POLITECNICO DI TORINO

Master's Degree Course in Mathematical Engineering

Master's Degree Thesis

Sleep stages classification using deep learning on wearable sensors data in patients with sleep disorders

ETH zürich

Supervisors
Prof. Tania Cerquitelli
Prof. Robert Riener
Dr. Diego Paez-Granados
Dr. Oriella Gnarra

Candidate
Camilla Massato

Academic Year 2024-2025

Summary

Sleep plays a fundamental role in physical and mental health, and poor-quality sleep is associated with a wide range of chronic conditions. However, sleep disorders remain underdiagnosed and undertreated in many cases, in part due to the limited accessibility of accurate sleep monitoring. Although Polysomnography (PSG) is the gold standard for sleep monitoring, it is limited to controlled clinical settings, requires trained professionals, and is typically restricted to a single night of sleep that may not reflect the typical sleep behavior of the patient.

Recent advances in wearable technology offer new opportunities for continuous, unobtrusive, and home-based sleep monitoring. This thesis explores a deep learning-based approach to automatic sleep staging using physiological signals collected from wearable devices to provide accessible tools for personalized sleep assessment, particularly for individuals with suspected or diagnosed sleep disorders.

The proposed method uses a Convolutional Neural Network (CNN) architecture, the U-Sleep, trained on multimodal data, namely Acceleration (ACC), Blood Volume Pulse (BVP), Electrodermal Activity (EDA), and Skin Temperature (TEMP), recorded using the Empatica E4 wristband. The dataset includes 127 participants with simultaneously recorded PSG and wearable data, allowing for direct performance comparison with the clinical standard.

Results show that the model can accurately estimate key sleep parameters, with Bland-Altman analysis revealing good agreement for Sleep Efficiency (SE), REM Latency (RL), Wake After Sleep Onset (WASO), and durations of REM and Deep sleep. Epoch-by-epoch concordance reached an accuracy of 0.87±0.07 for Wake, 0.90±0.04 for REM, 0.71±0.07 for Light, and 0.89±0.04 for Deep sleep. Overall accuracy and F1-score were 0.69±0.08 and 0.62±0.11 for the whole dataset, and 0.77±0.05 and 0.74±0.06 for healthy participants, respectively.

To assess the robustness and generalizability of the model, additional experiments were conducted on external datasets and with a different wearable device. These evaluations confirmed the adaptability of the model to various data sources and configurations, highlighting its potential for scalable application in real-world contexts.

In conclusion, this work demonstrates the feasibility of using multimodal wearable data for personalized sleep staging. By capturing individual-specific physiological patterns, the proposed approach supports the use of deep learning for precision sleep health, with potential applications in long-term monitoring, early detection, and management of sleep disorders in a clinical setting.

Acknowledgements

I would like to thank the Spinal Cord Injury & Artificial Intelligence (SCAI) Lab and the Sensory-Motor Systems (SMS) Lab at ETH Zürich for giving me the opportunity to do my thesis there, and above all, for allowing me to grow both professionally and personally.

My deepest gratitude goes to my supervisor, Dr. Oriella Gnarra, for being such a great mentor in both the good and challenging moments of my stay in Zürich, and for being an inspiration for my future career as a woman in STEM.

Finally, I want to thank Lala, Yoshi, Samuel, and Reja for contributing to my best memories in Switzerland — I will never forget any of it.

