WatchSleepNet: A Novel Model and Pretraining Approach for Advancing Sleep Staging with Smartwatches

Will Ke Wang KE.WANG064@DUKE.EDU
Bill Chen BILL.CHEN@DUKE.EDU
Jiamu Yang JERRY.YANG@DUKE.EDU
Hayoung Jeong HAYOUNG.JEONG@DUKE.EDU
Leeor Hershkovich LEEOR.HERSHKOVICH@DUKE.EDU
Shekh Md Mahmudul Islam SHEKHMDMAHMUDUL.ISLAM@DUKE.EDU
Mengde Liu MENGDE.LIU@DUKE.EDU
Ali R Roghanizad ALI.ROGHANIZAD@DUKE.EDU
Duke University, USA

Md Mobashir Hasan Shandhi MOBASHIR.SHANDHI@ASU.EDU
Arizona State University, USA

Andrew R Spector ANDREW.SPECTOR@DUKE.EDU
Jessilyn Dunn JESSILYN.DUNN@DUKE.EDU
Duke University, USA

Abstract

Sleep monitoring is essential for assessing overall health and managing sleep disorders, yet clinical adoption of consumer wearables remains limited due to inconsistent performance and scarce open source datasets and transparent codebase. In this study, we introduce WatchSleepNet, a novel, open-source three-stage sleep staging algorithm. The model uses sequence-to-sequence architecture integrating Residual Networks (ResNet), Temporal Convolutional Networks (TCN), and Long Short-Term Memory (LSTM) networks with self-attention to effectively capture both spatial and temporal dependencies crucial for sleep staging. To address the limited availability of high-quality wearable photoplethysmography (PPG) datasets, WatchSleepNet leveraged inter-beat interval (IBI) signals as a shared representation across polysomnography (PSG) and photoplethysmography (PPG) modalities. By pretraining on large PSG datasets and fine-tuning on wrist-worn PPG signals, the model achieved a REM F1 score of 0.631 $\pm$ 0.046 and a Cohen's Kappa of 0.554 $\pm$ 0.027, surpassing previous state-of-the-art methods. To promote transparency and further research, we publicly release our model and codebase, advancing reproducibility and accessibility in wearable sleep research and enabling the development for more robust, clinically viable sleep monitoring solutions.

Data and Code Availability The DREAMT dataset used in this study is fully available to the public on PhysioNet/DREAMT. In addition to DREAMT, we utilized the Sleep Heart Health Study (SHHS) and Multi-Ethnic Study of Atherosclerosis (MESA) datasets, both of which are accessible to the public through the National Sleep Research Resource (NSRR) platform. The code repository for this paper is available at GITHUB/WatchSleepNet_public.

Institutional Review Board (IRB) This study was IRB approved with IRB number: #Pro00108961

1. Introduction

Sleep is a critical biological process that plays a vital role in physical health, cognitive function, and emotional well-being. Adequate sleep is essential for numerous physiological processes, including immune function, memory consolidation, and metabolic regulation (de Zambotti et al. (2024)). The ability to accurately assess sleep patterns and stages is crucial for diagnosing sleep disorders, evaluating treatment efficacy, and understanding the relationship between sleep and health outcomes (Birrer et al. (2024); Sung et al. (2024)). The current gold standard for sleep

monitoring is polysomnography (PSG), which involves a comprehensive set of equipments that records brain activity, eye movements, muscle tone, and other physiological signals to determine sleep stages and perform diagnostic tests of sleep disorders (Birrer et al. (2024)). However, PSG is resource-intensive, requiring overnight laboratory visits, specialized equipments, and trained personnel to analyze data. Consequently, it is impractical for continuous monitoring, particularly in natural sleep environments (Peppard et al. (2013)). Wearable devices have emerged as an alternative for sleep monitoring. These devices offer non-invasive, continuous tracking of physiological signals in natural settings, making them an attractive option for both researchers and consumers (Lee (2024)). The rapid growth in consumer sleep technology is evident, with globally-shipped wearable devices increasing from 89 million in 2016 to 232 million in 2021 annually, reached 590.7 million by 2025, and projected to reach 645.7 million units by 2028 (Baumert et al. (2022); Pangarkar (2025)). Wearable devices typically use sensors like accelerometers and photoplethysmography (PPG) to infer sleep stages and calculate sleep metrics such as total sleep time, sleep onset latency, and sleep stages. Accelerometers detect motion to differentiate between rest and activity, while PPG measures blood volume changes to estimate heart rate and heart rate variability (HRV), which contains patterns that could indicate different sleep stages (Sung et al. (2024); Lee (2024)). The current state-of-the-art 4-stage wearable sleep staging algorithm comes from the Oura ring, whose clinical validation study on 96 healthy adults achieved a max Cohen's Kappa of 0.66 (Svensson et al. (2024)). However, the Oura Ring's performance is not directly comparable to that of wrist-based wearables due to differences in form factor and measurement location.

While wearable devices offer a promising approach to sleep staging, they often lack reliability, particularly for detecting REM and deep sleep, when compared to gold-standard PSG (Baumert et al. (2022)). Recent review and validation studies reported that certain consumer wrist-based wearables demonstrate high performance in detecting sleep versus wake, with accuracy reaching 91.5%, Cohen's kappa of 0.66 (Wulterkens et al. (2021)), and sensitivity ranging from 20% to 70% (de Zambotti et al. (2024)). However, in multi-stage classification, where sleep is separated into light, deep, and REM stages, performance declines substantially, with accuracy dropping to 76.4% and Cohen's kappa to 0.62 (Wulterkens et al. (2021)). Multiple reviews show that for multi-stage classification, there is a wide variability in performance (de Zambotti et al. (2024); Birrer et al. (2024); Chiang and Khosla (2023)). For instance, several studies evaluating consumer wearable devices reported epoch-to-epoch agreement rates ranging from 44-71% for light sleep, 28-72% for deep sleep, and 36-66% for REM sleep (Miller et al. (2022); Chinoy et al. (2020, 2022)). This variability is exacerbated in individuals with sleep disorders, such as sleep apnea, where validation studies are few, and in those that do exist, performance declines significantly for sleep stage classification in people with sleep disorders. Birrer et al. (2024)'s review found that only 4 out of 35 studies included both healthy participants and those with sleep disorders. Furthermore, Chiang and Khosla (2023) showed that Fitbit devices overestimated total sleep time by 59.8 minutes and underestimated wake after sleep onset by 36.1 minutes, with low wake specificity (0.44) in an adult population suspected of obstructive sleep apnea. Chiang and Khosla (2023) found that the sensitivity scores of REM sleep and deep sleep are significantly negatively correlated with apnea severity, indicating a decline in performance in sleep-disordered populations. This greatly limits the utility of wrist-based wearable devices for widespread sleep monitoring, particularly in populations with sleep apnea, which is widely prevalent but more than 80% of these patients are undiagnosed (Peppard et al. (2013)). Given the high prevalence of sleep apnea and its impact on sleep health, the need for accurate sleep staging in this population is both critical and particularly challenging (Peppard et al. (2013); Young et al. (1997))

Further hindering research and algorithmic development in wrist-based wearable sleep monitoring is the limited availability of high-quality, publicly accessible datasets and reproducible codebases. Unlike the more extensive and diverse datasets available for polysomnography (PSG), those using wearable devices tend to be smaller and less diverse. Wang et al. (2024) recently published the first of its kind wearable sleep dataset to begin to address this challenge. There is also a general lack of reproducibility in wearable sleep staging algorithm development due to the absence of open-source code, poorly described methods, and lack of standardized performance metrics used. Please refer to Section 1.1 for a detailed explanation. The inability to validate published works in this field makes it difficult to understand how well wearable sleep staging algorithms truly perform

for different users, compare performance between algorithms, and validate those results across different studies and conditions. Open access to data, alongside well-documented code and methodologies, is crucial for driving progress in this field and ensuring that sleep staging algorithms can be reliably evaluated, compared, and improved upon (de Zambotti et al. (2024); Birrer et al. (2024)).

1.1. Relevant Works

Wearable sleep staging research has employed various machine learning methods. A significant number of works utilize deep learning models with photoplethysmography (PPG) signals but lack publicly accessible datasets and available code. For example, Korkalainen et al. (2020); Fonseca et al. (2020, 2023); Heneghan et al. (2024); Silva et al. (2024); Attia et al. (2024), all developed PPG-based models achieving notable performances, but either the data or code or both are unavailable to support replication, benchmarking (comparing model performance across datasets and tasks), and building upon these important works. Olsen et al. (2023) et al.'s sleep staging algorithm demonstrates notable performance using raw PPG and accelerometry data from multiple sources and published their codebase, even though the datasets they used were not available. Radha et al. (2021) and Li et al. (2021) applied transfer learning from ECG to PPG signals. Kotzen et al. (2023) introduced SleepPPG-Net using PPG and IBI signals, achieving state-of-the-art performance. Wulterkens et al. (2021) and Topalidis et al. (2023) reported great sleep staging performances on small or proprietary wrist-watch datasets (Kappa = 0.62 and Kappa = 0.69 respectively). Although these works demonstrate significant promise for advancing wearable sleep staging algorithms, the absence of publicly accessible datasets and open source codebase from these studies hampers widespread reproducibility and further innovation.

