Sleep Staging Using PPG-Derived Heart Rate and Accelerometer Data from a Smart Ring with Lightweight Neural Networks

Roneel V. Sharan, *Senior Member, IEEE*, Hiroki Takeuchi, Akifumi Kishi, Jiabin Wang, Tatsuhiko Watanabe, and Yoshiharu Yamamoto, *Member, IEEE*

***Abstract*—Sleep staging plays an important role in assessing and diagnosing sleep disorders, however, polysomnography, the gold standard for sleep monitoring, is expensive, time-consuming, and often inaccessible. Wearable devices offer a promising alternative, providing a more convenient and scalable solution for long-term sleep monitoring. Compared to wrist-worn devices, smart rings offer a more stable signal acquisition environment, reducing motion artifacts and improving data reliability. This work investigates sleep staging using photoplethysmography (PPG)-derived instantaneous heart rate (IHR) signal and zero-crossing mode (ZCM) feature obtained from accelerometer data, both captured using a wearable smart ring. We propose a lightweight neural network model designed for wearable-based sleep staging, incorporating feature extraction, temporal modeling, and class balancing strategies. The IHR-based model consists of 503 k learnable parameters, while the ZCM-based model has 133 k learnable parameters, making them well-suited for efficient deployment on wearable devices. The proposed method is evaluated on a dataset of expert-annotated overnight sleep studies, using both IHR signal and ZCM features individually and in combination, in subject-independent cross-validation. The experimental results demonstrate that IHR features alone yield strong classification performance, achieving macro-average recall (unweighted average recall) values of 0.849, 0.805, 0.750, and 0.663 in two-class (wake vs. sleep), three-class (wake vs. non-rapid eye movement (NREM) sleep (N1, N2, N3) vs. REM sleep), four-class (wake vs. light sleep vs. deep sleep vs. REM sleep), and five-class (wake vs. N1 vs. N2 vs. N3 vs. REM) classification tasks, respectively. When combining IHR and ZCM features, classification performance improves further, reaching macro-average recall values of 0.866, 0.832, 0.772, and 0.671 in the respective tasks. These results highlight the effectiveness of IHR-based sleep staging and the additional benefit provided by movement-based ZCM features, particularly in two-, three-, and four-class sleep staging where we could achieve macro-average recall values of 0.750 or higher. The proposed smart ring-based system demonstrates strong potential for real-world sleep assessment by integrating multimodal physiological signals through lightweight neural networks, advancing non-invasive measurement and intelligent instrumentation.**

***Index Terms*—Convolutional neural networks, gated recurrent units, instantaneous heart rate, sleep staging, smart ring, wearables, zero-crossing**

This work was supported in part by the University of Essex, the JST FORESTO Program under Grant JPMJFR2156, and the JST MOONSHOT R&D Grants JPMJMS229B and JPMJMS2021.

Roneel V. Sharan is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom (e-mail: roneel.sharan@essex.ac.uk).

Hiroki Takeuchi and Yoshiharu Yamamoto are with the Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo 113-0033, Japan (e-mail: takeuchi@p.u-tokyo.ac.jp; yamamoto@p.u-tokyo.ac.jp).

Akifumi Kishi is with the Graduate School of Medicine, The University of Tokyo, Tokyo 113-0033, Japan (e-mail: kishi@p.u-tokyo.ac.jp).

Jiabin Wang and Tatsuhiko Watanabe are with SOXAI Inc., Yokohama, Kanagawa 231-0032, Japan (e-mail: jiabin.wang@soxai.co.jp; tatsuhiko.watanabe@soxai.co.jp).

I. Introduction

SLEEP plays an important role in overall health and well-being, yet sleep disorders and insufficient sleep remain prevalent [1]. Insufficient sleep has been linked to cognitive decline, increased risk of motor vehicle accidents, reduced quality of life, and various mental and physical health conditions [2]. Early detection and diagnosis of sleep disorders, such as insomnia and obstructive sleep apnea, are essential for effective intervention, as proper management can restore sleep quality, reduce daytime fatigue, and lower associated health risks [3].

The gold standard for sleep assessment is polysomnography (PSG), which involves overnight monitoring of physiological signals such as brain activity, eye movements, heart rhythm, and muscle activity. PSG data is segmented into epochs, and sleep stages, wakefulness (W), non-rapid eye movement (NREM) sleep (N1, N2, N3), and rapid eye movement (REM) sleep, are manually scored based on standard criteria. Sleep staging provides valuable metrics, including total sleep time, sleep efficiency, and sleep onset latency, which are essential for diagnosing sleep disorders [4]. However, PSG is expensive, labor-intensive, and requires specialized facilities, making it inaccessible for routine or large-scale sleep monitoring [5], [6].

Wearable devices have emerged as an attractive alternative for sleep monitoring [7], [8], [9], [10]. This is due to their ability to provide continuous, non-invasive data collection in a home setting [6]. Photoplethysmography (PPG) and accelerometer sensors, commonly integrated into wearable devices, enable heart rate and movement-based sleep monitoring. Smart rings, in particular, offer a promising form factor for unobtrusive, long-term sleep tracking. Compared to wrist-worn devices, smart rings provide a more stable signal acquisition environment, minimizing motion artifacts and improving data reliability [11]. However, due to constraints on computational resources [11], sleep staging models must be lightweight while maintaining high accuracy [12].

This work develops a neural network model for sleep staging using PPG-derived instantaneous heart rate (IHR)

and zero-crossing mode (ZCM) features extracted from an accelerometer in a smart ring. The network consists of a convolutional neural network (CNN) and a bidirectional gated recurrent unit (BiGRU), where the CNN captures short-term temporal patterns, and the BiGRU models long-term dependencies across sleep epochs [13], possibly learning the stage transition patterns observed in natural human hypnograms [14], [15]. The IHR and ZCM features are processed independently before being combined for final classification. To address class imbalance in sleep staging, we incorporate inverse class weighting. Notably, the model is designed to be computationally efficient, making the architecture well-suited for potential deployment on resource-constrained wearable devices. We evaluate the proposed model in subject-independent cross-validation and the model's effectiveness is demonstrated across different sleep stage classification tasks.

The main contributions of this work are as follows:

1) Lightweight sleep staging model: Development of a computationally efficient CNN-BiGRU architecture tailored for PPG-derived IHR signal from a smart ring, balancing accuracy with deployability in resource-constrained wearable devices.
2) Feature integration strategy: Independent processing and subsequent fusion of IHR and accelerometer-derived ZCM features, enabling complementary information from cardiac activity and movement to be effectively utilized for sleep stage classification.
3) Transfer learning across modalities: Pretraining on a large ECG-derived IHR dataset and fine-tuning on a small PPG-derived IHR dataset, demonstrating that cross-modality transfer is feasible and significantly improves performance.
4) Robust evaluation: Comprehensive assessment of the proposed framework under subject-independent cross-validation across multiple staging granularities, benchmarked against classical machine learning, feature-based, and deep learning baselines.
5) Practical relevance for wearables: Demonstration of strong performance with very low model complexity and short inference time, highlighting the suitability of the proposed method for integration into smart rings and similar wearable platforms.

In doing so, this work contributes to the development of intelligent instrumentation systems by advancing non-invasive measurement methods for continuous physiological monitoring. It also addresses an important gap in the instrumentation and measurement literature by demonstrating efficient signal processing and classification pipelines suitable for implementation in wearable health monitoring platforms.

The remainder of the paper is organized as follows. Section II describes the dataset, signal processing, and network architecture. Section III presents the experimental setup and results, and Section IV provides discussion and conclusions.

II. Materials and Methods

An overview of the proposed method for classifying sleep stages using PPG and accelerometer signals is shown in Fig. 1. We preprocess the PPG and accelerometer signals to extract the IHR signal and ZCM feature, respectively. These serve as inputs to separate neural networks, whose activations are then concatenated for sleep stage classification. More details are provided in the following subsections.

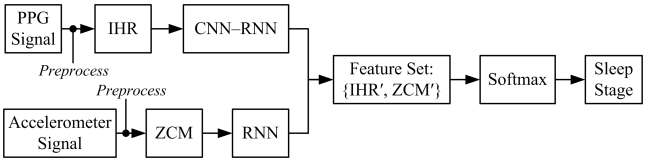

Fig. 1. Overview of the method developed for smart ring-based sleep staging.

A. Dataset

The dataset used in this study was obtained from the Sleep Laboratory at The University of Tokyo, Japan. It consists of 18 overnight sleep recordings collected from nine healthy adult subjects, with each subject contributing two separate recordings. Subjects were instructed to avoid napping, engaging in strenuous exercise, and consuming alcohol on the two experimental days. They arrived at the laboratory approximately two hours before their self-selected lights-off time. Upon arrival, they completed brief baseline questionnaires (Table I) and took a shower to remove sebum before undergoing nocturnal polysomnography (PSG). The PSG recording included electroencephalogram (F3/A2, F4/A1, C3/A2, C4/A1, O1/A2, and O2/A1), electrooculogram, submental electromyogram, lead II electrocardiogram, and pulse oximetry.