Contents

List of Tables

List of Figures

Acronyms

AASM American Academy of Sleep Medicine

ACC Acceleration

ACT Actigraphy

AI Artificial Intelligence

BVP Blood Volume Pulse

BD Breathing Disorders

CI Confidence Interval

CNN Convolutional Neural Network

DREAMT Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology

EEG Electroencephalogram

EOG Electrooculogram

E4 Empatica E4

EMG Electromyography

EDA Electrodermal Activity

HR Heart Rate

HRM Heart Rate Mean

HRSD Heart Rate Standard Deviation

IBI Inter-Beat Interval

IQR Interquartile Range

LLMs Large Language Models

LSTM Long Short-Term Memory

NIH National Institute of Health

NREM Non-Rapid Eye Movement

PPG Photoplethysmography

PSG Polysomnography

PTSD Post-Traumatic Stress Disorder

REM Rapid Eye Movement

RL REM Latency

RLS Restless Leg Syndrome

RNN Recurrent Neural Network

SLAMSS Sequence-to-sequence LSTM for Automated Mobile Sleep Staging

SA Sleep Apnea

OSA Obstructive Sleep Apnea

SE Sleep Efficiency

SMS Sleep Monitoring Study

SOL Sleep Onset Latency

SOTA State Of The Art

SD Standard Deviation

STFT Short-Time Fourier Transform

SVM Support Vector Machine

SWEZ Sleep-Wake Epilepsy Centre

SWS Slow-Wave Sleep

TEMP Skin Temperature

TST Total Sleep Time

WASO Wake After Sleep Onset

Chapter 1

Introduction

1.1 The importance of sleep

Sleep is an essential biological process marked by a substantial reduction in physical and mental activity, altered consciousness, and decreased muscle tone. It is regulated by circadian rhythms, influenced by environmental conditions such as light and temperature, and is governed by complex neurobiological mechanisms. On average, sleep occupies about one-third of human life, highlighting its fundamental role in maintaining physiological balance and overall health.

Sleep quality and structure, particularly the alternation between Non-Rapid Eye Movement (NREM) and Rapid Eye Movement (REM) stages, are crucial for cognitive and physical functioning. Restorative sleep supports memory consolidation, attention, emotional regulation, immune defense, tissue repair, and hormonal balance.

In contrast, poor or inadequate sleep has been linked to a wide range of negative effects. Zheng et al. [1] associate irregular sleep patterns with a higher risk of chronic conditions like cardiovascular disease, type 2 diabetes, and obesity. Cognitive problems, such as reduced concentration, slower reaction times, and memory lapses, are also common, and chronic sleep disorders can contribute to mental health disorders and increase the risk of accidents. Moreover, reductions in Slow-Wave Sleep (SWS) and REM sleep have been associated with structural brain changes in older adults, including brain atrophy [2], suggesting that sleep quality may be a risk factor for neurodegenerative diseases such as Alzheimer's.

Given its pervasive impact on health, monitoring sleep patterns is a key tool in the early detection and prevention of related disorders. Promoting good sleep should be a central goal in public health strategies.

1.1.1 Sleep Stages

According to the American Academy of Sleep Medicine (AASM), sleep is divided into five distinct stages that alternate cyclically throughout the night. These include a wake stage, three NREM stages—N1, N2, and N3—and one REM stage. Together, these stages form the sleep architecture, which typically repeats every 90 to 120 minutes in healthy adults.

The characteristics of each sleep stage are the following:

- Stage N1 marks the transition from wakefulness to sleep and represents light sleep, as individuals can be easily awakened in this phase. During this stage, brain activity begins to slow, muscle tone decreases slightly, and eye movements are slow and rolling.
- Stage N2 is a deeper form of light sleep and constitutes the largest proportion of total sleep time. It is characterized by a further slowing of brain activity and the appearance of reduced response to external stimuli. Body temperature drops, heart rate slows, and muscles become more relaxed.
- Stage N3, also known as SWS or deep sleep, is the most restorative stage. It is dominated by high-amplitude, low-frequency delta waves. During N3, muscle tone, breathing rate, and blood pressure reach their lowest levels. This stage is critical for physical recovery, immune function, and memory consolidation.
- REM sleep, contrarly to NREM stages, is characterized by rapid eye movements, low muscle tone (near paralysis), and brain activity similar to wakefulness. This is the stage where most vivid dreaming occurs. REM sleep plays a vital role in emotional regulation, mostly in the consolidation of procedural and emotional memories.

Sleep typically begins in N1 and progresses through N2 and N3 before entering REM. Over the course of the night, the proportion of REM sleep increases, while N3 sleep becomes less prominent. N1 and N2 are collectively referred to as light sleep, and N3 as deep sleep. For the purpose of this work, the four stages considered will be: Wake, REM, Light (N1 + N2), and Deep (N3).

1.1.2 Sleep Disorders

Sleep disorders are a group of medical conditions that alter sleep quality, timing, and duration, often leading to significant daytime distress. They can manifest as difficulty falling or staying asleep, excessive daytime sleepiness, abnormal behavior during sleep, or disruptions of the body's internal clock. According to the National Institute of Health (NIH), sleep disorders affect approximately 30–40% of the global population, making them a worldwide public health concern. These conditions can arise from various causes, including genetic predispositions, neurological disorders, lifestyle factors, or other medical conditions, and their impact ranges from mild disturbances to severe, chronic conditions.