Compared to wearable sleep staging algorithms, more studies have been published utilizing ECG-derived IBI or instantaneous heart rate (IHR) signals with larger amounts of accessible data for sleep staging. Notably, Nam et al. (2024) provided both code and data for their InsightSleepNet model using continuous PPG signals, which implements a seq-2-seq sleep stage classifier based on the InceptionTime Module on multiple datasets and achieves Cohen's Kappa scores > 0.74. Also of note, Sridhar et al. (2020) used IBI values calculated from publicly available SHHS and MESA datasets, with their model architecture clearly defined, enabling us to reproduce their methodology despite not having access to their code. We coined their model 'SleepConvNet' in this paper since their algorithm was mainly convolution based. Overall, the scarcity of accessible wrist-based PPG datasets and available code limits the adoption and advancement of existing works, directing our validation and advancement efforts toward works with datasets that were made available and similar overall task definitions or training approach.

1.2. Summary of Study and Contribution

In response to the challenges in wearable-based sleep staging, our study introduces a novel approach that leverages multiple datasets of varying sizes and characteristics, alongside a new deep learning architecture that we call WatchSleepNet. This architecture combines ResNet, Temporal Convolutional Networks (TCN), and Long Short-Term Memory (LSTM) networks with multi-headed attention mechanism for modeling sequential patterns of sleep stages throughout a night. We focus on improving sleep staging performance by utilizing wrist-based PPG signals and IBI values from a research-grade wearable device—the Empatica E4 smartwatch. By integrating insights from large publicly available polysomnography (PSG) datasets such as SHHS (Sleep Heart Health Study), Zhang et al. (2018); Quan et al. (1997) MESA (Multi-Ethnic Study of Atherosclerosis), Zhang et al. (2018); Chen et al. (2015) and the DREAMT Wang et al. (2024) dataset through model pretraining and fine-tuning, we aim to enhance WatchSleepNet's generalization and robustness.

This study makes several significant contributions to the field of wearable sleep staging:

- We are the first to publish the sleep staging codebase applied on a publicly available dataset for sleep staging using a wearable smartwatch (the Empatica E4), promoting transparency and reproducibility in a domain often lacking published or open source code for wearable-based sleep staging.

- We introduce WatchSleepNet, which effectively captures both spatial and temporal features from physiological signals by focusing on IBI signals. This allows the use of both ECG and PPG for pretraining, expanding the range of possible pre-

Table 1: Description of datasets used. DREAMT: Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology. SHHS: Sleep Heart Health Study. MESA: Multi-Ethnic Study of Atherosclerosis.

Dataset	**# subjects**	**Data samples used**	**Devices**	**No apnea**	**Mild apnea**	**Moderate apnea**	**Severe apnea**
DREAMT	100	100	PSG E4	26	25	24	25
SHHS	5791	7845	PSG	1467	2829	2062	1477
MESA	2055	2055	PSG PPG(Finger) ACC	415	569	474	478

training datasets and addressing the scarcity of PPG-based sleep staging resources.

- Our approach demonstrates how widely available signals can serve as a proxy for less accessible ones (e.g., using ECG to inform PPG-based models), thereby advancing transfer learning methodologies in biomedical informatics.

2. Methods

2.1. Datasets

This study utilizes three key datasets: SHHS, MESA, and DREAMT. SHHS and MESA are PSG datasets, and DREAMT contains both PSG and high-resolution wearable device signals. In Table 1) and the following section, we describe each dataset in detail, as well as how each dataset was used in this study. For additional details about the datasets, such as demographics and label support, please refer to Appendix A.

2.1.1. SHHS

The Sleep Heart Health Study (SHHS) (Zhang et al. (2018); Quan et al. (1997)) is a large-scale, multicenter cohort study initiated by the National Heart, Lung, and Blood Institute (NHLBI) to explore the relationship between sleep-disordered breathing (SDB), including obstructive sleep apnea (OSA), and cardiovascular outcomes. The study enrolled 6,441 participants aged 40 years and older from nine existing epidemiological cohorts between November 1995 and January 1998. The PSG recordings were conducted using a standardized configuration that included EEG, EOG, EMG, ECG, thoracic and abdominal excursions, airflow detection, pulse oximetry, heart rate, body position, and ambient light monitoring (Table 1). For our study, we combined the 5,791 ECG records from the SHHS1 dataset (N=5,791 unique individuals) and 2,054 ECG records from the SHHS2 dataset (N=2,054 unique individuals). In both datasets, the ECG was sampled at 125 Hz, and the length of each recording was, on average, approximately 10 hours. The distribution of sleep apnea severity among the combined SHHS1 and SHHS2 study participants is: 17.3% with no apnea, 33.8% with mild apnea, 24.7% with moderate apnea, and 17.0% with severe apnea. In Table 1, we report the exact numbers of unique participants in each sleep apnea severity category.

2.1.2. MESA

The Multi-Ethnic Study of Atherosclerosis (MESA) (Zhang et al. (2018); Chen et al. (2015)) is a comprehensive, multi-center cohort study sponsored by the NHLBI to investigate factors associated with the development and progression of subclinical cardiovascular disease (CVD) in a diverse population. The study initially enrolled 6,814 participants aged 45-84 years, representing Black, White, Hispanic, and Chinese-American ethnic groups from six U.S. centers between 2000 and 2002. To further explore the relationship between sleep disorders and cardiovascular outcomes, the MESA Sleep ancillary study was conducted between 2010 and 2012, enrolling 2,237 participants who underwent full overnight polysomnography, along with 7-day wrist-worn actigraphy and sleep questionnaires. We utilized the finger-tip PPG

signals sampled at 256 Hz from a subset of 2,055 participants in the MESA Sleep study, excluding participants whose raw data files could not be directly processed with the python package NeuroKit (Makowski et al. (2021)). The distribution of sleep apnea severity among these participants is: 20.2% with no apnea, 27.7% with mild apnea, 23.1% with moderate apnea, and 23.3% with severe apnea (Table 1).

2.1.3. DREAMT

The DREAMT dataset (Wang et al. (2024)) is a public dataset for wearable sleep staging research. The dataset includes raw physiological signals collected through both the Empatica E4, a research grade wearable device, and an overnight PSG, from 100 participants. The DREAMT dataset serves as a testbed for evaluating how well different sleep staging models generalize across different datasets and population demographics. The PPG signal from Empatica E4 was sampled at 64 Hz while the ECG signal from the PSG was sampled at 200 Hz. The distribution of sleep apnea severity among these participants is: 26% with no apnea, 25% with mild apnea, 24% with moderate apnea, and 25% with severe apnea.

2.1.4. Data Preprocessing

To ensure consistency and facilitate direct comparisons with DREAMT, we truncated each SHHS and MESA recording to 1,100 epochs, starting from the beginning of the signal and discarding any data thereafter. This approach preserves the most representative initial segment across participants, standardizes recording lengths for transfer learning tasks, and aligns with the maximum length used in the DREAMT dataset.

The preprocessing approach varied depending on the dataset. For SHHS, the raw ECG signals were read and IBI values extracted using NeuroKit (Makowski et al. (2021)). The MESA dataset's finger tip PPG signals were processed similarly. The DREAMT dataset, however, required extensive preprocessing due to its unique structure and noise characteristics. We calculated IBI values for each 30-second epoch of the blood volume pulse (BVP) signals from Empatica E4. This method minimizes the impact of irregularly tall peaks and troughs, which may be a result of motion artifacts or loose contact between the skin and the LED, from overshadowing the reliable peaks across the entire signal. Any IBI value larger than 2 seconds, which indicates a heart rate of less than 30 beats per minute, was considered physiologically abnormal and replaced with zero. Such anomalously large IBI values could occur between two extremely large peaks or when the device was not in proper contact with the skin. By implementing these preprocessing methods, we ensured that the extracted IBI signals were clean and reliable for further analysis.

2.2. Benchmark Models and Adaptations

InsightSleepNet was the only model found with open source code for sleep staging on publicly available databases (Nam et al. (2024)). The model integrates advanced components like local attention, InceptionTime (Fawaz et al. (2020)), and TCN to capture temporal features. While InsightSleepNet was designed for PPG, for our comparison, we adapted the model to use IBI signals. We used the same cross-entropy loss function and Adam optimizer (Kingma and Ba (2017)), but performed hyperparameter tuning on the pretraining dataset to find the best performing set of parameters for the task of using IBI to classify Wake, NREM sleep and REM sleep.

SleepPPGNet's architecture—combining residual convolutional layers with a TCN—efficiently captures both local morphological details and long-term temporal dependencies, critical for accurate sleep staging in ambulatory settings (Kotzen et al. (2023)). While there was no code or available dataset, SleepPPGNet was reproduced because of its applicability to wearable sleep staging algorithm development and its similarity to our designed pipeline. For our implementation, we also adapted SleepPPGNet to the wearable domain by fine-tuning its hyperparameters, similar to our implementation of InsightSleepNet.