In addition to PSG, subjects wore a SOXAI smart ring device on an arbitrary finger of the right hand to collect tri-wavelength PPG signals by green, red, and infrared LEDs and triaxial accelerometer data. PSG data were acquired using the Polymate Pro MP6000 system (Miyuki Giken Co., Ltd) at a sampling frequency of 500 Hz. The SOXAI ring transmitted data every 30 seconds to an in-house Android smartphone application, with PPG signals sampled at 200 Hz and accelerometer signals at 50 Hz. The clocks between the PSG system and smart ring were manually synchronized. Data recording began two minutes before the lights-off time and continued until two minutes after lights-on. Sleep stages were scored in 30-second non-overlapping epochs into standard sleep categories, Wake (W), NREM1 (N1), NREM2 (N2), NREM3 (N3), and REM (R), by a single technician blinded to the study objectives, following the criteria established by the American Academy of Sleep Medicine (AASM). Ethical approval for this study was granted by the Ethics Review Board at The University of Tokyo under Application No. 21-11, ensuring compliance with ethical research guidelines involving human subjects.

A breakdown of the dataset composition is provided in Table I. The dataset includes individuals between the ages of 23 and 36 years, with an average age of 27.7±3.5 years. Among the nine subjects, seven were male, and two were female. The Epworth Sleepiness Scale (ESS) scores [16], which provide a subjective measure of daytime sleepiness, had an average of 7.3±4.1 across the participants. A total of 15,295 epochs were available in the dataset. The distribution of sleep stages

TABLE I
DATASET CHARACTERISTICS

Parameter	Value
Number of subjects (recordings)	9 (18)
Age range (years)	23–36
Average age (years)	27.7±3.5
Sex (Male:Female)	7:2
Body mass index (kg/m^2)	22.8±2.8
Epworth sleepiness scale	7.3±4.1
Number of Wake (W) epochs	1,365 (8.92%)
Number of NREM1 (N1) epochs	901 (5.89%)
Number of NREM2 (N2) epochs	7,041 (46.03%)
Number of NREM3 (N3) epochs	2,890 (18.90%)
Number of REM (R) epochs	3,098 (20.26%)

is highly imbalanced, with N2 sleep comprising the largest proportion (7,041 epochs) and N1 sleep having the least representation (901 epochs).

B. Data Preprocessing

To ensure the reliability of the PPG data, a multi-step preprocessing pipeline was implemented. First, artifacts were mitigated using a Hampel filter, which detects and removes outlier values caused by motion artifacts or hardware issues. Subsequently, a low-pass filter with a cutoff frequency of 25 Hz was applied to remove high-frequency noise and prevent aliasing before downsampling. The PPG signal was then downsampled from its original 200 Hz sampling rate to 25 Hz, aligning it with the specifications of commercially available smart ring devices.

The accelerometer data were also downsampled to 25 Hz to maintain consistency with the PPG signal. Since the start and end times of PPG, accelerometer data, and PSG sleep stages were recorded separately, the signals and the corresponding sleep stages were synchronized between the light-off and light-on timestamps.

Four different sleep staging tasks were conducted to assess model performance under varying levels of complexity: a two-class classification where sleep was categorized into wake (W) and sleep (S), with sleep encompassing all NREM and REM sleep stages; a three-class classification that included wake (W), NREM sleep (N1, N2, N3), and REM sleep (R); a four-class classification that distinguished wake (W), light sleep (N1, N2), deep sleep (N3), and REM sleep (R); and a five-class classification that distinguished all individual sleep stages: W, N1, N2, N3, and R.

C. IHR Signal and ZCM Feature

The IHR was extracted from the PPG signal, captured using the green LED, using the event-related moving averages method [17]. The interbeat interval (IBI), which represents the time difference between consecutive PPG peaks, was computed. The IHR was then obtained as the reciprocal of the IBIs. To facilitate model training, the IHR values in each recording were normalized by subtracting the mean and dividing by the standard deviation. Finally, the IHR signals were resampled to 2 Hz using linear interpolation to maintain a consistent temporal resolution.

The ZCM features were derived from the accelerometer signals [18]. The three-dimensional body acceleration data were first filtered using a 6th-order Butterworth filter with the pass band between 2 Hz and 3 Hz. Then, the number of times the signal level crossed 0.01 G (above the noise threshold) within each minute for each axis was counted. Finally, the maximum value for three axes was adopted as the ZCM for that minute. This form of ZCM count is effective for distinguishing between sleep and wakefulness [18]; it rarely produces count values close to zero during wakefulness.

D. Neural Network Architecture for IHR Signal

The neural network designed for sleep staging using PPG-derived IHR signals follows a CNN-RNN hybrid architecture, as shown in Fig. 2. This model processes IHR signals over a 150-second time window centered on the current epoch [13]. Since the IHR is resampled at 2 Hz, this results in an input sequence length of 300 samples per epoch. This time-series data is then fed directly into the neural network, where it is first processed by convolutional layers, which operate on each 150-second window (300 samples) to extract spatial features, followed by a BiGRU that models the temporal dependencies across consecutive windows.

The CNN module consists of six one-dimensional convolutional layers. Each of these layers has 64 filters with a kernel size of 1×6 and a stride of 1×1, ensuring that fine-grained features from the IHR sequence are captured effectively. After each convolutional layer, a batch normalization layer [19] is applied to stabilize learning and accelerate convergence. Following this, a rectified linear unit (ReLU) activation function [20] introduces non-linearity into the network, enabling it to learn complex representations. A max-pooling layer [21] with a stride of 1×2 follows each convolutional operation, progressively reducing the dimensionality of the extracted features while preserving essential temporal patterns.

After the final max-pooling operation, a flattening layer transforms the feature maps into a one-dimensional vector, making the data suitable for sequential processing by the BiGRU. We chose GRUs instead of LSTMs because they achieve comparable performance in sequence modeling while being computationally lighter [22], making them better suited for resource-constrained wearable devices. A bidirectional model was adopted to fully capture temporal dependencies and stage transition dynamics [23], thereby providing an upper-bound on achievable performance. However, we also discuss uni-directional variants for real-time deployment trade-offs.

The BiGRU module is designed to model the temporal dependencies inherent in sleep stage transitions. A GRU unit processes the sequence from past to future, while another GRU unit simultaneously processes it in reverse, allowing the network to incorporate contextual information from both preceding and succeeding time steps. The hidden state at each time step is updated through a gating mechanism that includes an update gate, a reset gate, and a candidate state. These gates regulate the flow of information, ensuring that relevant past

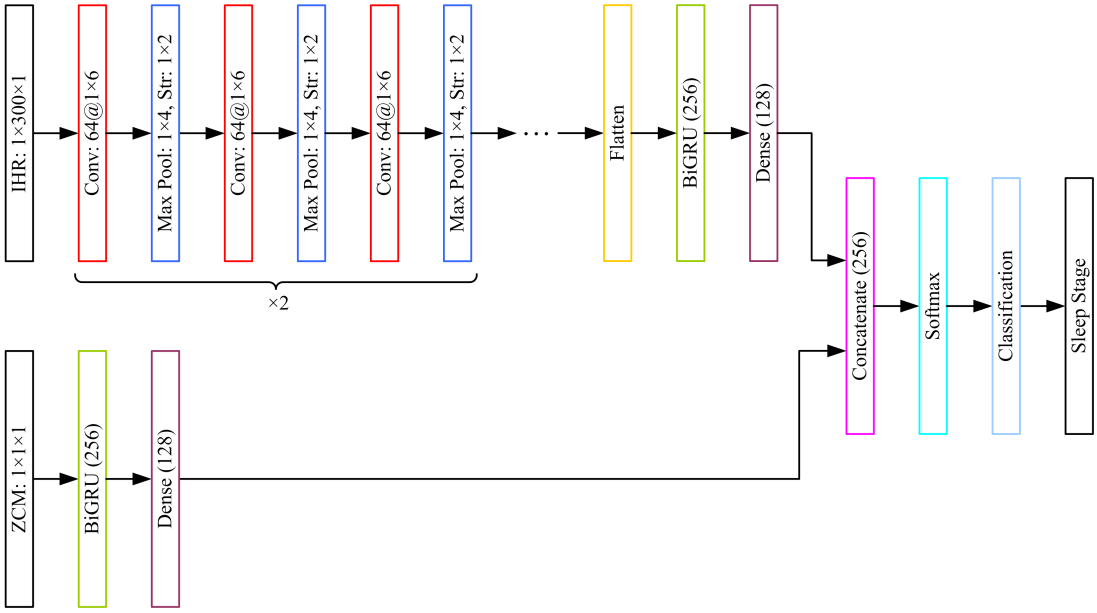

Fig. 2. Neural network architecture for sleep staging with IHR signal and ZCM feature derived from PPG and accelerometer signals, respectively, in a smart ring.

features are retained while redundant details are discarded. The hidden state at time step t is computed as:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \tag{1}$$

where z_t represents the update gate, $\hat{h}_t$ denotes the candidate hidden state, and h_{t-1} is the previous hidden state. By leveraging this bidirectional structure, the BiGRU effectively captures long-range dependencies, improving sleep stage classification accuracy.