The most common sleep disorders are the following:

- Sleep-related breathing disorders: conditions characterized by abnormal respiration patterns or insufficient ventilation during sleep. These disorders can lead to fragmented sleep, reduced oxygen levels, and daytime fatigue. One common example is Obstructive Sleep Apnea (OSA), which causes relaxation of the tongue that blocks the airway during sleep.
- Insomnia: difficulty falling, staying asleep, or waking up too early and being unable to return to sleep. People with insomnia often experience daytime fatigue, irritability, poor concentration, and reduced performance.
- Central disorder of Hypersomnolence: difficulty staying awake during the day despite getting a normal or long amount of nighttime sleep.

- Parasomnias: collection of unusual sleep behaviors, movements, emotions, perceptions, or dreams that occur while falling asleep, during sleep, or while waking. These events are typically involuntary and can be disruptive to the individual or bed partner. For instance, REM sleep behavior disorder, often coexisting with other neurological disorders, causes people to act out during dreams.

- Circadian rhythm disorders: shifted sleeping hours, mismatching the normal day-night schedule. This misalignment can lead to difficulty falling asleep, waking up at socially acceptable times, or maintaining regular sleep-wake patterns.

Understanding and identifying these disorders is essential not only for clinical diagnosis but also for the development of specific interventions and personalized therapies. Since many sleep disorders alter the architecture and distribution of sleep stages, several sleep measures - such as sleep onset latency and REM duration - can be valuable biomarkers. These quantitative indicators provide insights into underlying pathological processes and can be used for early detection, treatment monitoring, and risk stratification in both clinical and research contexts.

1.2 Polysomnography

PSG is considered the gold standard for sleep monitoring and clinical diagnosis of sleep disorders. It involves the simultaneous recording of multiple physiological signals during sleep, including Electroencephalogram (EEG), Electrooculogram (EOG), and Electromyography (EMG), among others. These signals provide a comprehensive overview of brain activity, eye movements, and muscle tone, which are essential for the accurate sleep staging.

Despite its diagnostic accuracy, PSG presents several limitations. The process requires sleep technicians or clinicians to manually annotate sleep stages by visually inspecting the recorded signals. This manual scoring is time-consuming and inevitably subjective, leading to variability in classification. For instance, inter-rater agreement in four-stage sleep classification has been reported to be around 83.7%, according to Nikkonen et al. [3], who analyzed 50 PSG recordings scored by 10 experts from 7 different sleep centers, involving both healthy individuals and patients with sleep disorders. Furthermore, PSG is typically conducted in a controlled hospital or sleep laboratory environment, requiring the attachment of numerous sensors to the patient's head, face, chest, and limbs. This setup can interfere with natural sleep patterns, as the unfamiliar setting and the physical discomfort of the equipment may lead to altered or unrepresentative sleep behavior [4]. In addition, the procedure is labor-intensive, expensive, and not scalable for big populations and long-term use [5].

As a result, recent research has increasingly focused on developing more accessible and less intrusive alternatives to traditional PSG. These approaches aim to facilitate large-scale, cost-effective sleep monitoring while maintaining clinical relevance. Such advancements are particularly important not only for the early detection of sleep disorders, but also for understanding their potential associations with other medical conditions, including neurodegenerative diseases [6].

1.3 Wearable Devices

The increasing demand for accessible and non-intrusive sleep monitoring in real-world settings has driven the development and spread of wearable and nearable technologies. Devices such as smartwatches, wristbands, sensor-embedded mattresses, and bedside radars have emerged as promising tools for capturing sleep-related physiological data outside the clinical environment. These technologies can be classified as either consumer or research devices, depending on their use, data accessibility, and validation. They typically record signals such as Heart Rate (HR), ACC, TEMP, and in some cases EDA or Photoplethysmography (PPG), from which sleep stages or measures can be estimated.

While such technologies offer the potential for scalable and cost-effective sleep assessment, significant challenges remain. The most challenging among them is improving the accuracy of sleep detection algorithms and maximizing the clinical relevance of the data these devices produce. Many commercial devices rely on proprietary algorithms and have limited ability to distinguish between light, deep, and REM sleep stages with sufficient precision. These limitations are highlighted by a prospective multicenter validation study that evaluated 11 commercially available sleep trackers, revealing substantial variability in their performance and limited agreement with gold-standard PSG scoring for sleep stage classification [7].