SleepConvNet (Sridhar et al. (2020)) is one of the few models explicitly designed to utilize IBI signals allowing a direct comparison to WatchSleepNet. The model employs convolutional and dilated convolutional layers to capture both local and long-range features from the IBI input. We applied cross-entropy loss function with the Adam optimizer. We performed hyperparameter tuning of the SleepConvNet on the pretraining dataset with 5-fold cross-validation to find the best performing set of hyperparameters. Aside from hyperparameters, we added an additional convolutional layer to avoid shape mismatches during matrix operations. This issue is most likely due to the difference in computational tools and infras-

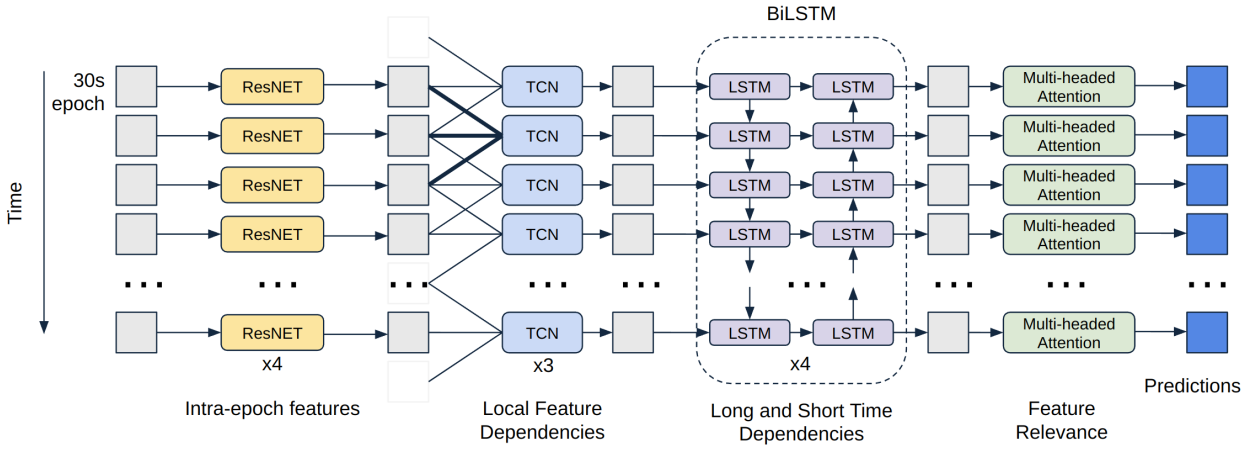


Figure 1: Abstracted schematic of the neural network architecture used for sleep staging. The figure provides a high-level overview of the major components, including ResBlocks, TCN layers, LSTM with multihead attention, and Linear layers, without showing detailed input and output shapes.

tructure. The original model was developed using proprietary tools and infrastructure while our implementation was carried out in PyTorch. These modifications allowed us to reproduce the model's core functionality as faithfully as possible while adapting it to our experimental framework.

2.3. Model Architecture

WatchSleepNet is designed to capture both spatial and temporal features from wearable-derived physiological signals for sleep stage classification. (Figure 1) First, the residual convolutional blocks (inspired by ResNet) extract multi-level spatial features while preserving crucial information through skip connections, which helps mitigate vanishing gradients. These CNN-based layers excel at detecting local patterns but need further temporal modeling to fully leverage the sequential nature of sleep data.

To address longer-range dependencies, we incorporate a Temporal Convolutional Network (TCN). The TCN employs dilated and causal convolutions, enabling the model to process extended temporal contexts while preserving the order of events. This is especially beneficial in sleep staging, where distant epochs can still influence current sleep states. Next, the bidirectional LSTM provides forward and backward sequence modeling, capturing contextual information both before and after each time step. This complements the TCN by refining any uncovered dependencies that might extend beyond typical convolutional receptive fields.

Finally, we add a multi-head self-attention mechanism, allowing the model to selectively weight crucial moments in the time series. The multi-head attention mechanism helps the model focus on different parts of the sequence simultaneously, which is particularly well-suited for applications like sleep staging where information is dispersed over time. The output of the attention layer then passes through a fully connected layer with softmax to produce per-epoch sleep-stage predictions. This end-to-end architecture—combining CNN, TCN, LSTM, and attention—provides a robust, multi-scale approach that captures short-term fluctuations, longer temporal trends, and the critical events most relevant to accurate sleep staging. For more details, please refer to Appendix C.

Hyperparameters such as TCN layers, LSTM layers, and the number of attention heads heavily influences model performance. Increasing the number of layers led to a larger number of parameters, which can cause overfitting and increased computational demands, while decreasing the layers can result in underfitting and insufficient modeling of temporal dependencies. Through a grid search of hyperparameter values, we identified the optimal settings that yielded the best trade-off between model complexity, training efficiency, and predictive accuracy (Appendix C).

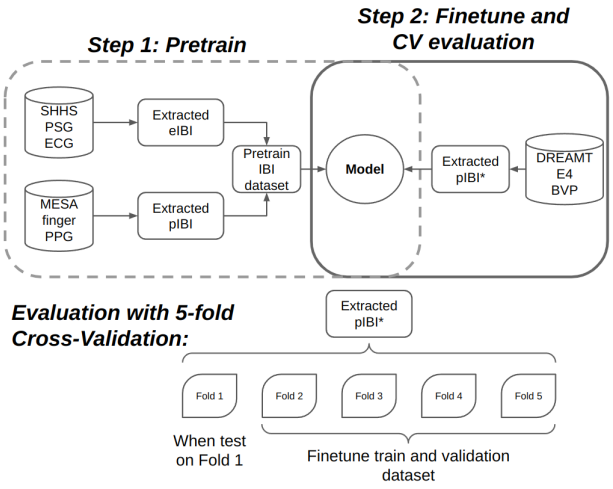


Figure 2: Pretraining a model on a combined dataset of IBI extracted from SHHS (using ECG) and MESA (using finger PPG) signals, then fine-tuning on the DREAMT dataset. Model evaluation is performed using 5-fold cross-validation on the DREAMT dataset.

2.4. Pre-training and model finetuning

In our study, we employed transfer learning by fine-tuning of a pre-trained model to enhance performance on the sleep stage classification task using data from the DREAMT dataset (Figure 2). All models were trained for 200 epochs with early stopping implemented with a patience of 20 epochs. While our main experiment focuses on finetuning on all layers of the models for a fair comparison between the benchmarking models and our WatchSleepNet, we later conducted an ablation study freezing different layers from WatchSleepNet during finetuning (See Appendix D and Supplementary Table 6).

2.5. Performance Metrics

To comprehensively evaluate the model's effectiveness, we assessed its performance using multiple metrics: Accuracy, macro F1-score, Cohen's Kappa, AUROC, and the F1-score specifically for the REM sleep stage (i.e., REM F1). These metrics were chosen to reflect the model's capability to handle imbalanced data and its proficiency in classifying different sleep stages, particularly the challenging REM stage.

3. Results

3.1. WatchSleepNet performance compared to benchmark algorithms

Under the same pretraining and finetuning approach, WatchSleepNet showed significant improvement in performance over SleepConvNet, SleepPPGNet and InsightSleepNet, for 3-stage (Wake vs NREM vs REM) sequence-to-sequence classification (Table 2), in all performance metrics, including overall accuracy, macro F1, REM F1, AUROC, and Cohen's Kappa, with average and standard deviation REM F1-score of 0.631 $\pm$ 0.046 and average and standard deviation Cohen's Kappa of 0.553 $\pm$ 0.027. InsightSleepNet had the lowest REM F1-score of 0.0029 $\pm$ 0.047 and Cohen's Kappa of 0.131 $\pm$ 0.0402. SleepConvNet performed moderately better, achieving an REM F1-score of 0.433 $\pm$ 0.053 and Cohen's Kappa of 0.404 $\pm$ 0.047. SleepPPGNet achieved an REM F1-score of 0.409 $\pm$ 0.055 and Cohen's Kappa of 0.426 $\pm$ 0.031. The summary of metrics comparing the performance of these three models can be found in Table 2. Additional model architecture and model training details, such as tunable parameters and training time for WatchSleepNet and benchmarking algorithms can be found in Appendix C.

3.2. Ablation of WatchSleepNet layers

We also performed an ablation study by toggling the TCN layers and attention layers. All ablation experiments were conducted using the best performance hyperparameter settings determined from the full architecture to ensure that any performance changes are due solely to the inclusion or exclusion of the TCN and attention modules. We specifically chose to toggle only TCN and attention because the ResNet-like convolutional blocks and the LSTM component are fundamental for robust feature extraction and sequence-to-sequence classification, respectively, while the TCN layers and attention layers are non-essential. Figure 3 visualizes the ablation results comparing four configurations of our model: (1) With TCN and with attention, (2) With TCN but without attention, (3) Without TCN but with attention, and (4) Without TCN and without attention. Each bar shows mean performance on two evaluation methods: Method (A) corresponds to 5-fold cross-validation (CV) on the large SHHS + MESA dataset (i.e., pretraining results shown in dark blue), while Method (B) corresponds to 5-fold

Table 2: Performance comparison of different models when finetuned and cross-validated on the DREAMT dataset. Performance metrics are reported as means and standard deviations where applicable.

Model	Acc	F1	REM F1	AUROC	Kappa
WatchSleepNet	**0.785 ± 0.016**	**0.780 ± 0.017**	**0.631 ± 0.046**	**0.888 ± 0.015**	**0.553 ± 0.027**
SleepConvNet	0.730 ± 0.016	0.713 ± 0.022	0.433 ± 0.053	0.825 ± 0.028	0.404 ± 0.047
SleepPPGNet	0.739 ± 0.029	0.722 ± 0.021	0.409 ± 0.055	0.823 ± 0.028	0.426 ± 0.031
InsightSleepNet	0.671 ± 0.027	0.590 ± 0.031	0.0029 ± 0.0047	0.606 ± 0.021	0.131 ± 0.042

CV on the smaller DREAMT dataset after pretraining on the entire SHHS + MESA dataset (shown in light blue). The gray, dashed trend lines in each subplot highlight that performance tends to decrease (slope downwards) when moving from left (with TCN and/or attention) to right (without TCN and/or attention). This degradation is more pronounced for DREAMT (Method B), as shown by the lower bars and more negative slope in the "light blue" series for both REM F1 and Cohen's Kappa. This underscores how the TCN layers and the attention layers, while non-essential to our sequence-to-sequence sleep-staging task, help maintain higher consistency between training on the larger dataset and fine-tuning on the smaller one. For more details on the performance metrics, please see Supplementary Table 5 in Appendix D.