A fully connected (dense) layer of size 128 follows the BiGRU, refining the extracted features before the final classification layer. The network applies a softmax classifier to predict sleep stages based on the processed IHR features.

To address the significant class imbalance in sleep staging, inverse class weights are incorporated into the network's loss function. The weighted cross-entropy loss ensures that underrepresented sleep stages contribute proportionally to the training process, preventing the network from being biased toward majority classes. The weight assigned to each class is computed as:

$$w_i = \frac{N}{K \sum_{n=1}^{N} \mathrm{t}_{ni}}. \tag{2}$$

where N is the total number of samples, K is the number of sleep stage classes, and t_{ni} represents whether sample n belongs to class i. This weighting scheme encourages the network to learn balanced representations across all sleep stages.

As such, the IHR-based CNN-BiGRU architecture leverages convolutional layers to extract spatial features from heart rate signals, while the recurrent layers model sleep stage transitions over time. The 150-second input window ensures that the network has sufficient temporal context to classify the current epoch accurately. The combination of feature extraction, temporal modeling, and class balancing strategies, along with a lightweight architecture of only 35 layers and 503 k learnable parameters, makes this model well-suited for sleep staging using wearable devices.

E. Neural Network Architecture for ZCM

The neural network architecture developed for sleep staging using ZCM values is based on a BiGRU and follows a similar classification approach as the IHR network, as shown in Fig. 2. However, unlike the IHR network, which includes convolutional layers for feature extraction, the ZCM network directly processes the time-series input using recurrent layers.

The input to the network consists of ZCM values extracted from accelerometer signals, formatted as sequential data. The BiGRU layer processes this input to capture temporal dependencies and learn sleep stage transitions effectively. This layer is identical in structure to the one used in the IHR network, allowing bidirectional processing to incorporate both past and future context.

Following the BiGRU, a fully connected layer with 128 units refines the extracted features. The final classification is performed using another fully connected layer with output units corresponding to the number of sleep stages (two, three, four, or five). This layer is followed by a softmax activation function, converting the outputs into probability distributions for classification. To address the class imbalance, inverse class weights are applied in the classification layer's loss function. The weight computation follows the same methodology as in the IHR network. The ZCM network has a total of 10 layers and 133 k learnable parameters, making it a lightweight model for sleep staging.

F. Neural Network Architecture for IHR & ZCM

The final classification network integrates features extracted from both the IHR and ZCM networks to improve sleep stage classification performance. The combination of these two modalities allows the model to leverage complementary information from cardiovascular and movement-based signals, enhancing its ability to differentiate between sleep stages.

To achieve this, the activations from the fully connected layers of the IHR and ZCM networks, each consisting of 128 units, are concatenated to form a combined feature representation of size 256. This combined feature vector serves as the input to the final classification network. A fully connected layer follows, with output units corresponding to the number of sleep stages in the specific classification task (two, three, four, or five classes). A softmax activation function is applied to transform the network's outputs into probability scores for each class. The final classification is performed using a classification layer, which employs inverse class weights to mitigate the impact of class imbalance, as described in the previous subsections.

G. Evaluation Metrics

To evaluate the performance of the proposed method in sleep staging, we utilize recall, precision, F_1 score, and Cohen's kappa (κ).

Recall, also referred to as class accuracy, measures the proportion of correctly identified instances for a given class out of all actual instances of that class. It is computed as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

where TP (true positives) represents correctly classified instances of a class, and FN (false negatives) represents instances that belong to the class but were misclassified.

Precision evaluates how many of the predicted instances for a class were actually correct. It is given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

where FP (false positives) denotes instances that were incorrectly classified as belonging to the class.

F_1 score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. It is computed as:

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{5}$$

In addition, we assess the agreement between the predicted sleep stage labels and the ground truth annotations using Cohen's kappa (κ) [24]. Kappa takes into account the possibility of agreement occurring by chance and is computed as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{6}$$

where P_o is the observed agreement between the model predictions and the ground truth, and P_e is the expected agreement under random chance. The interpretation of κ values follows established guidelines [25].

Since class distributions are highly imbalanced in sleep staging [26], we also report the macro-average, also referred to as unweighted average, of recall, F_1 score, and κ values across all sleep stages. The macro-average recall is also known as unweighted average recall (UAR). The macro-average provides a more balanced performance measure by ensuring that underrepresented sleep stages are given equal importance in performance evaluation, rather than being dominated by majority classes.

III. Experimental Evaluation

A. Experimental Setup

To evaluate the performance of the proposed models, a leave-one-subject-out cross-validation approach was implemented. In this setup, data from one subject was designated as the test set, while data from the remaining eight subjects were partitioned such that seven were used for training and one was used for validation. This process was repeated iteratively until every subject had been used as the test set once, ensuring a comprehensive assessment of model generalizability across individuals.

The optimization of all three networks, namely, the IHR-based model, the ZCM-based model, and the combined IHR & ZCM model, was performed using the adaptive moment estimation algorithm [27]. The initial learning rate for all models was set at 0.001, with a maximum of 30 training epochs to allow sufficient convergence while preventing overfitting.

To enhance the robustness of the IHR-based model, we pretrained and validated it on the larger MGH dataset [28], comprising ECG-derived IHR from 993 subjects [13], where R-peaks were detected using the Pan–Tompkins algorithm [29]. This dataset was divided into three subsets: 557 subjects for training, 238 subjects for validation, and 198 subjects for testing. The pretraining on this larger dataset ensured that the IHR-based model learned generalizable representations before being fine-tuned on the sleep staging dataset collected for this study.

B. Results for Sleep Staging Using ECG-Derived IHR on the Pretraining (MGH) Dataset

Table II summarizes the macro-average recall, F_1 score, and κ achieved by the IHR model on the MGH test set across different sleep staging tasks. In the two-class task (W vs. S), the model achieved a recall of 0.824, F_1 score of 0.885, and a κ value of 0.509. As expected, performance declined with increasing classification granularity. For the five-class task (W, N1, N2, N3, R), the model attained a recall of 0.630, F_1 score of 0.586, and κ of 0.482. These results indicate that the pretrained model captures meaningful sleep stage information from ECG-derived IHR and serves as a strong foundation for subsequent training with PPG-based signals.

C. Results for Sleep Staging Using Smart Ring Dataset

The macro-average recall, F_1 score, and κ values in sleep staging with different number of classes and different inputs (IHR, ZCR, and IHR & ZCM) in leave-one-subject-out cross-validation on the smart ring dataset are presented in Table III.

TABLE II
RESULTS FOR SLEEP STAGING WITH DIFFERENT NUMBER OF CLASSES USING ECG-DERIVED IHR ON THE PRETRAINING (MGH) DATASET

Number of classes	Macro-average values		
	Recall	F_1	κ
2 (Wake, Sleep)	0.824	0.885	0.509
3 (Wake, NREM, REM)	0.787	0.715	0.570
4 (Wake, Light, Deep, REM)	0.732	0.646	0.523
5 (Wake, N1, N2, N3, REM)	0.630	0.586	0.482

TABLE III
RESULTS FOR SLEEP STAGING WITH DIFFERENT NUMBER OF CLASSES AND DIFFERENT INPUT SIGNALS ON THE SMART RING DATASET

Number of classes	Classifier input	Macro-average values		
		Recall	F_1	κ
2 (Wake, Sleep)	IHR	0.849	0.905	0.414
	ZCM	0.839	0.897	0.389
	IHR & ZCM	0.866	0.921	0.469
3 (Wake, NREM, REM)	IHR	0.805	0.762	0.653
	ZCM	0.632	0.522	0.286
	IHR & ZCM	0.832	0.770	0.662
4 (Wake, Light, Deep, REM)	IHR	0.750	0.729	0.635
	ZCM	0.525	0.453	0.279
	IHR & ZCM	0.772	0.738	0.647
5 (Wake, N1, N2, N3, REM)	IHR	0.663	0.636	0.554
	ZCM	0.454	0.403	0.265
	IHR & ZCM	0.671	0.637	0.555

TABLE IV
CONFUSION MATRIX FOR 2-CLASS CLASSIFICATION (WAKE (W) VS SLEEP (S))

Actual \ Predicted	W	S	Recall
W	**1,185**	180	0.868
S	1,885	**12,045**	0.865
Precision	0.366	0.985	

1) Two-Class Classification: The results for two-class classification task indicate that the combination of IHR & ZCM outperforms the individual models in all metrics, achieving the highest macro-average recall (0.866), F_1 score (0.921), and κ value (0.469). The improvement observed when combining IHR and ZCM suggests that integrating both heart rate and motion-based features enhances the network's ability to distinguish between wake and sleep states more accurately.