Moreover, a substantial number of wearable systems lack proper validation in clinical populations, particularly among individuals with sleep disorders, who have atypical sleep architecture. As emphasized by Zambotti et al. [8], precise benchmarking against PSG in diverse datasets is essential to establish the reliability and generalizability of these technologies across different applications.

In this study, two research wearable devices are employed to explore the feasibility of sleep monitoring through wrist-worn sensors:

- Empatica E4: a wristband device designed for research applications that provides high-resolution raw physiological data, including 3-axis ACC, BVP, EDA, and TEMP.

- Corsano: a research-grade wristband providing both raw physiological signals and sleep stage estimations, offering an integrated approach to sleep tracking.

1.4 Research Goals

The main objective of this study is to develop a deep learning-based model capable of classifying sleep stages using physiological data collected from wrist-worn research devices, with a particular focus on individuals affected by sleep disorders. This research is motivated by the need to bridge the gap between the high accuracy of clinical sleep monitoring methods and the practicality of wearable technologies. Specifically, the aim is to design a model that surpasses the classification performance of currently available consumer-grade devices, while maintaining scalability, accessibility, and reduced obtrusiveness, which are key limitations of standard PSG.

Given the increasing relevance of real-world sleep monitoring, the study also emphasizes clinical applicability. By focusing on populations with diagnosed sleep disorders, it addresses a significant gap in wearable validation studies, which often rely on healthy participants. The

methodology integrates multiple sources of physiological data to improve stage discrimination and tests the robustness of the proposed solution across datasets and device types.

The specific research objectives can be summarized as follows:

1. Integration of EDA and TEMP: Investigate the impact of adding EDA and TEMP to traditional sensor modalities (e.g., ACC and PPG), focusing on signal preprocessing and model optimization. The goal is to evaluate whether these additional signals enhance the model's ability to differentiate between sleep stages.

2. Diagnosis-specific performance analysis: Assess the model's performance across different diagnostic subgroups (e.g., insomnia, sleep apnea, parasomnias), together with healthy participants, to identify potential strengths and limitations. This analysis will help understand how sleep disorders affect wearable-based stage classification for future model adaptations in clinical settings.

3. Use of augmented training data: Combine data from the initial dataset with an open-source dataset acquired using the same research device, in order to improve model generalization and robustness through exposure to a wider range of sleep patterns and inter-subject variability.

4. Cross-device generalization: Evaluate the trained model on a separate dataset acquired from a different wrist-worn device, to explore the model's generalizability across devices.

Overall, this work aims to contribute to the development of more accurate and clinically relevant wearable-based sleep monitoring systems, potentially paving the way for future applications in remote diagnostics, personalized sleep health, and early detection of sleep conditions.

Chapter 2

Material and Methods

2.1 Datasets

2.1.1 SMS

The first dataset analyzed in this work is part of the Sleep Monitoring Study (SMS) [9], made up of 127 participants recorded for one night in an examination room at the Sleep-Wake Epilepsy Centre (SWEZ) of the University Hospital of Bern, Switzerland. Clinical metadata with demographic information, including the number of participants, percentage of females, average age, and percentage of each diagnosis of sleep disorder, is presented in Table 2.1.

The prevalent diagnosis in the dataset is sleep-related breathing disorders, such as OSA, and central disorders of hypersomnolence. These are followed by insomnia, parasomnias (such as REM sleep behavior disorder), sleep-related movement disorders (such as periodic limb movement disorder), and some cases of circadian disorder. Moreover, the dataset contains missing diagnoses and includes participants belonging to other unspecified sleep disorders. Healthy controls refer to individuals who were evaluated in the sleep clinic but did not receive a formal diagnosis of any sleep disorder. Although some of these participants may have atypical sleep-related characteristics, they are considered within the range of normal physiological variation and classified as healthy for the purposes of this study.

The participants' sleep was monitored with a wristband, Empatica E4 (E4), placed on the non-dominant arm together with PSG. One night of sleep is recorded for each participant. The choice of the non-dominant wrist aims to reduce movement artifacts and improve the reliability of the collected data. The E4 suits this work due to its non-invasive nature, portability, and ability to continuously collect multiple physiological signals, making it a suitable device for sleep research conducted in both clinical and research environments.