3.3. Additional Ablation Experiments

Under the best hyperparameters identified via tuning, we conducted additional ablation studies to assess the impact of freezing various network layers during finetuning. Specifically, we experimented with no finetuning, and gradually allowing more layers to be finetuned with DREAMT IBI data. Our findings indicate that full finetuning of all parameters yields the best performance—evidenced by higher REM F1 and Cohen's Kappa scores—compared to partial finetuning. For detailed explantion and results, plase see Supplementary Table 6 in Appendix D). Additional experiments with different pretraining dataset combinations (i.e., with only MESA, only MESA, SHHS+MESA, as well as no pretraining at all) further confirmed that leveraging a combined SHHS+MESA dataset markedly improves the model's performance over single-source or no pretraining approaches, likely due to the abundance of pretraining data allowing the model to better capture generalizable sleep patterns. (see Supplementary Table 7 in in Appendix D).

4. Discussion

4.1. Model Performances

We conducted preliminary experiments on using PPG signals for sleep staging in exactly the same pipeline as Figure 2. The results (see Supplementary Table 3) reveal that both InsightSleepNet and SleepPPGNet demonstrated strong performance on the MESA PPG dataset, with mean REM F1 scores of 0.668 and 0.710, respectively, and high accuracy and Kappa values. However, when these models were finetuned on the DREAMT dataset—comprising wearable wrist-watch BVP signals—their performance dropped markedly (REM F1 scores fell to 0.0026 and 0.200, respectively). This discrepancy underscores the significant signal characteristic differences between fingertip PPG data from MESA and wearable BVP signals from DREAMT (see Appendix B for further details). Furthermore, when applied to IBI signals, both models suffered from overparameterization, which likely led to overfitting and, consequently, suboptimal performance.

In contrast, SleepConvNet, with its simpler convolutional architecture, was more effective at capturing local features from IBI signals. Its moderate performance—evidenced by a REM F1 score of 0.433 and a Cohen's Kappa of 0.404—suggests that while its streamlined design promotes robustness and some degree of generalization, it lacks the recurrent and

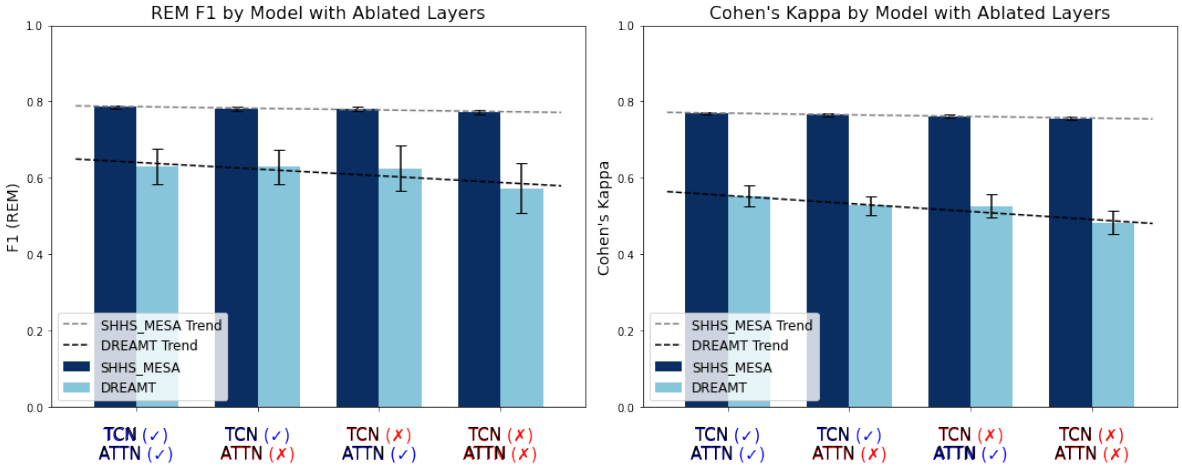


Figure 3: WatchSleepNet performance metrics of REM F1 (left) and Cohen's Kappa (right) with and without TCN layers and/or attention layers. Evaluation Strategy: "SHHS+MESA" means 5-fold CV within the SHHS + MESA pretraining dataset, and "DREAMT" means 5-fold CV with finetuning on DREAMT. Gray lines are fitted linear regression of the model performance scores as the specific layers of the models are removed.

attention-based mechanisms needed to fully capture long-range temporal dependencies. This trade-off between efficiency and performance is inherent to its design, making it computationally efficient but less capable of handling the variability and complexity of sleep-stage dynamics compared to more sophisticated architectures.

Among the models we have tested, WatchSleepNet emerged as the best-performing model, striking a balance between model complexity, training efficiency, and generalization. Its architecture integrates ResNet-inspired feature extraction, a Temporal Convolutional Network (TCN), and an LSTM with multi-head attention, enabling it to capture both local and long-range dependencies in IBI signals. This comprehensive design not only results in great performance on IBI signals—with a mean REM F1 score of 0.631 and a Cohen's Kappa of 0.553—but also ensures effective transferability when finetuned on smaller, more variable datasets like DREAMT. The ablation studies further reinforce that the inclusion of both TCN and attention layers is crucial for maintaining performance consistency across datasets, highlighting WatchSleepNet's robustness and adaptability in practical sleep staging applications.

4.2. Choice of Signal

Our decision to use IBI signals for sleep staging in this study is strongly driven by the significant challenges in harmonizing wearable-based PPG and fingertip PPG waves, as well as the lower availability of publicly available fingertip PPG datasets with clinical gold-standard sleep labels, compared to exisint datasets containing ECG signals. There are numerous benefits to using IBI signals for sleep staging. First, there is significantly more data available for pre-training with IBI, as it can be derived from both ECG and PPG datasets, expanding the range of available pretraining datasets, compared to limiting our pretraining datasets to PPG only, which could boost the generalizability of pretrained models. In addition, the use of IBI signals minimizes the impact of the differences in PPG and ECG morphology on model pretraining and fine-tuning, making them a more reliable biometric when working with heterogeneous and large external datasets. This is because raw PPG signals are more prone to motion artifacts and device-specific variations such as the heterogeneity of PPG morphology, particularly when recorded from different body locations such as the wrist or finger (Castaneda et al. (2018)). Choosing well-processed IBI signals allows

us to leverage the breadth of existing ECG and PPG datasets while maintaining robustness and scalability.

4.3. Experimentation Transparency

In the field of wearable sleep staging, a significant challenge lies in researchers and developers not making their codebase and data available, which hinders iterative and cumulative advancements in algorithm performance and generalizability. This lack of openness has limited the reproducibility and validation of many published methods, making it difficult for researchers to build upon existing work. Providing detailed architecture and (hyper)parameter descriptions as written methods without the accompanying codebase is insufficient– there is no guarantee of accurately reproducing models reported in this way. Further, the absence of standard datasets for benchmarking further complicates efforts to compare models and evaluate performance across studies.

In this context, our work is significant as it addresses these challenges directly. We contribute a complete and publicly available codebase for sleep staging using a dataset containing high-resolution smartwatch data (Wang et al. (2024)), ensuring that our work can be reproduced and built upon by other researchers and offering a tool for future benchmarking. By providing these resources, we aim to foster transparency and collaboration in the field, setting a new standard for reproducibility in wearable sleep staging research.

4.4. Limitations

In this study, we aimed to demonstrate the applicability and potential of deep learning algorithms developed for 3-class sleep staging (Wake, NREM, REM) using the data from Empatica E4. Our study does not differentiate between light and deep sleep stages or the finer granularity of N1, N2, and N3 stages present in 4- or 5-class sleep staging models that may provide more detailed insights into sleep patterns. Future work could expand WatchSleepNet to more detailed 4-class and 5-class sleep staging, enhancing the model's capability to capture finer sleep stage transitions.

The study also does not incorporate other physiological signals collected from the Empatica E4 device, such as accelerometry, electrodermal activity, and skin temperature. A key reason for this omission is the lack of external datasets with gold standard sleep labels annotated by sleep technicians based on PSG for pretraining on these signal types. Nonetheless, accelerometry, being the second most common signal type used for sleep staging algorithms, is a candidate for future investigation to complement the IBI signals used in this study.

We recognize that our study's benchmarking is incomplete, due to the significant challenges while reproducing pipelines without published codebase or public data. For instance while we attempted to replicate Li et al. (2021)'s CNN+SVM approach for a three-stage sleep staging task (WAKE vs. NREM vs. REM), our extensive hyperparameter tuning yielded unsatisfactory results—with the CNN-only baseline achieving an accuracy of approximately 65.9%, a Cohen's kappa of 0.442, a macro F1-score of 0.563, and a REM F1-score of only 0.257 on the pretraining dataset, far below what we expect to perform well on the pretraining dataset before transferred on a smaller and more noisy dataset like DREAMT. This discrepancy underscores the challenges inherent in reproducing pipelines that depend on custom-engineered features and undisclosed preprocessing steps. In contrast, we successfully reproduced SleepPPGNet and benchmarked our model against other approaches such as SleepConvNet and InsightSleepNet, which offer clearer descriptions and/or reproducible implementations. Additionally, our study omits spectrogram-based models like Olsen's deep learning pipeline due to their substantially different methodological approach, suggesting that future work should directly compare these frameworks.