To further analyze the classification performance, the confusion matrix for the IHR & ZCM model is presented in Table IV. The matrix shows the number of correctly and incorrectly classified instances for both wake and sleep.

From the confusion matrix, it is evident that the model correctly classified 1,185 wake epochs while misclassifying 180 wake epochs as sleep. Similarly, for the sleep class, 12,045 epochs were correctly classified, whereas 1,885 sleep epochs were misclassified as wake. The class-wise recall (accuracy) for wake and sleep were computed as 0.868 (wake) and 0.865 (sleep). This indicates that the model performs comparably well for both classes.

TABLE V
CONFUSION MATRIX FOR 3-CLASS CLASSIFICATION (WAKE (W) VS NREM (N) VS REM (R))

Actual \ Predicted	W	N	R	Recall
W	**1,126**	161	78	0.825
N	881	**9,126**	825	0.843
R	157	377	**2,564**	0.828
Precision	0.520	0.944	0.740	

2) Three-Class Classification: The results for the three-class classification task indicate that the IHR & ZCM model achieves the highest classification performance across all metrics, with a macro-average recall of 0.832, an F_1 of 0.770, and a κ value of 0.662. In comparison, the IHR-only model achieves slightly lower performance, with a macro-average recall of 0.805 and a κ value of 0.653. The ZCM-only model performs the worst among the three, demonstrating notably lower classification performance, with a macro-average recall of 0.632 and a κ value of 0.286. These findings suggest that the combination of IHR and ZCM features improves classification performance, providing a more comprehensive representation of sleep stages compared to using either input modality alone.

The confusion matrix for the IHR & ZCM model shown in Table V provides further insight into the classification results. Among the wake epochs, 1,126 were correctly classified, while 161 were misclassified as NREM sleep and 78 as REM sleep. NREM sleep exhibited the highest number of correctly classified epochs, with 9,126 instances correctly identified, but 881 were misclassified as wake and 825 as REM sleep. Similarly, REM sleep had 2,564 correctly classified epochs, but 157 were misclassified as wake and 377 as NREM sleep. These results indicate that some degree of confusion exists between NREM and REM sleep, which is expected given the natural transitions between these stages.

The class-wise recall values further highlight the effectiveness of the IHR & ZCM model in distinguishing between wake, NREM, and REM sleep stages. The recall values for each class were 0.825 for wake, 0.843 for NREM sleep, and 0.828 for REM sleep, indicating that the model achieves a balanced classification performance across all sleep stages. The κ values for each class were 0.594 for wake, 0.670 for NREM sleep, and 0.722 for REM sleep, suggesting that the model's ability to differentiate REM sleep from other stages is the most reliable, while wake stage classification is the least robust.

The hypnograms shown in Fig. 3 represent the sleep stage

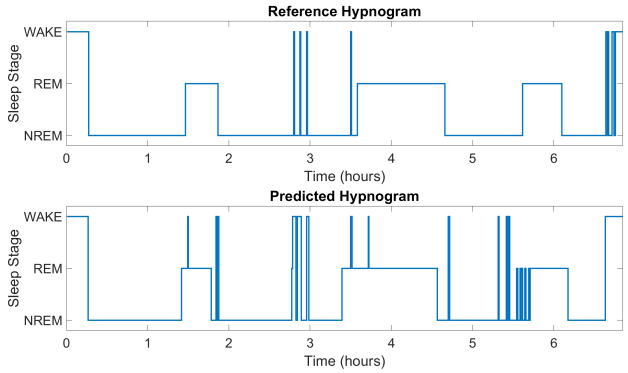

Fig. 3. Actual and predicted hypnograms of a subject in three-class classification.

TABLE VI
CONFUSION MATRIX FOR 4-CLASS CLASSIFICATION (WAKE (W) VS LIGHT SLEEP (L) VS DEEP SLEEP (D) VS REM (R))

Actual \ Predicted	W	L	D	R	Recall
W	**975**	212	63	115	0.714
L	634	**5,723**	936	649	0.721
D	35	401	**2,424**	30	0.839
R	120	453	8	**2,517**	0.812
Precision	0.553	0.843	0.706	0.760	

transitions over time for a subject in the three-class classification task. The top panel illustrates the reference hypnogram, which corresponds to the ground truth sleep staging provided by human expert. The bottom panel presents the predicted hypnogram generated by the IHR & ZCM sleep staging model.

A comparison of the two hypnograms reveals that the model is able to capture the general structure of sleep stage transitions, including prolonged periods of wakefulness at the beginning and intermittent transitions between REM and NREM sleep. However, discrepancies exist, particularly in the form of misclassifications where brief awakenings or REM episodes are either over-predicted or missed. Notably, the predicted hypnogram exhibits more fragmented wake episodes, suggesting that the model may struggle with distinguishing transient wake states from sleep. Despite these differences, the model aligns well with the reference in major sleep stage transitions, indicating its overall effectiveness in automated sleep staging.

3) Four-Class Classification: The results for the four-class classification task demonstrate that the IHR & ZCM model achieves the highest classification performance, with a macro-average recall of 0.772, an F_1 of 0.738, and a κ value of 0.647. In comparison, the IHR-only model achieves a macro-average recall of 0.750 and a κ value of 0.635, whereas the ZCM-only model performs significantly worse, with a macro-average recall of 0.525 and a κ value of 0.279. These findings reinforce the earlier observations that combining IHR and ZCM features leads to more accurate sleep staging performance compared to using either feature set independently.

The confusion matrix shown in Table VI further illustrates the classification performance of the IHR & ZCM model. Among the wake epochs, 975 were correctly classified, while 212 were misclassified as light sleep, 63 as deep sleep, and 115 as REM sleep. Light sleep had 5,723 correctly classified epochs, but 634 were misclassified as wake, 936 as deep sleep, and 649 as REM sleep. The classification of deep sleep was the most robust, with 2,424 correctly classified epochs, though 401 were misclassified as light sleep, 35 as wake, and 30 as REM sleep. Similarly, REM sleep exhibited strong classification performance, with 2,517 correctly identified epochs, though 120 were misclassified as wake, 453 as light sleep, and 8 as deep sleep.

Examining the class-wise recall values provides additional insight into the classification performance. The recall values for wake, light sleep, deep sleep, and REM sleep using the IHR & ZCM model were 0.714, 0.721, 0.839, and 0.812, respectively. This suggests that the model performs particularly well in distinguishing deep sleep and REM sleep, while classification of wake and light sleep is comparatively more challenging. The κ values for each class were 0.581 for wake, 0.572 for light sleep, 0.707 for deep sleep, and 0.729 for REM sleep, indicating that REM sleep and deep sleep classification are the most reliable, whereas wake and light sleep classification are more prone to errors.

These results indicate that light sleep is the most frequently misclassified stage, particularly being confused with both wake and deep sleep. This is expected, as light sleep serves as a transitional stage between wakefulness and deeper sleep stages. The model shows strong performance in differentiating deep sleep and REM sleep, which aligns with their distinct physiological characteristics. However, classification errors between wake and light sleep suggest that further refinements to feature extraction or model architecture may help improve the accuracy of distinguishing these closely related stages.

4) Five-Class Classification: In five-class sleep staging, as expected, the IHR & ZCM model achieves the highest classification performance, with a macro-average recall of 0.671, an F_1 score of 0.637, and a κ value of 0.555. The IHR-only model performs slightly worse, achieving a macro-average recall of 0.663 and a κ value of 0.554, while the ZCM-only model struggles significantly, with a macro-average recall of just 0.454 and a κ of 0.265. These results suggest that ZCM features alone are not sufficient for accurate five-class sleep staging, but their combination with IHR features provides marginal improvement over using IHR alone.

