Redline, Conrad Iber, George O'connor, and Javier Nieto. 2002. Underdiagnosis of sleep apnea syndrome in US communities. *Sleep and Breathing* 6, 02 (2002), 049–054.
- [28] Kevin Kotzen, Peter H Charlton, Sharon Salabi, Lea Amar, Amir Landesberg, and Joachim A Behar. 2022. SleepPPG-Net: A deep learning algorithm for robust sleep staging from continuous photoplethysmography. *IEEE Journal of Biomedical and Health Informatics* 27, 2 (2022), 924–932.
- [29] Olave E Krigolson, Mathew R Hammerstrom, Wande Abimbola, Robert Trska, Bruce W Wright, Kent G Hecker, and Gordon Binsted. 2021. Using Muse: Rapid mobile assessment of brain performance. *Frontiers in Neuroscience* 15 (2021), 634147.
- [30] Qiao Li, Qichen Li, Ayse S Cakmak, Giulia Da Poian, Donald L Bliwise, Viola Vaccarino, Amit J Shah, and Gari D Clifford. 2021. Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables. *Physiological measurement* 42, 4 (2021), 044004.
- [31] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [32] Yujie Luo, Junyi Li, Kejing He, and William Cheuk. 2022. A Hierarchical Attention-Based Method for Sleep Staging Using Movement and Cardiopulmonary Signals. *IEEE Journal of Biomedical and Health Informatics* 27, 3 (2022), 1354–1363.
- [33] Yaopeng JX Ma, Johannes Zschocke, Martin Glos, Maria Kluge, Thomas Penzel, Jan W Kantelhardt, and Ronny P Bartsch. 2023. Automatic sleep-stage classification of heart rate and actigraphy data using deep and transfer learning approaches. *Computers in Biology and Medicine* 163 (2023), 107193.
- [34] Ulysses J Magalang, Ning-Hung Chen, Peter A Cistulli, Annette C Fedson, Thorarinn Gíslason, David Hillman, Thomas Penzel, Renaud Tamisier, Sergio Tufik, Gary Phillips, et al. 2013. Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep* 36, 4 (2013), 591–596.
- [35] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior research methods* (2021), 1–8.
- [36] Bryce A Mander, Joseph R Winer, and Matthew P Walker. 2017. Sleep and human aging. *Neuron* 94, 1 (2017), 19–36.
- [37] Janna Mantua, Antigone Grillakis, Sanaa H Mahfouz, Maura R Taylor, Allison J Brager, Angela M Yarnell, Thomas J Balkin, Vincent F Capaldi, and Guido Simonelli. 2018. A systematic review and meta-analysis of sleep architecture and chronic traumatic brain injury. *Sleep medicine reviews* 41 (2018), 61–77.
- [38] Sebastiano Massaro and Leandro Pecchia. 2019. Heart rate variability (HRV) analysis: A methodology for organizational neuroscience. *Organizational research methods* 22, 1 (2019), 354–393.
- [39] Matteo Migliorini, Anna M Bianchi, Domenico Nisticò, Juha Kortelainen, Edgar Arce-Santana, Sergio Cerutti, and Martin O Mendez. 2010. Automatic sleep staging based on ballistocardiographic signals recorded through bed sensors. In *2010 annual international conference of the IEEE engineering in medicine and biology*. IEEE, 3273–3276.
- [40] Yasue Mitsukura, Brian Sumali, Masaki Nagura, Koichi Fukunaga, and Masato Yasui. 2020. Sleep stage estimation from bed leg ballistocardiogram sensors. *Sensors* 20, 19 (2020), 5688.
- [41] Seiichi Morokuma, Toshinari Hayashi, Masatomo Kanegae, Yoshihiko Mizukami, Shinji Asano, Ichiro Kimura, Yuji Tateizumi, Hitoshi Ueno, Subaru Ikeda, and Kyuichi Niizeki. 2023. Deep learning-based sleep stage classification with cardiorespiratory and body movement activities in individuals with suspected sleep disorders. *Scientific reports* 13, 1 (2023), 17730.
- [42] Maria Pallyova, Viliam Donic, Sona Gresova, Igor Peregrin, and Zoltan Tomori. 2010. Do differences in sleep architecture exist between persons with type 2 diabetes and nondiabetic controls? *Journal of Diabetes Science and Technology* 4, 2 (2010), 344–352.
- [43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
- [44] Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2022. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering* 69, 8 (2022), 2456–2467.
- [45] Wei Qu, Zhiyong Wang, Hong Hong, Zheru Chi, David Dagan Feng, Ron Grunstein, and Christopher Gordon. 2020. A residual based attention model for eeg based sleep staging. *IEEE journal of biomedical and health informatics* 24, 10 (2020), 2833–2843.
- [46] Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. 1997. The sleep heart health study: design, rationale, and methods. *Sleep* 20, 12 (1997), 1077–1085.
- [47] Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M Aarts. 2021. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *NPJ digital medicine* 4, 1 (2021), 135.
- [48] Shashank Rao, Abdallah El Ali, and Pablo Cesar. 2019. DeepSleep: a ballistocardiographic deep learning approach for classifying sleep stages. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 187–190.
- [49] Susan Redline, Peter V Tishler, Tor D Tosteson, John Williamson, Kenneth Kump, Ilene Browner, Veronica Ferrette, and Patrick Krejci. 1995. The familial aggregation of obstructive sleep apnea. *American journal of respiratory and critical care medicine* 151, 3 (1995), 682–687.
- [50] Dieter Riemann, Mathias Berger, and Ulrich Voderholzer. 2001. Sleep and depression—results from psychobiological studies: an overview. *Biological psychology* 57, 1-3 (2001), 67–103.