Lastly, while our participant cohort included individuals with varying degrees of sleep apnea and some other sleep disorders, we did not conduct a comprehensive analysis of how these conditions might influence model performance. We recognize the importance of understanding sleep disorder–specific effects, and we have provided some preliminary results in Appendix E that illustrate performance differences across no apnea, mild, moderate, and severe apnea groups. However, a more in-depth investigation into these and other sleep disorders remains an essential avenue for future work, as it could further refine our algorithm's applicability and reliability in real-world and clinical contexts. Another potential future direction is to consider building models that account for each participant's sleep disorders and health conditions, for example by using mixed-effects modeling or by explicitly incorporating these variables as input features to the model.

5. Conclusion

In this work, we introduced a novel sleep staging algorithm that leverages interbeat interval signals for robust model pretraining and fine-tuning, significantly improving wrist-based wearables' performance in 3-stage sleep classification. Our results highlight that combining large ECG datasets with smaller PPG data can effectively capture both local and long-range temporal dependencies, particularly in challenging scenarios like REM detection. By publicly releasing our code and benchmarking data, we aim to foster transparency and collaboration in wearable sleep research. This approach not only addresses long-standing reproducibility gaps but also sets a foundation for more advanced, clinically applicable sleep staging solutions.

Acknowledgments

Acknowledgments redacted

References

Shirel Attia, Revital Shani Hershkovich, Alissa Tabakhov, Angeleene Ang, Sharon Haimov, Riva Tauman, and Joachim A. Behar. SleepPPG-Net2: Deep learning generalization for sleep staging from photoplethysmography. (arXiv:2404.06869), April 2024. doi: 10.48550/arXiv.2404.06869. URL http://arxiv.org/abs/2404.06869. Publisher: arXiv.

Mathias Baumert, Martin R Cowie, Susan Redline, Reena Mehra, Michael Arzt, Jean-Louis Pépin, and Dominik Linz. Sleep characterization with smart wearable devices: a call for standardization and consensus recommendations. *Sleep*, 45(12):zsac183, December 2022. ISSN 0161-8105. doi: 10.1093/sleep/zsac183.

Vera Birrer, Mohamed Elgendi, Olivier Lambercy, and Carlo Menon. Evaluating reliability in wearable devices for sleep staging. *NPJ Digital Medicine*, 7:74, March 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01016-9.

Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, 4(4):195–202, 2018. ISSN 2573-2838. doi: 10.15406/ijbsbe.2018.04.00125.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline. Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, June 2015. ISSN 1550-9109. doi: 10.5665/sleep.4732.

Ambrose A. Chiang and Seema Khosla. Consumer Wearable Sleep Trackers: Are They Ready for Clinical Use? *Sleep Medicine Clinics*, 18(3): 311–330, September 2023. ISSN 1556-4088. doi: 10.1016/j.jsmc.2023.05.005.

Evan D Chinoy, Joseph A Cuellar, Kirbie E Huwa, Jason T Jameson, Catherine H Watson, Sara C Bessman, Dale A Hirsch, Adam D Cooper, Sean P A Drummond, and Rachel R Markwald. Performance of Seven Consumer Sleep-Tracking Devices Compared with Polysomnography. *Sleep*, (zsaa291), December 2020. ISSN 0161-8105. doi: 10.1093/sleep/zsaa291. URL https://doi.org/10.1093/sleep/zsaa291.

Evan D Chinoy, Joseph A Cuellar, Jason T Jameson, and Rachel R Markwald. Performance of Four Commercial Wearable Sleep-Tracking Devices Tested Under Unrestricted Conditions at Home in Healthy Young Adults. *Nature and Science of Sleep*, 14:493–516, March 2022. ISSN 1179-1608. doi: 10.2147/NSS.S348795.

Massimiliano de Zambotti, Cathy Goldstein, Jesse Cook, Luca Menghini, Marco Altini, Philip Cheng, and Rebecca Robillard. State of the science and recommendations for using wearable technology in sleep and circadian research. *Sleep*, 47(4):zsad325, April 2024. ISSN 0161-8105. doi: 10.1093/sleep/zsad325.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for Time Series Classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, November 2020. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-020-00710-y.

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. Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population. *Sleep*, 43(9):zsaa048, September 2020. ISSN 0161-8105. doi: 10.1093/sleep/zsaa048.

Pedro Fonseca, Marco Ross, Andreas Cerny, Peter Anderer, Fokke van Meulen, Hennie Janssen, Angelique Pijpers, Sylvie Dujardin, Pauline van Hirtum, Merel van Gilst, and Sebastiaan Overeem. A computationally efficient algorithm for wearable sleep staging in clinical populations. *Scientific Reports*, 13(1):9182, June 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-36444-2. Publisher: Nature Publishing Group.

Conor Heneghan, Ryan Gillard, Logan Niehaus, Logan Schneider, and Marius Guerard. Sleep Staging Classification from Wearable Signals Using Deep Learning. *Sleep*, 47(Supplement_1):A130, May 2024. ISSN 0161-8105. doi: 10.1093/sleep/zsae067.0302.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. (arXiv:1412.6980), January 2017. doi: 10.48550/arXiv.1412.6980. URL http://arxiv.org/abs/1412.6980. Publisher: arXiv.

Henri Korkalainen, Juhani Aakko, Brett Duce, Samu Kainulainen, Akseli Leino, Sami Nikkonen, Isaac O Afara, Sami Myllymaa, Juha Töyräs, and Timo Leppänen. Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea. *Sleep*, 43(11):zsaa098, November 2020. ISSN 0161-8105. doi: 10.1093/sleep/zsaa098.

Kevin Kotzen, Peter H. Charlton, Sharon Salabi, Lea Amar, Amir Landesberg, and Joachim A. Behar. SleepPPG-Net: A Deep Learning Algorithm for Robust Sleep Staging From Continuous Photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 27(2):924–932, February 2023. ISSN 2168-2208. doi: 10.1109/JBHI.2022.3225363.

Pei-Lin Lee. Wearable sleep tracker in clinical settings: challenges and promise. *Journal of Clinical Sleep Medicine*, 20(7):1221–1221, 2024. doi: 10.5664/jcsm.11148. Publisher: American Academy of Sleep Medicine.

Qiao Li, Qichen Li, Ayse S. Cakmak, Giulia Da Poian, Donald L. Bliwise, Viola Vaccarino, Amit J. Shah, and Gari D. Clifford. Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables. *Physiological Measurement*, 42(4):044004, May 2021. ISSN 0967-3334. doi: 10.1088/1361-6579/abf1b0. Publisher: IOP Publishing.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, feb 2021. doi: 10.3758/s13428-020-01516-y. URL https://doi.org/10.3758%2Fs13428-020-01516-y.

Dean J. Miller, Charli Sargent, and Gregory D. Roach. A Validation of Six Wearable Devices for Estimating Sleep, Heart Rate and Heart Rate Variability in Healthy Adults. *Sensors (Basel, Switzerland)*, 22(16):6317, August 2022. ISSN 1424-8220. doi: 10.3390/s22166317.

Borum Nam, Beomjun Bark, Jeyeon Lee, and In Young Kim. InsightSleepNet: the interpretable and uncertainty-aware deep learning network for sleep staging using continuous Photoplethysmography. *BMC Medical Informatics and Decision Making*, 24(1):50, February 2024. ISSN 1472-6947. doi: 10.1186/s12911-024-02437-y.

Mads Olsen, Jamie M. Zeitzer, Risa N. Richardson, Polina Davidenko, Poul J. Jennum, Helge B. D. Sørensen, and Emmanuel Mignot. A flexible deep learning architecture for temporal sleep stage classification using accelerometry and photoplethysmography. *IEEE Transactions on Biomedical Engineering*, 70(1):228–237, 2023. doi: 10.1109/TBME.2022.3187945.

Tajammul Pangarkar. Smart Wearables Statistics and Facts (2025), January 2025. URL https://scoop.market.us/smart-wearables-statistics/.

Paul E. Peppard, Terry Young, Jodi H. Barnet, Mari Palta, Erika W. Hagen, and Khin Mae Hla. Increased Prevalence of Sleep-Disordered Breathing in Adults. *American Journal of Epidemiology*, 177 (9):1006–1014, May 2013. ISSN 0002-9262. doi: 10.1093/aje/kws342.

S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl. The Sleep Heart Health Study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, December 1997. ISSN 0161-8105.

Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M. Aarts. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *npj Digital Medicine*, 4(1):1–11, September 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00510-8. Publisher: Nature Publishing Group.

Fernanda B. Silva, Luisa F. S. Uribe, Felipe X. Cepeda, Vitor F. S. Alquati, João P. S. Guimarães, Yuri G. A. Silva, Orlem L. dos Santos, Alberto A. de Oliveira, Gabriel H. M. de Aguiar, Monica L. Andersen, Sergio Tufik, Wonkyu Lee, Lin Tzy Li, and Otávio A. Penatti. Sleep staging algorithm based on smartwatch sensors for healthy and sleep apnea populations. *Sleep Medicine*, 119:535–548, July 2024. ISSN 1389-9457. doi: 10.1016/j.sleep.2024.05.033.

Niranjan Sridhar, Ali Shoeb, Philip Stephens, Alaa Kharbouch, David Ben Shimol, Joshua Burkart, Atiyeh Ghoreyshi, and Lance Myers. Deep learning for automated sleep staging using instantaneous heart rate. *npj Digital Medicine*, 3(1):1–10, August 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0291-x. Publisher: Nature Publishing Group.