The confusion matrix shown in Table VII further illustrates classification performance and misclassification trends. Among wake epochs, 870 were correctly classified, but 334 were misclassified as N1, and smaller numbers were misclassified as N2 (59), N3 (44), and REM (58). The poor classification of N1 is evident, as 449 epochs were correctly classified,

TABLE VII
CONFUSION MATRIX FOR 5-CLASS CLASSIFICATION (WAKE (W) VS NREM 1 (N1) VS NREM 2 (N2) VS NREM 3 (N3) VS VS REM (R))

Actual \ Predicted	W	N1	N2	N3	R	Recall
W	**870**	334	59	44	58	0.637
N1	130	**449**	194	39	89	0.498
N2	155	1,194	**4,314**	834	544	0.613
N3	13	63	493	**2,287**	34	0.791
R	88	167	255	70	**2,518**	0.813
Precision	0.693	0.203	0.812	0.699	0.776	

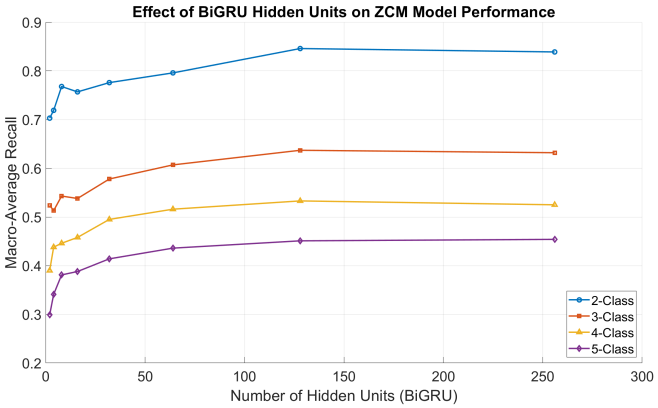

Fig. 4. Macro-average recall as a function of BiGRU hidden units in the ZCM model.

but 130 were misclassified as wake, 194 as N2, 39 as N3, and 89 as REM. The difficulty in differentiating N1 from adjacent stages (W and N2) is expected, as N1 is a transitional stage between wakefulness and deeper sleep. N2 sleep, which is the most frequent sleep stage, was correctly classified in 4,314 instances but often confused with N1 (1,194 misclassified epochs) and N3 (834 misclassified epochs). Meanwhile, N3 sleep was classified with high accuracy, with 2,287 correctly classified epochs, though some misclassifications occurred with N2 (493 epochs) and a few with N1 (63) and REM (34). REM sleep was also well classified, with 2,518 correctly labeled epochs, though 255 were misclassified as N2 and 167 as N1.

A deeper look at class-wise recall values shows variations in classification accuracy across different sleep stages. The recall values for W, N1, N2, N3, and REM using the IHR & ZCM model were 0.637, 0.498, 0.613, 0.791, and 0.813, respectively. This indicates that N3 (deep sleep) and REM sleep were classified with the highest accuracy, whereas N1 (lightest sleep stage) showed the lowest recall at just 0.498, confirming that N1 remains the most challenging stage to classify accurately. The κ values for each class further highlight this difficulty, with N1 achieving the lowest κ at 0.224, whereas N3 and REM showed the highest reliability with κ values of 0.677 and 0.740, respectively.

D. Effect of BiGRU Hidden Units on ZCM Performance

Although the ZCM model takes a single scalar input per epoch (1×1), the BiGRU does not operate on this value in isolation. Instead, it models the sequential relationship of ZCM values across epochs, thereby capturing temporal dependencies that reflect activity patterns over time. To evaluate the impact of model capacity, we varied the number of hidden units in the BiGRU from 2 to 256. As shown in Fig. 4, performance generally improves with increasing hidden units, with macro-average recall rising from 2 units to 128 units, and stabilizing at 256 units. This demonstrates that the BiGRU can successfully leverage low-dimensional ZCM inputs by modeling their temporal evolution across epochs, and that a larger hidden state enhances its ability to capture these dependencies.

E. Effect of Pretraining IHR Model

To evaluate the impact of pretraining on a large external dataset, we compared the performance of the IHR-based CNN-BiGRU model when trained from scratch on the smart ring dataset versus when pretrained on the MGH dataset (ECG-derived IHR) and subsequently fine-tuned on the smart ring dataset (PPG-derived IHR). As shown in Table VIII, training from scratch on the small smart ring dataset achieved macro-average recall values of 0.819, 0.736, 0.682, and 0.588 for two-, three-, four-, and five-class sleep staging tasks, respectively. When the model was pretrained on the MGH dataset and fine-tuned, performance improved consistently across all tasks, reaching macro-average recall values of 0.849, 0.805, 0.750, and 0.663, respectively.

These results clearly demonstrate that pretraining on a large ECG-derived IHR dataset provides a strong initialization that transfers effectively to PPG-derived IHR, significantly improving performance in subject-independent cross-validation on the smart ring dataset. This highlights the importance of transfer learning for addressing data scarcity in wearable sleep staging research.

F. IHR Model Performance with Different CNN and RNN Architectures

To evaluate architectural trade-offs, we experimented with both CNN backbones (ResNet50 [30], Xception [31], and the proposed lightweight 1-D CNN) and RNN variants (LSTM, GRU, and their bidirectional extensions), the results for which are presented in Table IX. ResNet50 and Xception are 2-D CNNs which were transformed to 1-D CNN for IHR signal classification as per the procedure described in [13].

In terms of macro-average recall, the proposed 1-D CNN-BiGRU consistently outperformed the alternatives across all classification granularities, achieving 0.849, 0.805, 0.750, and 0.663 for the two-, three-, four-, and five-class tasks, respectively. While deeper CNN backbones such as ResNet50-BiGRU and Xception-BiGRU provided competitive performance (e.g., recalls of 0.733–0.747 in four-class staging),

TABLE VIII
RESULTS FOR SLEEP STAGING WITH SMART RING-DERIVED IHR MODEL WITH AND WITHOUT PRETRAINING

Model	2-Class			3-Class			4-Class			5-Class		
	Recall	F_1	κ	**Recall**	F_1	κ	**Recall**	F_1	κ	**Recall**	F_1	κ
Without pretraining	0.819	0.905	0.392	0.736	0.689	0.543	0.682	0.651	0.528	0.588	0.557	0.452
With pretraining	0.849	0.905	0.414	0.805	0.762	0.653	0.750	0.729	0.635	0.663	0.636	0.554

TABLE IX
RESULTS FOR SLEEP STAGING WITH SMART RING-DERIVED IHR AND DIFFERENT CNN AND RNN ARCHITECTURES

Model	2-Class			3-Class			4-Class			5-Class		
	Recall	F_1	κ	**Recall**	F_1	κ	**Recall**	F_1	κ	**Recall**	F_1	κ
ResNet50-BiGRU	0.812	0.903	0.381	0.746	0.720	0.593	0.733	0.714	0.614	0.621	0.591	0.497
Xception-BiGRU	0.809	0.886	0.346	0.726	0.718	0.596	0.747	0.736	0.644	0.655	0.630	0.545
CNN-LSTM	0.836	0.913	0.424	0.764	0.740	0.612	0.718	0.700	0.593	0.620	0.591	0.495
CNN-BiLSTM	0.846	0.899	0.398	0.759	0.740	0.616	0.731	0.708	0.605	0.635	0.599	0.507
CNN-GRU	0.814	0.891	0.359	0.732	0.711	0.580	0.722	0.706	0.603	0.643	0.619	0.532
CNN-BiGRU (Proposed)	0.849	0.905	0.414	0.805	0.762	0.653	0.750	0.729	0.635	0.663	0.636	0.554

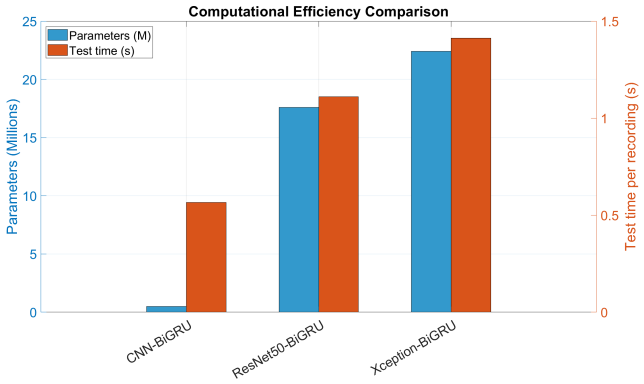

Fig. 5. Comparison of model complexity and efficiency across CNN-BiGRU, ResNet50-BiGRU, and Xception-BiGRU.

they did not surpass the proposed lightweight CNN-BiGRU, highlighting the importance of balancing accuracy with model efficiency for wearable deployment.

To assess the computational efficiency of the proposed framework, we compared the number of learnable parameters and the inference time per recording against deeper CNN-based variants (ResNet50-BiGRU and Xception-BiGRU). All models were implemented in MATLAB and evaluated on an NVIDIA GeForce RTX 4060 GPU to ensure consistency.

Fig. 5 presents the results. The proposed CNN-BiGRU contains only 0.503 M parameters and requires 0.57 seconds per recording during testing. In contrast, ResNet50-BiGRU and Xception-BiGRU are substantially heavier, with 17.6 M and 22.4 M parameters, and testing times of 1.11 s and 1.41 s per recording, respectively.

These results highlight the efficiency advantage of the proposed model. Despite having far fewer parameters and lower inference time, CNN-BiGRU achieved comparable or superior sleep staging performance relative to the deeper CNN-based models. This balance between accuracy and deployability makes it particularly suitable for integration into resource-constrained wearable devices such as smart rings.