- [51] A Roebuck, V Monasterio, E Geder, M Osipov, J Behar, A Malhotra, T Penzel, and GD Clifford. 2013. A review of signals used in sleep analysis. *Physiological measurement* 35, 1 (2013), R1.
- [52] Carol L Rosen, Dennis Auckley, Ruth Benca, Nancy Foldvary-Schaefer, Conrad Iber, Vishesh Kapur, Michael Rueschman, Phyllis Zee, and Susan Redline. 2012. A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: the HomePAP study. *Sleep* 35, 6 (2012), 757–767.
- [53] Carol L Rosen, Emma K Larkin, H Lester Kirchner, Judith L Emancipator, Sarah F Bivins, Susan A Surovec, Richard J Martin, and Susan Redline. 2003. Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity. *The Journal of pediatrics* 142, 4 (2003), 383–389.
- [54] Ibrahim Sadek, Jit Biswas, and Bessam Abdulrazak. 2019. Ballistocardiogram signal processing: A review. *Health information science and systems* 7, 1 (2019), 10.
- [55] Jonathan RL Schwartz and Thomas Roth. 2008. Neurophysiology of sleep and wakefulness: basic science and clinical implications. *Current neuropharmacology* 6, 4 (2008), 367–378.
- [56] Catherine Siengsukon, Mayis Al-Dughmi, Alham Al-Sharman, and Suzanne Stevens. 2015. Sleep parameters, functional status, and time post-stroke are associated with offline motor skill learning in people with chronic stroke. *Frontiers in neurology* 6 (2015), 225.
- [57] Niranjan Sridhar, Ali Shoeb, Philip Stephens, Alaa Kharbouch, David Ben Shimol, Joshua Burkart, Atiyeh Ghoreyshi, and Lance Myers. 2020. Deep learning for automated sleep staging using instantaneous heart rate. *NPJ digital medicine* 3, 1 (2020), 106.
- [58] Lars St, Svante Wold, et al. 1989. Analysis of variance (ANOVA). *Chemometrics and intelligent laboratory systems* 6, 4 (1989), 259–272.
- [59] Ambra Stefani and Birgit Högl. 2020. Sleep in Parkinson’s disease. *Neuropsychopharmacology* 45, 1 (2020), 121–128.
- [60] Haoqi Sun, Wolfgang Ganglberger, Ezhil Panneerselvam, Michael J Leone, Syed A Quadri, Balaji Goparaju, Ryan A Tesh, Oluwaseun Akeju, Robert J Thomas, and M Brandon Westover. 2020. Sleep staging from electrocardiography and respiration with deep learning. *Sleep* 43, 7 (2020), zsz306.
- [61] WE Thomson. 1949. Delay networks having maximally flat frequency characteristics. *Proceedings of the IEE-Part III: Radio and Communication Engineering* 96, 44 (1949), 487–490.
- [62] Eleonora Tobaldini, Lino Nobili, Silvia Strada, Karina R Casali, Alberto Braghiroli, and Nicola Montano. 2013. Heart rate variability in normal and pathological sleep. *Frontiers in physiology* 4 (2013), 62099.