Ee Rah Sung, Zakaa Hassan, and J. Shirine Allam. Emerging Technologies and Wearables for Monitoring and Managing Sleep Disorders in Patients with Cardiovascular Disease. *Current Sleep Medicine Reports*, 10(2):158–168, June 2024. ISSN 2198-6401. doi: 10.1007/s40675-024-00280-1.

Thomas Svensson, Kaushalya Madhawa, Hoang Nt, Ung-il Chung, and Akiko Kishi Svensson. Validity and reliability of the Oura Ring Generation 3 (Gen3) with Oura sleep staging algorithm 2.0 (OSSA 2.0) when compared to multi-night ambulatory polysomnography: A validation study of 96 participants and 421,045 epochs. *Sleep Medicine*, 115:251–263, March 2024. ISSN 13899457. doi: 10.1016/j.sleep.2024.01.020.

Pavlos Topalidis, Dominik P. J. Heib, Sebastian Baron, Esther-Sevil Eigl, Alexandra Hinterberger, and Manuel Schabus. The Virtual Sleep Lab—A Novel Method for Accurate Four-Class Sleep Staging Using Heart-Rate Variability from Low-Cost Wearables. *Sensors*, 23(55):2390, January 2023. ISSN 1424-8220. doi: 10.3390/s23052390. Publisher: Multidisciplinary Digital Publishing Institute.

Will Ke Wang, Jiamu Yang, Leeor Hershkovich, Hayoung Jeong, Bill Chen, Karnika Singh, Ali R Roghanizad, Mobashir Hasan Shandhi, Andrew R Spector, and Jessilyn Dunn. Addressing wearable sleep tracking inequity: a new dataset and novel methods for a population with sleep disorders. 2024.

Bernice M Wulterkens, Pedro Fonseca, Lieke W A Hermans, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, Johannes P van Dijk, Nele Vandenbussche, Sigrid Pillen, Merel M van Gilst, and Sebastiaan Overeem. It is All in the Wrist: Wearable Sleep Staging in a Clinical Population versus Reference Polysomnography. *Nature and Science of Sleep*, 13:885–897, June 2021. ISSN null. doi: 10.2147/NSS.S306808. Publisher: Dove Medical Press.

Terry Young, Linda Evans, Laurel Finn, and Mari Palta. Estimation of the Clinically Diagnosed Proportion of Sleep Apnea Syndrome in Middle-aged Men and Women. *Sleep*, 20(9):705–706, September 1997. ISSN 0161-8105. doi: 10.1093/sleep/20.9.705.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association: JAMIA*, 25(10):1351–1358, October 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy064.

Appendix A. Additional Information on Datasets Used

Supplementary Table 1 summarizes key demographic details for the DREAMT, MESA, and SHHS datasets. It shows the minimum and maximum ages, along with the mean age and its standard deviation, and breaks down subject counts into three age groups: Young (18–39), Mid (40–64), and Old ($\geq$ 65). Although DREAMT features a broader age range with a lower mean age (56.24 years), MESA and SHHS include predominantly older participants (mean ages of 69.60 and 65.30 years, respectively). The gender distribution is fairly balanced across all datasets. Overall, despite differences in absolute numbers, the age group distributions and gender proportions are comparable, supporting consistent cross-dataset analyses.

Supplementary Table 1: Demographic and medical information from DREAMT

Dataset	**Min Age**	**Max Age**	**Mean Age**	**Age Std**	**Count Young (18–39)**	**Count Mid (40-64)**	**Count Old ($\geq$ 65)**	**Count Male**	**Count Female**
DREAMT	21	87	56.24	16.6	20	44	33	45	55
MESA	54	95	69.6	9.2	0	810	1427	1039	1198
SHHS	39	90	65.3	11.2	6	4693	5185	4626	5258

Supplementary Table 2 shows the sample support for sleep stage labels in the DREAMT, SHHS, and MESA datasets. Despite variations in total epoch counts, the relative distributions of wake, NREM, and REM epochs are quite similar. Both DREAMT and MESA have NREM epochs contributing roughly 47–50%, with REM epochs consistently representing the smallest fraction (10–14%). Although SHHS exhibits a slightly higher share of REM epochs (13.57%), the overall label support remains comparable across all three datasets, ensuring a balanced representation of sleep stages for cross-dataset analyses.

Supplementary Table 2: Datasets labels sample support

Dataset	**# Wake Epochs**	**% Wake Epochs**	**# NREM Epochs**	**% NREM Epochs**	**# REM Epochs**	**% REM Epochs**
DREAMT	18,709	23.84%	51,399	65.50%	8,365	10.66%
SHHS	2,898,168	32.03%	4,922,985	54.40%	1,227,594	13.57%
MESA	1,106,467	42.92%	1,203,872	46.69%	267,797	10.39%

Appendix B. Preliminary Experiments with PPG

The table summarizes experiments using two PPG-based models—InsightSleepNet Nam et al. (2024) and SleepPPGNet (with raw PPG signal) Kotzen et al. (2023)—with a consistent preprocessing pipeline applied to both MESA (fingertip PPG) and DREAMT (wrist BVP) data based on Nam et al. (2024). When trained and tested on MESA PPG (80%/20% split), both models achieve high accuracy, F1), and AUROC, with SleepPPGNet showing slightly higher REM F1 (0.710 vs. 0.668) and Cohen's Kappa (0.719 vs. 0.703) than InsightSleepNet. However, when models are trained on 100% MESA PPG and then finetuned on DREAMT via 5-fold cross-validation, performance drops significantly—particularly in REM F1 (around 0.0026 for InsightSleepNet and 0.200 SleepPPGNet) and Cohen's Kappa—highlighting challenges in generalizing from fingertip PPG to wrist-based BVP. This suggests that while both models learn PPG features well in the source dataset, additional preprocessing or domain adaptation techniques may be needed to enhance cross-dataset and cross-modality generalization, and is out of scope of the direction of our paper.

Supplementary Table 3: Datasets labels sample support

Model	Training Dataset	Testing Approach	Acc	F1	REM F1	AUROC	Cohen's Kappa
InsightSleepNet	MESA PPG (80%)	MESA PPG (20%)	0.829	0.828	0.668	0.936	0.703
InsightSleepNet	MESA PPG (100%)	Finetuned on DREAMT through 5-fold CV	0.658 ± 0.018	0.584 ± 0.018	0.0026 ± 0.0028	0.621 ± 0.014	0.150 ± 0.025
SleepPPGNet	MESA PPG (80%)	MESA PPG (20%)	0.840	0.838	0.710	0.943	0.719
SleepPPGNet	MESA PPG (100%)	Finetuned on DREAMT through 5-fold CV	0.719 ± 0.011	0.692 ± 0.013	0.200 ± 0.092	0.789 ± 0.026	0.392 ± 0.034

Appendix C. Additional Modeling Details

C.1. Model Architecture: WatchSleepNet

This appendix includes more detailed description about the architecture of WatchSleepNet, which is designed to capture both spatial and temporal features from wearable-derived physiological signals for sleep stage classification. The architecture comprises three primary components: (i) a series of residual convolutional blocks, (ii) a Temporal Convolutional Network (TCN), and (iii) a bidirectional LSTM augmented with multi-head attention. An overview of the model is depicted in Figure 1.

Input Dimensions and Reshaping Our input data are organized into 30-second epochs, each containing 750 features derived from physiological signals (e.g., IBI measurements). Formally, the input has shape $(\text{batch\_size}, \text{seq\_len}, 750)$. To facilitate convolutional operations, we merge the batch and sequence dimensions into $(\text{batch\_size} \times \text{seq\_len}, 1, 750)$, so that the temporal dimension of each epoch can be processed by 1D convolutions.

Residual Convolutional Blocks We employ four residual blocks, inspired by ResNet, to extract spatial features at multiple scales. Each block consists of convolutional layers, batch normalization, and ReLU activation, with skip connections that add the block's input directly to its output. This structure preserves important low-level features and mitigates vanishing gradients. The blocks progressively reduce the temporal dimension (e.g., using strides) while increasing the number of feature maps. After the final residual block, the feature maps have shape $(\text{batch\_size} \times \text{seq\_len}, 256, 3)$.

Transition to TCN A linear layer is then used to reduce the spatial dimension, collapsing the 3 points per feature map into a single scalar, resulting in $(\text{batch\_size} \times \text{seq\_len}, 256)$. We reshape this to $(\text{batch\_size}, 256, \text{seq\_len})$ so that the TCN can process the feature vectors over the temporal dimension. The TCN uses causal and dilated convolutions to capture both short-term and long-range dependencies, employing residual connections within each TCN layer to facilitate stable gradient flow.

Bidirectional LSTM with Multi-Head attention Following the TCN, we feed the sequences into a bidirectional LSTM network. This LSTM processes data in both forward and backward directions, preserving context that might appear before or after a given time step. We augment this LSTM with multi-head attention, allowing the model to assign learnable weights to various time steps for each attention head. This mechanism is particularly valuable in sleep staging, where key signal patterns may be scattered throughout the sequence. The combined output is reshaped to $(\text{batch\_size}, \text{seq\_len}, 512)$.

Output Layer Finally, a fully connected linear layer applies a softmax activation to produce per-epoch class probabilities. The output shape is $(\text{batch\_size}, \text{seq\_len}, 3)$, corresponding to the three sleep stage classes (e.g., wakefulness, light sleep, and deep sleep).