The data provided by the wristband are shown in Table 2.2 with their respective range, unit of measures, frequencies, and type of data, including the ACC expressed as a 3-axis accelerometer, which captures body movements and can help infer activity levels or restlessness during sleep; the BVP derived from the PPG signal, which uses light absorption to measure blood volume changes and allows for the estimation of heart rate and heart rate variability; the HR estimated from the BVP signal; the Inter-Beat Interval (IBI) which is the time interval between

Table 2.1: Demographics of the SMS dataset

Dataset	**Full**	**Train**	**Evaluation**	**Test**
Participants (n)	127	86	15	26
Females (%)	54.89	56.04	53.33	51.85
Age (mean ± SD)	45.3±16.2	46.2±16.4	44.6±15.7	42.2±16.1
Diagnosis (%)				
Breathing disorders	59.1	60.5	60.0	53.9
Hypersomnolence	15.0	14.0	13.3	19.2
Insomnia	5.5	5.8	6.7	3.9
Missing diagnosis	4.7	4.7	6.7	3.9
Parasomnias	4.7	4.7	6.7	3.9
Healthy controls	4.7	3.5	0.0	11.5
Other disorders	4.7	5.8	6.7	3.9
Movement disorders	1.6	1.2	0.0	3.9
Circadian disorders	1.5	2.2	0.0	0.0

Demographics of SMS dataset, including number of participants, percentage of females, age as mean ± Standard Deviation (SD) expressed in years, and percentage of sleep disorders diagnoses. The values are shown for the full dataset and for the training, evaluation, and test sets.

individual beats, derived from the HR; the EDA, measured by a perspiration sensor, which reflects changes in skin conductance associated with sympathetic nervous system arousal; and the TEMP, measured by an infrared thermopile, which tracks peripheral skin temperature and may provide insight into circadian rhythms and thermoregulation. These signals offer a comprehensive overview of the participants' physiological state during sleep, allowing for the identification of patterns related to autonomic nervous system activity and therefore to sleep stages. Moreover, the E4 data were later synchronized with the PSG recordings to allow comparative analysis and assess the potential of wearable technology to approximate gold standard sleep measurements.

A visual representation of the physiological signals collected through the E4 is provided in Figure 2.1. The plot illustrates the temporal evolution of each signal over the course of the recorded night. On the x-axis, time is expressed in hours, allowing for an intuitive understanding of how the physiological parameters vary throughout the sleep period. The y-axis reports the respective units of measurement for each signal, corresponding to those listed in Table 2.2. To facilitate interpretation and highlight potential correlations between physiological dynamics and sleep architecture, the signals are segmented and colored according to the alternating sleep stages identified through PSG. Each color segment corresponds to a specific sleep stage (e.g.,