- [63] Erdenebayar Urtnasan, Jong-Uk Park, Eun Yeon Joo, and Kyoung-Joung Lee. 2022. Deep convolutional recurrent model for automatic scoring sleep stages based on single-lead ECG signal. *Diagnostics* 12, 5 (2022), 1235.
- [64] Mahtab Vaezi and Mehdi Nasri. 2023. AS3-SAE: Automatic Sleep Stages Scoring using Stacked Autoencoders. *Frontiers in Biomedical Technologies* (2023).
- [65] Fokke B van Meulen, Angela Grassi, Leonie Van den Heuvel, Sebastiaan Overeem, Merel M van Gilst, Johannes P van Dijk, Henning Maass, Mark JH Van Gastel, and Pedro Fonseca. 2023. Contactless camera-based sleep staging: The healthbed study. *Bioengineering* 10, 1 (2023), 109.
- [66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [67] Bernard L Welch. 1947. The generalization of ‘STUDENT’S’ problem when several different population variances are involved. *Biometrika* 34, 1-2 (1947), 28–35.
- [68] M. B. Westover, V. Moura Junior, R. Thomas, S. Cash, S. Nasiri, H. Sun, A. Gupta, J. Rosand, M. Ghanta, W. Ganglberger, U. Katwa, K. Stone, Z. Zhang, G. Ganjoo, T. E. Nassi PhD Candidate, R. Wei, D. Hwang, L. M. Trotti, A. Parekh, D. Rapoport, et al. 2023. The Human Sleep Project (version 2.0). *Brain Data Science Platform* 2, 0 (2023). <https://doi.org/10.60508/qjbv-hg78>
- [69] Longwen Wu, Pengcheng Ren, Yaqin Zhao, Ruchen Lv, Qinyu Ding, and Yirui Zuo. 2023. Sleep Stage Classification based on BCG using Improved Deep Convolutional Generative Adversarial Networks. In *2023 7th International Conference on Imaging, Signal Processing and Communications (ICISPC)*. IEEE, 65–69.
- [70] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhersch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. 2021. TorchAudio: Building Blocks for Audio and Speech Processing. *arXiv preprint arXiv:2110.15018* (2021).
- [71] Ruhan Yi, Moein Enayati, James M Keller, Mihail Popescu, and Marjorie Skubic. 2019. Non-invasive in-home sleep stage classification using a ballistocardiography bed sensor. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 1–4.
- [72] Bing Zhai, Yu Guan, Michael Catt, and Thomas Plötz. 2021. Ubi-SleepNet: Advanced multimodal fusion techniques for three-stage sleep classification using ubiquitous sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–33.
- [73] Bing Zhai, Ignacio Perez-Pozuelo, Emma AD Clifton, Joao Palotti, and Yu Guan. 2020. Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–33.