Architectural Considerations In determining the number of convolutional blocks, TCN layers, and LSTM layers, we balanced model capacity against training complexity. Larger models can learn more intricate patterns but risk overfitting and increased computational overhead, while smaller models may struggle to capture the full range of temporal dependencies inherent in sleep data. The final configuration offers a robust end-to-end solution that addresses both local feature extraction (via the residual blocks) and broader temporal context (via the TCN, LSTM, and attention). This design empowers WatchSleepNet to accurately classify sleep stages from wearable signals while maintaining computational feasibility.

C.2. WatchSleepNet Hyperparameters

To find the best number of layers to use for LSTM and TCN, we performed hyperparameter tuning using 5-fold cross validation on the pretrain (SHHS+MESA) dataset, while using both the TCN layers and attention layers. The best WatchSleepNet model configuration was determined through hyperparameter tuning to optimize performance on the sleep stage classification task. The final model parameters are as follows:

- $NUM\_INPUT\_CHANNELS$: 1
- $NUM\_CHANNELS$: 256
- $KERNEL\_SIZE$: 5
- $HIDDEN\_DIM$: 256
- TCN_LAYERS: 3
- NUM_LAYERS: 4
- NUM_HEADS: 16

Later on, the ablation experiments were performed based on this best configuration.

Supplementary Table 4: Comparison of saved model size, tunable parameters, GPU memory usage, training times, and inference times, for four sleep staging models: WatchSleepNet, InsightSleepNet, SleepConvNet, and SleepPPGNet.

Model name	**Saved model size (MB)**	**# Tunable parameters**	**GPU memory usage**	**Total Training Time (HH:MM:SS)**	**Training time per epoch (MM:SS)**	**Inference Time (ms)**
WatchSleepNet	33.9	8,456,131	15,590 MB at batch_size = 16	07:12:59	03:01	37.71
SleepConvNet	4.8	1,205,155	17,364 MB at batch_size = 16	03:04:33	02:05	2.22
SleepPPGNet	14.0	3,628,547	15,700 MB at batch_size = 4	12:29:44	07:23	13.34
InsightSleepNet	18.0	4,464,580	15,492 MB at batch_size = 4	26:23:24	10:59	54.82

C.3. Model size and training times

The table provides a detailed comparison of four models—WatchSleepNet, InsightSleepNet, SleepConvNet, and SleepPPGNet—across several critical metrics related to their storage, complexity, and training efficiency. The Saved Model Size (MB) indicates the disk footprint of the model once saved, where WatchSleepNet is

the largest at 33.7 MB and SleepConvNet is the smallest at 4.8 MB. The Tunable Parameters column lists the number of parameters the model can learn during training; here, WatchSleepNet has the highest count at 8,456,131, suggesting a more complex architecture with greater capacity to capture detailed features, while SleepConvNet, with only 1,205,155 parameters, is comparatively simpler. The GPU Memory Usage column shows the memory required during training, with the values noted alongside the batch size used (for example, WatchSleepNet uses 15,590 MB at a batch size of 16). This metric is influenced by both the model architecture and the chosen batch size, which explains why models trained with a smaller batch size (InsightSleepNet and SleepPPGNet) have lower memory usage figures despite their differing complexities. The Total Training Time (HH:MM:SS) provides an overall measure of how long each model takes to train from start to finish. Notably, WatchSleepNet trains in just over 7 hours (07:12:59), faster than SleepPPGNet's 12 hours and 29 minutes and InsightSleepNet's lengthy 16 hours and 23 minutes. However, it's worth noting that total training time inherently exhibits a degree of randomness due to factors like early stopping and randomness in gradient descent. Some models may halt training earlier than expected, while others might persist longer, depending on these intrinsic dynamics. Finally, the Training Time per Epoch (HH:MM:SS) indicates the duration of a single training cycle; WatchSleepNet again stands out by requiring only 03:01 per epoch compared to InsightSleepNet's 10:59 per epoch and SleepPPGNet's 07:23 per epoch. This per-epoch time, too, varies depending on the specific hardware, software, and environment in use, making direct comparisons context-dependent. This analysis underscores that, despite having a higher number of tunable parameters, WatchSleepNet is capable of training faster and more efficiently than its counterparts, striking a reasonable balance between model complexity and training performance.

Appendix D. Further Ablation Performances

Supplementary Table 5: WatchSleepNet performance metrics with and without TCN layers and or the attention layers. Evaluation Strategie: (A) 5-fold CV within the SHHS + MESA dataset, (B) 5-fold CV with finetuning on DREAMT

Use TCN	Use attention	Evaluation Method	Accuracy	F1	REM F1	AUROC	Cohen's Kappa
Yes	Yes	(A)	0.867 ± 0.0016	0.867 ± 0.0015	0.786 ± 0.0040	0.962 ± 0.0008	0.769 ± 0.0026
Yes	Yes	(B)	0.785 ± 0.016	0.780 ± 0.017	0.631 ± 0.046	0.888 ± 0.015	0.553 ± 0.027
Yes	No	(A)	0.865 ± 0.0022	0.865 ± 0.0023	0.782 ± 0.0055	0.961 ± 0.0011	0.766 ± 0.0042
Yes	No	(B)	0.779 ± 0.012	0.770 ± 0.016	0.629 ± 0.045	0.883 ± 0.014	0.528 ± 0.026
No	Yes	(A)	0.863 ± 0.0018	0.863 ± 0.0019	0.781 ± 0.0047	0.960 ± 0.0008	0.762 ± 0.0036
No	Yes	(B)	0.775 ± 0.016	0.768 ± 0.019	0.626 ± 0.060	0.878 ± 0.014	0.527 ± 0.031
No	No	(A)	0.860 ± 0.0020	0.859 ± 0.0022	0.772 ± 0.0064	0.958 ± 0.0012	0.756 ± 0.0043
No	No	(B)	0.756 ± 0.025	0.747 ± 0.022	0.573 ± 0.066	0.856 ± 0.019	0.484 ± 0.031

D.1. Toggling TCN and Attention Layers

In addition to the bar plot Figure 3 in main texts to show the ablation results and the importance of TCN layers and the attention layer, we also show here, the ablation results in comprehensive performance metrics for the model under different configuations (with or without TCN or attention layers). Supplementary Table 5 presents an ablation study comparing four configurations of our model: (1) With TCN and with attention, (2) With TCN but without attention, (3) Without TCN but with attention, and (4) Without TCN and without attention. Two evaluation methods are listed. Method (A) corresponds to 5-fold cross-validation (CV) on the large SHHS + MESA dataset (i.e., pretraining results), while Method (B) is the 5-fold CV on the smaller DREAMT dataset after pretraining on the entire SHHS + MESA dataset. We observe that, in all configurations, the performance on DREAMT (Method B) is lower than on SHHS + MESA (Method A). This is illustrated by negative values in the Delta REM F1 and Delta Cohen's Kappa columns, both of which quantify the performance difference between (A) and (B). Notably, the greatest decreases in REM F1 (-0.199) and Cohen's Kappa (-0.272) occur when neither TCN nor attention is used while the performance drop is mitigated when both TCN and attention layers are included (REM F1 (-0.155) and Cohen's Kappa (-0.216)).These patterns indicate that including TCN and attention helps maintain higher consistency between training on the larger dataset and fine-tuning on the smaller dataset.

D.2. Ablation Experiments with different layers frozen

Supplementary Table 6: WatchSleepNet performance metrics with different layers fixed during finetuning.

Finetuning Specifics	**Accuracy**	**F1**	**REM F1**	**AUROC**	**Cohen's Kappa**
No finetuning	0.740	0.741	0.572	0.853	0.489
Finetune on Classifier Only	0.745 ± 0.015	0.749 ± 0.016	0.591 ± 0.053	0.862 ± 0.011	0.505 ± 0.022
Finetune on Attention+Classifier	0.757 ± 0.017	0.752 ± 0.020	0.581 ± 0.058	0.863 ± 0.012	0.493 ± 0.032
Finetune on LSTM+Attention +Classifier	0.773 ± 0.012	0.766 ± 0.016	0.611 ± 0.039	0.877 ± 0.012	0.524 ± 0.029
Finetune on TCN+LSTM +Attention+Classifier	0.771 ± 0.015	0.764 ± 0.019	0.599 ± 0.041	0.872 ± 0.014	0.521 ± 0.034
Finetune on All Layers	0.785 ± 0.016	0.780 ± 0.017	0.631 ± 0.046	0.888 ± 0.015	0.553 ± 0.027

Supplementary Table 6 presents a comprehensive evaluation of WatchSleepNet's performance across various finetuning strategies, with results reported for Accuracy, F1 Score, REM F1 Score, AUROC, and Cohen's Kappa. The baseline configuration (No finetuning) yields an accuracy of 0.740, an F1 score of 0.741, a REM F1 score of 0.572, an AUROC of 0.853, and a Cohen's Kappa of 0.489, serving as a reference point without any additional layer adjustments. When only the classifier is finetuned, the overall accuracy and F1 score increase slightly to 0.745 ± 0.015, while the REM F1 score improves to 0.591 ± 0.053, with an AUROC of 0.862 ± 0.011 and Cohen's Kappa of 0.505 ± 0.022. Finetuning the attention mechanism along with the classifier results in further improved overall accuracy (0.757 ± 0.017) and F1 (0.752 ± 0.020), but the REM F1 score decreases to 0.581 ± 0.058, suggesting that adjustments in the attention layers may slightly favor REM (a minor class) at the expense of overall performance. Expanding the finetuning to include the LSTM, in addition to the attention and classifier, leads to substantial gains with accuracy and F1 rising to 0.773 ± 0.012 and 0.766 ± 0.016, respectively, and improvements in AUROC (0.877 ± 0.012) and Cohen's