G. Effect of Fusion Strategy

To investigate how multimodal fusion affects sleep staging performance, we compared the simple feature-level concatenation used in the main experiments (late fusion) with a cross-attention fusion module that explicitly models interactions between the IHR and ZCM embeddings. All fusion experiments used the same leave-one-subject-out cross-validation protocol, data splits, and training hyperparameters as the main study to ensure a fair comparison.

Using the original late-fusion scheme (independent IHR and ZCM branches whose 128-dimensional activations are concatenated and classified), the macro-average recall was 0.866, 0.832, 0.772, and 0.671 for the two-, three-, four-, and five-class tasks, respectively. Replacing the simple concatenation with a cross-attention fusion module produced marginal changes with macro-average recall values of 0.867, 0.836, 0.758, and 0.667 for the two-, three-, four-, and five-class tasks, respectively. In other words, cross-attention yielded a slight improvement in the coarser two- and three-class problems, but it led to small degradations in the finer four- and five-class settings. Importantly, the cross-attention module introduces additional parameters and computational overhead. Given these trade-offs and the relatively small performance gains for the two- and three-class tasks, we retain the simpler feature concatenation approach in the main experiments.

We considered the option of early fusion, but this was judged impractical in our current framework because the IHR branch is pretrained on a large ECG-derived IHR dataset and expects a single-channel input. Early fusion would change the input statistics and require reinitializing or retraining the convolutional front end, which would negate the benefits of pretraining. For scenarios where pretraining is not used, or where sufficient paired multimodal training data exist, early fusion and hybrid architectures remain promising avenues for future work.

TABLE X
RESULTS FOR SLEEP STAGING USING BASELINE AND PROPOSED METHODS WITH IHR AND ZCM

Classifier	2-Class			3-Class			4-Class			5-Class		
	Recall	F_1	κ	**Recall**	F_1	κ	**Recall**	F_1	κ	**Recall**	F_1	κ
RF[a]	0.777	0.861	0.284	0.631	0.573	0.375	0.560	0.511	0.343	0.456	0.410	0.272
SVM[a]	0.746	0.831	0.229	0.650	0.584	0.390	0.579	0.503	0.341	0.480	0.419	0.289
MLP[a]	0.798	0.879	0.326	0.635	0.571	0.374	0.569	0.494	0.328	0.465	0.400	0.266
BiGRU[a]	0.795	0.878	0.322	0.668	0.644	0.465	0.620	0.589	0.443	0.533	0.505	0.384
Fusion[a]	0.848	0.916	0.442	0.783	0.699	0.552	0.702	0.649	0.525	0.609	0.566	0.463
Fusion[b]	0.866	0.921	0.469	0.832	0.770	0.662	0.772	0.738	0.647	0.671	0.637	0.555

[a]Trained from scratch.
[b]IHR model is pretrained.

H. Results Using Baseline and Proposed Methods

Table X presents the results of various baseline methods, inspired from earlier works [8], [9], [13], [32], [33], under a leave-one-subject-out cross-validation scheme, consistent with the evaluation of the proposed framework. The classical feature-based approaches (random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP)), which relied on handcrafted time- and frequency-domain features from the IHR signal as described in [13] together with the ZCM feature, generally underperformed across all staging tasks, with macro-average recall falling below 0.5 for five-class classification. The feature-based BiGRU showed modest improvements, particularly in three-class and four-class staging, but remained limited in discriminative ability.

The proposed fusion model when trained from scratch, that is, without pretraining, outperformed these feature-engineered models, highlighting the benefit of joint representation learning directly from IHR signals. Importantly, while all these models were trained from scratch, the proposed fusion model where the IHR-based CNN-BiGRU is pretrained achieved the best performance across all staging granularities, underscoring the effectiveness of leveraging pretraining for robust sleep stage classification from smart ring signals.

IV. DISCUSSION AND CONCLUSION

The proposed study investigates sleep staging using a lightweight deep learning framework based on IHR and ZCM features extracted from PPG and accelerometer signals recorded by a smart ring. The results demonstrate that the classification of deep sleep (N3) and REM sleep is the most reliable, while N1 sleep remains the most challenging to distinguish. This is consistent with previous sleep staging studies [13], where N1 is often misclassified due to its transitional nature between wakefulness and deeper sleep stages. In our dataset, N1 represents only 5.9% of all epochs, making it the least represented class. This class imbalance likely also contributes to the reduced performance in N1 detection. The lower performance of the ZCM-based model alone, particularly in three-, four-, and five-class staging, further highlights the importance of IHR in improving classification accuracy. Moreover, combining IHR and ZCM features provides only a marginal improvement over the IHR-only model, suggesting that IHR is the dominant feature in sleep staging.

A general trend observed in the classification results is the increased misclassification rate between adjacent sleep stages. In the five-class classification task, misclassifications frequently occur between neighboring sleep stages, such as N1 and N2 or N2 and N3, reflecting the gradual and continuous nature of sleep transitions. The results also indicate that wakefulness is often confused with N1 sleep, which aligns with the challenge of distinguishing brief arousals from sustained wakefulness in sleep scoring. Similarly, REM sleep is occasionally misclassified as N2 sleep, likely due to similar autonomic and movement-related characteristics in both stages.

The classification agreement between the proposed models and the ground truth sleep labels was evaluated using κ. In the five-class classification task, we achieve κ values of 0.632, 0.224, 0.500, 0.677, and 0.740 for wake, N1, N2, N3, and REM sleep, respectively. Compared to a meta-analysis on manual sleep staging [34], which reported κ values of 0.70, 0.24, 0.57, 0.57, and 0.69 for wake, N1, N2, N3, and REM sleep, respectively, the agreement interpretation of our algorithm matches that of manual scoring for wake (substantial agreement), N1 (fair argeement), N2 (moderate agreement), and REM (substantial agreement). For the N3 class, our algorithm achieves substantial agreement, whereas manual scoring achieves only moderate agreement.

When comparing the performance of the IHR network on the ECG-derived pretraining dataset and the PPG-derived fine-tuning dataset from the smart ring, several key differences emerge. While the overall trends are similar, with performance decreasing as the number of sleep stage classes increases, the smart ring dataset consistently outperforms the pretraining dataset across all tasks. For instance, in the five-class task, the macro-average recall improved from 0.630 to 0.663, F_1 score increased from 0.586 to 0.636, and κ value rose from 0.482 to 0.554. These improvements are particularly interesting given that the pretraining dataset (MGH) consisted of clinical sleep studies, potentially including subjects with various sleep disorders, while the smart ring dataset comprised only healthy individuals. The cleaner and more stable sleep patterns in the healthy population, along with the domain-specific fine-tuning, likely contributed to the observed performance gains.

Although the proposed framework leverages pretraining on ECG-derived IHR and fine-tuning on PPG-derived IHR, the

modality mismatch is inherently limited because IHR reflects approximately the same physiological process (beat-to-beat intervals) irrespective of the acquisition modality. R-peaks (ECG) and pulse peaks (PPG) were detected using established algorithms, and the resulting IBIs were resampled at 2 Hz and z-score normalized. This preprocessing step harmonized scale and temporal resolution while largely eliminating modality-specific morphology, allowing the pretrained model to transfer effectively across ECG- and PPG-derived IHR without the need for explicit domain adaptation.

The models developed for sleep staging are computationally efficient, with the IHR model containing 503 k learnable parameters and the ZCM model containing 133 k learnable parameters. Although current smart rings typically rely on cloud-based processing for sleep staging, the lightweight neural networks developed in this work offer advantages even in this setup. Their compact architecture enables faster inference, reduced latency, and lower computational costs, making them well-suited for scalable cloud deployment, particularly important when serving many users simultaneously. Smaller models also streamline deployment, updates, and maintenance, enabling rapid iteration and continuous model improvement. Moreover, their efficiency supports more sustainable computing practices by minimizing energy usage on cloud servers. As smart ring hardware continues to evolve, there is potential for on-device sleep staging. In such future scenarios, the lightweight nature of these models would be even more advantageous, enabling efficient, real-time analysis directly on the device without sacrificing battery life or performance.

The comparative analysis against prior wearable-based sleep staging studies in Table XI highlights the effectiveness of our proposed method, particularly in the context of healthy subject populations. In two-class classification, our model using heart rate alone (macro-average recall or UAR: 0.849) already outperforms earlier works such as [8] (0.754) and [32] (0.820), wrist and finger-worn devices, respectively, both of which used a wider range of physiological features. The addition of accelerometer-derived ZCM data further improves our model's performance to 0.866, suggesting the complementary nature of heart rate and movement features and the advantage of our lightweight neural network architecture.