- [74] Feng Zhang, Rujia Zhong, Song Li, Zhenfa Fu, Renfei Wang, Tianxiao Wang, Zhili Huang, and Weidong Le. 2019. Alteration in sleep architecture and electroencephalogram as an early sign of Alzheimer’s disease preceding the disease pathology and cognitive decline. *Alzheimer’s & Dementia* 15, 4 (2019), 590–597.
- [75] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. 2018. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association* 25, 10 (2018), 1351–1358.
- [76] Zhiwei Zhang and Minfang Tang. 2023. A Domain-Based, Adaptive, Multi-Scale, Inter-Subject Sleep Stage Classification Network. *Applied Sciences* 13, 6 (2023), 3474.
- [77] Dechun Zhao, Renpin Jiang, Mingyang Feng, Jiaxin Yang, Yi Wang, Xiaorong Hou, and Xing Wang. 2022. A deep learning algorithm based on 1D CNN-LSTM for automatic sleep staging. *Technology and Health Care* 30, 2 (2022), 323–336.
- [78] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*. PMLR, 4100–4109.

A THE PHYSIOLOGICAL MEANINGS OF THE SLEEP STAGES

- **Wake (W):** The state of being awake, which typically occurs at the start and end of the sleep record, as well as during brief awakening periods throughout the night.
- **Non-REM Stage 1 (N1):** The lightest stage of sleep, often considered a transition phase between wakefulness and more profound sleep stages.
- **Non-REM Stage 2 (N2):** A stage of sleep that represents a deeper sleep than N1 and accounts for the majority of sleep time. It is characterized by specific brain waveforms such as sleep spindles and K-complexes.
- **Non-REM Stage 3 (N3):** The deepest stage of non-REM sleep, often referred to as slow-wave sleep due to the presence of high amplitude, low-frequency brain waves. It plays a crucial role in physical recovery and memory consolidation.
- **REM Sleep (R):** The stage of sleep associated with rapid eye movements, where most vivid dreaming occurs, and is linked to brain regions involved in learning and memory.

B MORE DATASET INFORMATION

The age and sex distributions of the NSRR dataset are listed in Tables 11 and 12, respectively.

Age Group	Percentage
0–20	3.9%
20–40	1.6%
40–60	22.6%
60–80	59.2%
80+	12.7%

Table 11. The age distribution of the NSRR dataset. Most (71.9%) subjects are over 60 years old. This is because some of the datasets only enroll elderly people.

Sex	Percentage
Male	60.3%
Female	39.6%

Table 12. The sex distribution of the NSRR dataset. Males are more than females.

In the BCG265 dataset, the distribution of class labels is detailed in Table 13.

C A FURTHER DISCUSSION ON THE BCG SENSOR

System Design. The system design of our monitoring pads is shown in Figure 7.

Comparison to other non-contact sensors. In similar scenarios, sensors include optical fibers and piezoresistive sensors. Compared to piezoresistive sensors, our sensors have higher precision and sensitivity, and they are

Stage	W	N1	N2	N3	R
Percentage	25%	12%	34%	15%	14%

Table 13. Stage distribution of the BCG265 dataset.

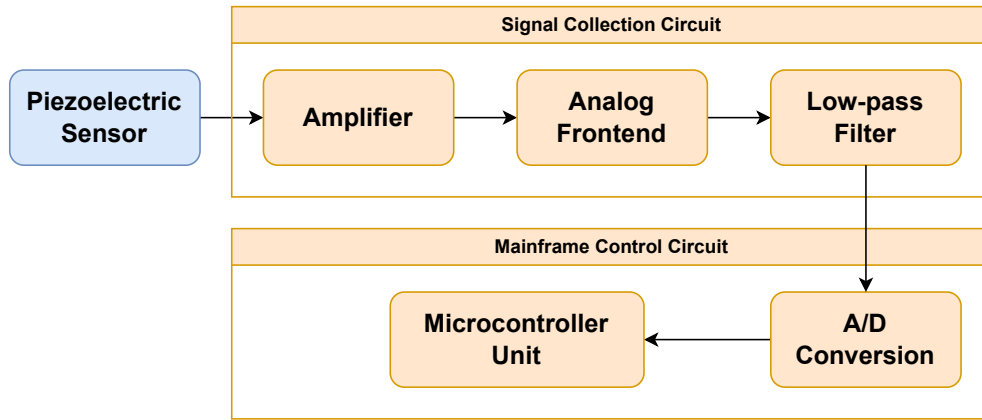


Fig. 7. The system design of our monitoring pads.

capable of capturing physiological information related to breathing and heartbeats. The accuracy and sensitivity of optical fiber sensors are somewhat higher than our piezoelectric sensors. However, they are more expensive, have complex encapsulation processes (requiring weaving into a mesh), and cannot withstand significant external pressure. Compared to these similar sensors, the piezoelectric sensors used in this paper meet the requirements for the precision and sensitivity of perceiving sleep activities and can still remain operational after external pressure.

Influence of co-sleepers. In scenarios with co-sleepers, if the user themselves is also sleeping in the bed, the sensor's signal is mainly influenced by the user, with minimal impact from the co-sleeper. If the user is away from the bed, the sensor signal may be influenced by the co-sleeper, which is also one of the issues we need to address next. To obtain reliable data for testing the model, we currently require that only one person be present during the data collection process.

D CODE AVAILABILITY

Source code is available at <https://github.com/nealchen2003/IMWUT2024-SleepNetZero>.

Received 1 May 2024; revised 1 August 2024; accepted 20 September 2024