Kappa (0.524 ± 0.029), alongside a moderate increase in REM F1 to 0.611 ± 0.039. However, when the TCN is also finetuned (i.e., finetuning on TCN+LSTM+Attention+Classifier), overall performance remains similar (accuracy of 0.771 ± 0.015, F1 of 0.764 ± 0.019, AUROC of 0.872 ± 0.014, and Cohen's Kappa of 0.521 ± 0.034), while the REM F1 score decreases to 0.599 ± 0.041, indicating that the TCN layers might not favor REM detection. The best performance is achieved when all layers are finetuned, yielding an accuracy of 0.785 ± 0.016, an F1 score of 0.780 ± 0.017, a REM F1 score of 0.631 ± 0.046, an AUROC of 0.888 ± 0.015, and a Cohen's Kappa of 0.553 ± 0.027. This analysis reveals the trend that as more layers are made tunable, overall performance metrics tend to improve; however, the REM F1 score does not increase monotonically. This discrepancy suggests that certain layers—potentially those within the feature extractor or sequence modeling modules—may favor the detection of the minority REM class, while others primarily boost general classification performance. Such trade-offs are crucial when optimizing models for both overall accuracy and class-specific detection.

D.3. Alblation: Different Pretraining Dataset Combinations

Supplementary Table 7: WatchSleepNet performance metrics with different combinations of pretraining datasets, including without pretraining.

Pretraining Dataset	Acc	F1	REM F1	AUROC	Cohen's Kappa
SHHS+MESA	0.785 ± 0.016	0.780 ± 0.017	0.631 ± 0.046	0.888 ± 0.015	0.553 ± 0.027
SHHS Only	0.761 ± 0.019	0.751 ± 0.024	0.574 ± 0.062	0.870 ± 0.020	0.492 ± 0.043
MESA Only	0.761 ± 0.0134	0.752 ± 0.017	0.561 ± 0.050	0.871 ± 0.011	0.496 ± 0.030
No pretraining	0.655 ± 0.019	0.519 ± 0.024	0.000 ± 0.000	0.498 ± 0.025	0.000 ± 0.000

Supplementary Table 7 provides an ablation study evaluating WatchSleepNet's performance when pretrained on different datasets, demonstrating the effect of combining multiple pretraining sources on final model performance. The table reports five key metrics: Accuracy, F1 Score, REM F1 Score, AUROC, and Cohen's Kappa, each highlighting distinct aspects of the model's classification ability. When pretrained on both SHHS and MESA datasets, WatchSleepNet achieves the highest performance, with an accuracy of 0.785 ± 0.016, an F1 score of 0.780 ± 0.017, a REM F1 score of 0.631 ± 0.046, an AUROC of 0.888 ± 0.015, and a Cohen's Kappa of 0.553 ± 0.027, indicating strong generalization and agreement, alongside robust REM detection. Using only the SHHS dataset for pretraining yields a lower but still respectable performance, with accuracy at 0.761 ± 0.019, F1 at 0.751 ± 0.024, REM F1 at 0.574 ± 0.062, AUROC at 0.870 ± 0.020, and Cohen's Kappa at 0.492 ± 0.043, showing a noticeable drop in REM-specific performance. Pretraining solely on MESA results in further declines, with accuracy at 0.761 ± 0.0134, F1 at 0.752 ± 0.017, REM F1 at 0.561 ± 0.050, AUROC at 0.871 ± 0.011, and Cohen's Kappa at 0.496 ± 0.030, revealing a substantial weakening in REM classification and overall agreement. Without any pretraining, performance deteriorates significantly, with accuracy at 0.655 ± 0.019, F1 at 0.519 ± 0.024, REM F1 at a near-zero 0.000 ± 0.000, AUROC at 0.498 ± 0.025, and Cohen's Kappa at 0.000 ± 0.000, underscoring the critical role of pretraining for effective feature learning, particularly for REM stages. This analysis clearly demonstrates that combining multiple pretraining sources—SHHS and MESA—enhances WatchSleepNet's final performance across all metrics, most notably in REM F1 and Cohen's Kappa, likely due to the broader diversity of the combined datasets. This suggests that leveraging varied pretraining data helps the model better capture the

underlying patterns in the target domain, leading to greater predictive power compared to single-source or no pretraining approaches.

Appendix E. WatchSleepNet Performances for Different Apnea Severity Groups

Supplementary Table 8 presents the performance metrics of a 3-stage sleep stage classification model tested across four populations with varying sleep apnea severity levels: normal, mild, moderate, and severe. The model utilizes the IBI signal type captured through a wristwatch and finetunes all layers in the DREAMT dataset. Overall, the metrics such as accuracy, F1 score, AUROC, and Cohen's Kappa show minimal variance across different apnea levels, maintaining robust performance regardless of apnea severity. However, the REM F1 score decreases from 0.6863 in moderate apnea cases to 0.5696 in severe cases, highlighting the difficulty in consistently predicting REM sleep stages in individuals with severe apnea. This decline is likely due to the low percentage of REM epochs in people with severe apnea, which poses a challenge for the model in accurately detecting and classifying REM sleep.

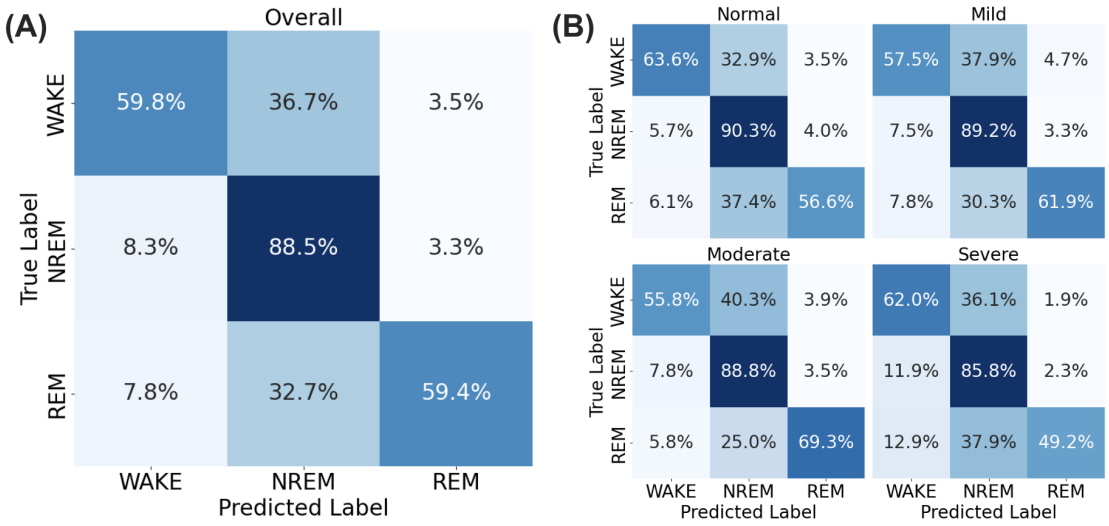
Supplementary Table 8: Performance comparison of models finetuned on the DREAMT dataset using wristwatch IBI signals. Key metrics include accuracy, F1-score, REM F1, AUROC, and Cohen's Kappa, providing insights into model performance for different apnea levels when predicting sleep stages. Results are presented for each severity group

Testing Dataset	Apnea Severity Category	Accuracy	F1	REM F1	AUROC	Kappa
DREAMT	Normal	0.7939	0.7878	0.6156	0.9009	0.5901
	Mild	0.7902	0.7849	0.6525	0.8918	0.5587
	Moderate	0.7884	0.7827	0.6863	0.8862	0.5574
	Severe	0.7690	0.7657	0.5696	0.8650	0.5058

E.1. WatchSleepNet performance across participants with different severity levels

Supplementary Figure 1 (A) displays the confusion matrix of the model's performance for the entire DREAMT population, while Supplementary Figure 1 (B) shows the confusion matrices for subpopulations categorized by different sleep apnea severity levels. Each cell in these matrices represents the frequency with which one sleep stage is classified as another, offering insight into patterns of error. The confusion matrices reveal a consistent trend where approximately 35% of the misclassifications occur between NREM and REM sleep stages. These misclassifications are likely due to physiological similarities in certain features of the IBI signals during the transitions between these stages.

In patients with severe apnea, the model struggles more to accurately classify REM sleep. The precision of REM is 56.8% in No Apnea group, 61.9% in the Mild Apnea group, an nd 69.3% in the Moderate Apnea group while the precision of REM is 49.2% in the severe apnea group. The substantial rise in misclassifying REM as NREM sleep suggests that severe apnea alters REM sleep characteristics. This sharp increase in misclassification in severe apnea may be potentially due to fragmented REM sleep or overlapping features with NREM sleep in this population, and deserve further investigation. Overall with all apnea severity groups, the highest misclassifications occur between REM and NREM stages, indicating that distinguishing between these two stages is particularly challenging for the model in cases of severe apnea.



Supplementary Figure 1: Confusion matrices illustrating the performance of the sleep stage classification model on the entire DREAMT dataset. (a) Confusion matrix displaying the model's classification accuracy across all sleep stages for the overall population. (b) Confusion matrices highlighting the model's performance across sleep stages and providing insights into potential misclassifications between different stages.