In the three-class task, a similar trend is observed. While [8] reported a macro-average recall of 0.615 and [35] reported 0.670 on the same dataset, our method achieves 0.805 with heart rate alone and improves further to 0.832 when acceleration data are included. These improvements reinforce the value of using tailored neural network architectures and fine-tuning on specific wearable signal modalities like PPG-derived IHR, especially when compared to traditional classifiers.

The four-class classification task further demonstrates the strength of our approach. Prior studies using devices like a Happy Ring (0.670) [32], Fitbit Surge (0.681) [36], Actiwatch (0.686) [37], Samsung smartwatch (0.717) [9], and wrist-worn wearable (0.746) [33] show lower macro-average recall values. Our heart rate-only model surpasses these with a score of 0.750, and the inclusion of acceleration features lifts performance to 0.772. This is noteworthy considering our model is trained on a smaller dataset of 18 recordings but still outperforms studies with much larger subject cohorts, indicating strong model generalizability and robustness. Notably, [33] utilized a model pretrained on ECG-derived IHR features, achieving slightly lower performance than ours, underscoring the importance of pretraining on high-quality ECG-based signals for downstream tasks using wearable-derived data. While earlier works rely on handcrafted features extracted from the heart rate signal, our method leverages a CNN to learn heart rate variability patterns directly from the IHR signal, enabling more effective representation learning.

Finally, in the five-class setting, our study is among the first to report macro-average recall values using smart ring data, achieving 0.663 with heart rate alone and 0.671 with the addition of acceleration. While direct comparisons with past work are not possible due to lack of available results in this class structure, these values establish a strong baseline for future research in fine-grained sleep staging using ring-based wearables.

The ability to accurately classify sleep stages using PPG and accelerometer signals has significant potential applications. The proposed lightweight deep learning models can be deployed in wearable devices such as smart rings to enable unobtrusive, continuous sleep monitoring. This could provide valuable insights into sleep patterns, aiding in the diagnosis and management of sleep disorders. Additionally, integrating these models into consumer wearables could facilitate large-scale sleep studies by providing reliable sleep stage estimates without the need for cumbersome polysomnography recordings.

Despite its promising performance, this study has certain limitations. The dataset consisted of only 18 recordings from 9 healthy subjects, which restricts the generalizability of the findings, particularly to clinical populations with sleep disorders or cardiovascular conditions that may influence heart rate and movement patterns during sleep. Nonetheless, our evaluation employed subject-independent cross-validation, ensuring that the reported results reflect robustness across unseen subjects rather than overfitting to individuals. Furthermore, the proposed CNN-BiGRU demonstrated good performance when pretrained and validated on a large ECG-derived IHR dataset that included subjects with sleep disorders, suggesting potential applicability beyond healthy cohorts. Future work should focus on validating the framework on larger and more diverse populations, including individuals with insomnia, sleep apnea, and other sleep disorders. Additionally, incorporating complementary physiological signals, such as respiratory effort or electrodermal activity, could further improve classification performance.

While the bidirectional GRU improves classification accuracy by leveraging both past and future context, its use raises deployability concerns since future data is unavailable in real-time streaming. However, real-time sleep staging may not always be necessary for smart ring applications, where the primary goal is nightly sleep assessment and trend monitoring rather than on-the-fly interventions. In such cases, data can be processed retrospectively, allowing bidirectional models to be deployed without affecting usability. For scenarios that demand real-time detection, unidirectional models may be

TABLE XI
SUMMARY OF STUDY CHARACTERISTICS AND CLASSIFICATION PERFORMANCE (MACRO-AVERAGE RECALL OR UNWEIGHTED AVERAGE RECALL (UAR)) OF STUDIES USING WEARABLES FOR SLEEP STAGING

Number of classes	Reference	Method	UAR
2 (Wake, Sleep)	Walch *et al.* 2019 [8]	Device: Apple watch No. of subjects: 31 (healthy) Features: heart rate and acceleration Classifier: neural network	0.754
	Grandner *et al.* 2022 [32]	Device: Happy ring No. of subjects: 36 (healthy) Features: heart rate, acceleration, electrodermal activity, and skin temperature Classifier: random forest	0.820
	This work	Device: SOXAI smart ring No. of subjects (recordings): 9 (18) (healthy) Input: heart rate Classifier: neural network	0.849
	This work	Same as above but with input: heart rate and acceleration	0.866
3 (Wake, NREM, REM)	Walch *et al.* 2019 [8]	Same as 2-class classification	0.615
	Zhai *et al.* 2022 [35]	Device: Apple watch No. of subjects: 31 (healthy) Features: heart rate and acceleration Classifier: neural network	0.670
	This work	Same as 2-class classification (heart rate only)	0.805
	This work	Same as 2-class classification (heart rate and acceleration)	0.832
4 (Wake, Light, Deep, REM)	Beattie *et al.* 2017 [36]	Device: Fitbit Surge No. of subjects: 60 (healthy) Features: heart rate and acceleration Classifier: linear discriminant analysis	0.681
	Radha *et al.* 2021 [33]	Device: Wrist-worn wearable (Royal Philips) No. of subjects: 60 (healthy) Features: heart rate Classifier: RNN	0.746
	Grandner *et al.* 2022 [32]	Same as 2-class classification	0.670
	Liu *et al.* 2023 [37]	Device: Actiwatch No. of subjects: 75 (healthy) Features: heart rate and acceleration Classifier: linear discriminant analysis	0.686
	Silva *et al.* 2024 [9]	Device: Samsung smartwatch No. of subjects: 1430 (healthy and sleep apnea) Features: heart rate and acceleration Classifier: RNN	0.717
	This work	Same as 2-class classification (heart rate only)	0.750
	This work	Same as 2-class classification (heart rate and acceleration)	0.772
5 (Wake, N1, N2, N3, REM)	This work	Same as 2-class classification (heart rate only)	0.663
	This work	Same as 2-class classification (heart rate and acceleration)	0.671

more appropriate despite slightly reduced accuracy. Thus, the choice between uni- and bidirectional models reflects a trade-off with bidirectional models offering higher staging performance in retrospective monitoring while unidirectional models enabling real-time applications.

In conclusion, this work demonstrates that lightweight neural network models using heart rate and movement signals from smart rings can achieve strong performance in sleep staging, particularly in coarser classification tasks like two-, three-, and four-class staging. Our models consistently outperform previous wearable-based approaches in these settings, highlighting the benefits of using PPG-derived IHR with data-driven neural representations. While performance in the more granular five-class task is more limited, the results nonetheless establish a useful baseline for future work with smart ring data. These findings suggest that compact, efficient models can enable accurate and scalable sleep monitoring using consumer-grade wearables, with potential applications in both population-level studies and individual health tracking. Further research incorporating more diverse populations and additional physiological signals may help close the gap in finer sleep stage differentiation.

REFERENCES

[1] M. M. Pankowska, H. Lu, A. G. Wheaton, Y. Liu, B. Lee, K. J. Greenlund, and S. A. Carlson, "Prevalence and geographic patterns of self-reported short sleep duration among US adults, 2020," *Preventing Chronic Disease*, vol. 20, no. 2, 2023, Art. no. 220400.

[2] D. C. Lim, A. Najafi, L. Afifi, C. L. Bassetti, D. J. Buysse, F. Han, B. Högl, Y. A. Melaku, C. M. Morin, A. I. Pack, D. Poyares, V. K. Somers, P. R. Eastwood, P. C. Zee, and C. L. Jackson, "The need to promote sleep health in public health agendas across the globe," *The Lancet Public Health*, vol. 8, no. 10, pp. e820–e826, 2023.

[3] J. T. Maurer, "Early diagnosis of sleep related breathing disorders," *GMS Current Topics in Otorhinolaryngology - Head and Neck Surgery*, vol. 7, 2008, Art. no. Doc03.

[4] D. Shrivastava, S. Jung, M. Saadat, R. Sirohi, and K. Crewson, "How to interpret the results of a sleep study," *Journal of Community Hospital Internal Medicine Perspectives*, vol. 4, no. 5, 2014, Art. no. 24983.

[5] D.-W. Chang, Y.-D. Liu, C.-P. Young, J.-J. Chen, Y.-H. Chen, C.-Y. Chen, Y.-C. Hsu, F.-Z. Shaw, and S.-F. Liang, "Design and implementation of a modularized polysomnography system," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 7, pp. 1933–1944, 2012.
[6] S. Djanian, A. Bruun, and T. D. Nielsen, "Sleep classification using consumer sleep technologies and AI: A review of the current landscape," *Sleep Medicine*, vol. 100, pp. 390–403, 2022.
[7] S.-F. Liang, C.-E. Kuo, Y.-C. Lee, W.-C. Lin, Y.-C. Liu, P.-Y. Chen, F.-Y. Cherng, and F.-Z. Shaw, "Development of an EOG-based automatic sleep-monitoring eye mask," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 11, pp. 2977–2985, 2015.
[8] O. Walch, Y. Huang, D. Forger, and C. Goldstein, "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device," *Sleep*, vol. 42, no. 12, 2019, Art. no. zsz180.
[9] F. B. Silva, L. F. Uribe, F. X. Cepeda, V. F. Alquati, J. P. Guimarães, Y. G. Silva, O. L. dos Santos, A. A. de Oliveira, G. H. de Aguiar, M. L. Andersen, S. Tufik, W. Lee, L. T. Li, and O. A. Penatti, "Sleep staging algorithm based on smartwatch sensors for healthy and sleep apnea populations," *Sleep Medicine*, vol. 119, pp. 535–548, 2024.
[10] Y.-H. Wang, I.-Y. Chen, H. Chiueh, and S.-F. Liang, "A low-cost implementation of sample entropy in wearable embedded systems: An example of online analysis for sleep EEG," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021, Art. no. 4002412.
[11] Z. Wang, R. Yu, X. Wang, J. Ding, J. Tang, J. Fang, Z. He, Z. Li, T. Röddiger, W. Xu, X. Zhang, huan-ang Gao, N. Gao, C. Yu, Y. Shi, and Y. Wang, "Computing with smart rings: A systematic literature review," 2025. [Online]. Available: https://arxiv.org/abs/2502.02459
[12] W. Huang, L. Zhang, W. Gao, F. Min, and J. He, "Shallow convolutional neural networks for human activity recognition using wearable sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021, Art. no. 2510811.
[13] R. V. Sharan, H. Takeuchi, A. Kishi, and Y. Yamamoto, "Macro-sleep staging with ECG-derived instantaneous heart rate and respiration signals and multi-input 1-D CNN–BiGRU," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024, Art. no. 2535212.
[14] A. Kishi, Z. R. Struzik, B. H. Natelson, F. Togo, and Y. Yamamoto, "Dynamics of sleep stage transitions in healthy humans and patients with chronic fatigue syndrome," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 294, no. 6, pp. R1980–R1987, 2008.
[15] A. Kishi, H. Yasuda, T. Matsumoto, Y. Inami, J. Horiguchi, M. Tamaki, Z. R. Struzik, and Y. Yamamoto, "NREM sleep stage transitions control ultradian REM sleep rhythm," *Sleep*, vol. 34, no. 10, pp. 1423–1432, 2011.
[16] M. W. Johns, "A new method for measuring daytime sleepiness: The Epworth sleepiness scale," *Sleep*, vol. 14, no. 6, pp. 540–545, 1991.
[17] M. Elgendi, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans, "Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions," *PLOS ONE*, vol. 8, no. 10, 2013, Art. no. e76585.
[18] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms," *Sleep*, vol. 26, no. 3, pp. 342–392, 2003.
[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.
[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.
[21] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2146–2153.
[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proceedings of the NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.
[23] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, 2019.
[24] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
[25] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
[26] D. Zhou, Q. Xu, J. Wang, H. Xu, L. Kettunen, Z. Chang, and F. Cong, "Alleviating class imbalance problem in automatic sleep stage classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022, Art. no. 4006612.
[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015.
[28] M. M. Ghassemi, B. E. Moody, L.-W. H. Lehman, C. Song, Q. Li, H. Sun, R. G. Mark, M. B. Westover, and G. D. Clifford, "You snooze, you win: The physionet/computing in cardiology challenge 2018," in *Proceedings of the 2018 Computing in Cardiology Conference (CinC)*, vol. 45, 2018, pp. 1–4.
[29] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.
[32] M. A. Grandner, Z. Bromberg, A. Hadley, Z. Morrell, A. Graf, S. Hutchison, and D. Freckleton, "Performance of a multisensor smart ring to evaluate sleep: In-lab and home-based evaluation of generalized and personalized algorithms," *Sleep*, vol. 46, no. 1, 2022, Art. no. zsac152.
[33] M. Radha, P. Fonseca, A. Moreau, M. Ross, A. Cerny, P. Anderer, X. Long, and R. M. Aarts, "A deep transfer learning approach for wearable sleep stage classification with photoplethysmography," *npj Digital Medicine*, vol. 4, 2021, Art. no. 135.
[34] Y. J. Lee, J. Y. Lee, J. H. Cho, and J. H. Choi, "Interrater reliability of sleep stage scoring: A meta-analysis," *Journal of Clinical Sleep Medicine*, vol. 18, no. 1, pp. 193–202, 2022.
[35] B. Zhai, Y. Guan, M. Catt, and T. Plötz, "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*, vol. 5, no. 4, 2022, Art. no. 191.
[36] Z. Beattie, Y. Oyang, A. Statan, A. Ghoreyshi, A. Pantelopoulos, A. Russell, and C. Heneghan, "Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals," *Physiological Measurement*, vol. 38, no. 11, pp. 1968–1979, 2017.
[37] P.-K. Liu, N. Ting, H.-C. Chiu, Y.-C. Lin, Y.-T. Liu, B.-W. Ku, and P.-L. Lee, "Validation of photoplethysmography- and acceleration-based sleep staging in a community sample: Comparison with polysomnography and Actiwatch," *Journal of Clinical Sleep Medicine*, vol. 19, no. 10, pp. 1797–1810, 2023.

Roneel V. Sharan (Senior Member, IEEE) received the Ph.D. degree in engineering from the Auckland University of Technology, Auckland, New Zealand, in 2016. From 2016 to 2019, he was a Postdoctoral Research Fellow at the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia. From 2019 to 2023, he was a Research Fellow with the Australian Institute of Health Innovation, Macquarie University, Sydney, Australia. He is currently a Senior Lecturer at the School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom. His research interests include biomedical signal processing and machine learning.

Hiroki Takeuchi received the B.Sc. degree from Tohoku University, Miyagi, Japan, and M.Sc. and Ph.D. degrees in Education from The University of Tokyo, Tokyo, Japan, respectively in 2018, 2020, and 2024. He is currently a Project Researcher at the Graduate School of Education, University of Tokyo. His current research interests include wearable sleep monitoring and its applications to health science.

Akifumi Kishi received the B.Sc., M.Sc. and Ph.D. degrees in Education from The University of Tokyo, Tokyo, Japan, respectively in 2006, 2008 and 2011. From 2010 to 2014, he was working as a Postdoctoral Fellow at the Sleep Disorders Center at the New York University School of Medicine, New York, NY, USA, and in the Pain & Fatigue Study Center at the Beth Israel Medical Center, New York, NY, USA. From 2014 to 2022, he was an Assistant Professor in the Division of Physical and Health Education, Graduate School of Education, The University of Tokyo. From 2022 to 2025, he served as a Project Lecturer in the Department of Systems Pharmacology, Graduate School of Medicine, The University of Tokyo. Since April 2025, he has been serving as a Lecturer in the same department. He has a long-term interest in understanding the mechanism and function of sleep in humans. His current research involves the analysis, assessment, modeling, and control of the human sleep dynamics.

Jiabin Wang received the B.E. and M.E. degrees in Electronic Engineering from Tsinghua University, Beijing, China, and the Ph.D. degree in Electronic Engineering from the University of Tokyo, Tokyo, Japan. He is currently a Research Engineer with SOXAI Inc., Tokyo, Japan. His research interests include wearable sensors, embedded systems, wireless communication, and the application of machine learning in embedded systems.

Tatsuhiko Watanabe received the B.E., M.E., and Ph.D. degrees in electrical engineering from Yokohama National University, Yokohama, Japan, in 2012, 2013, and 2016, respectively. From 2014 to 2017, he was a Research Fellow of the Japan Society for the Promotion of Science (JSPS), Tokyo, Japan. From 2016 to 2019, he was a Postdoctoral Fellow at the Institute of Electromagnetic Fields, ETH Zurich, Zurich, Switzerland, where he focused on research in the field of nanophotonics. In 2017, he was awarded the ETH Fellow. Since 2021, he has been the founder and CEO of SOXAI Inc., Yokohama, Japan.

Yoshiharu Yamamoto received his B.Sc., M.Sc. and Ph.D. degrees in educational sciences from the University of Tokyo, Tokyo, Japan, respectively in 1984, 1986 and 1990. From 1989 to 1993, he was working as a Postdoctoral Researcher at the Faculty of Applied Health Sciences, University of Waterloo, Waterloo, Ontario, Canada, mainly on bio-signal processing for human cardiovascular and autonomic physiology. In 1993, he was granted a faculty position at the Graduate School of Education, the University of Tokyo, and has been a Full Professor at the Educational Physiology Laboratory, Graduate School of Education, the University of Tokyo since 2000, where he is teaching and researching physiological bases and data analytics in various fields of health sciences. His current research interest includes biomedical signal processing, nonlinear and statistical bio-dynamics, and health informatics. He is currently a senior editor of *IEEE Transactions on Biomedical Engineering* and an editorial board member of *Biomedical Physics & Engineering Express*.