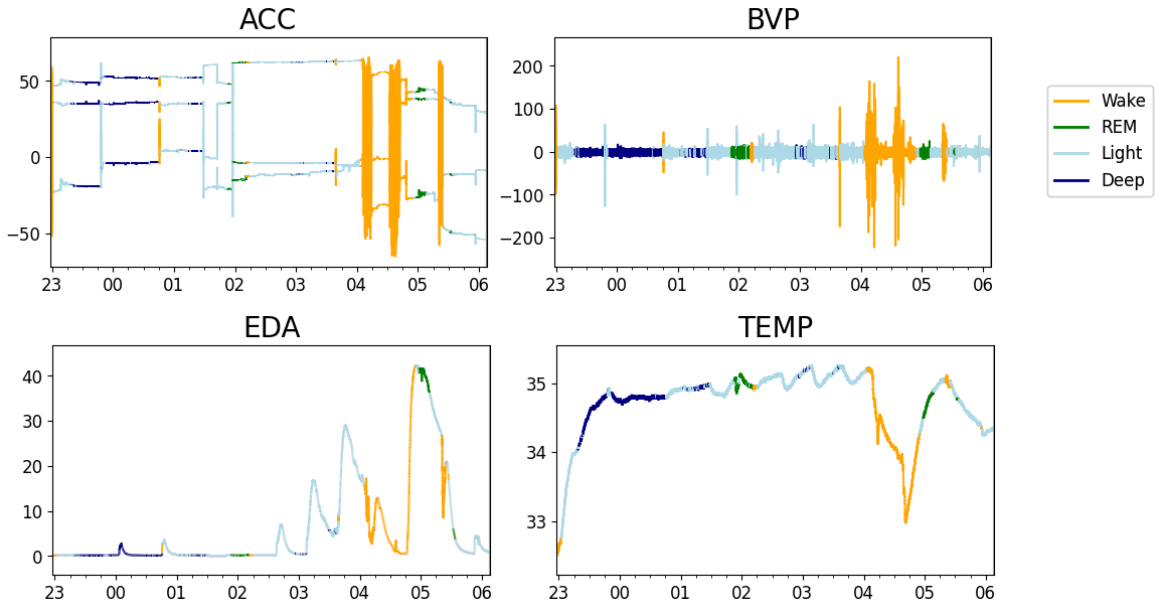
Table 2.2: Empatica E4 signals

Signal	Range	Unit of measure	Frequency	Type of signal
ACC	±2	g	32 Hz	3-axial
BVP	±500	a.u.	64 Hz	PPG sensor-derived
HR	60–200	bpm	1 Hz	BVP-derived
IBI	Varies with HR	ms	Not fixed	HR-derived
EDA	0.01–100	µS	4 Hz	Continuous signal
TEMP	16–42	°C	4 Hz	Continuous signal

Overview of the signals provided by the Empatica E4, including range, unit of measure, frequency, and type of signal. Units of measure: g = gravitational acceleration; a.u. = arbitrary units; bpm = beats per minute; ms = milliseconds; µS = microsiemens; °C = degrees Celsius.

Wake, REM, Light, Deep), enabling a direct visual association between the physiological variations and the sleep pattern. This representation supports a more comprehensive analysis of how certain biosignals fluctuate in relation to the sleep cycle.

Figure 2.1: Signal plots coloured according to sleep stages



Physiological signals recorded by the device worn by one participant suffering from OSA; each signal is coloured according to the corresponding sleep stages.

The plot includes only the ACC, BVP, EDA, and TEMP signals, which are the primary focus of this work. The derived signals HR and IBI, although available, are excluded from the study due to their direct dependence on the BVP signal and their lower sampling frequency, which makes them redundant or less informative in this context.

2.1.2 DREAMT

The Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology (DREAMT) presented in [10] is an open-source dataset from the Physionet challenge. It collects a night of sleep from 100 participants recruited from the Duke University Health System (DUHS) Sleep Disorder Lab, including signals from the E4, together with sleep technician-annotated sleep stage labels based on PSG recordings. Clinical metadata are also available, with information about participants' health and sleep disorders as shown in Table 2.3. Only the primary diagnosis is reported for each participant. Given that the sleep study protocol is specifically designed to detect and monitor apnea events during sleep, a large proportion of the cohort had already been diagnosed with OSA. Other sleep disorders such as hypersomnolence, Restless Leg Syndrome (RLS), and insomnia are also represented, together with a small group of healthy controls.

Table 2.3: Demographic of the DREAMT dataset

Metric	**Full dataset**
Participants (n)	100
Females (%)	55
Age (mean ± SD)	56.2 ± 16.6
Diagnosis (%)	
Breathing disorders	66
Hypersomnolence	20
Healthy controls	9
Restless leg syndrome	2
Insomnia	2

Demographics of DREAMT dataset, including number of participants, percentage of females, age as mean ± SD expressed in years, and percentage of sleep disorders diagnoses. The values are shown for the full dataset.

All six raw signals from the E4, 3-axial ACC, BVP derived from the PPG sensor, EDA, and TEMP, HR and IBI estimated from the raw BVP are provided by the dataset. All the data are available in the highest sampling frequency of 64 Hz, from the original frequencies already shown in Table 2.2. The actual timestamp is time-shifted and starts with 0 to preserve privacy.

Sleep stage labels are provided every 30 seconds and are derived from technician-annotated PSG data. The following stages are included: Wake (W), Non-REM stages N1, N2, and N3,

and REM (R). An additional label, Preparation (P), marks the initial period before the PSG recording officially begins. In this work, the data were trimmed to exclude the preparation phase, considering the lights-off time as the first annotated stage following the "P" label, and the lights-on time as the last available stage. This ensures that only the actual sleep period is analyzed, removing non-representative segments.

A special label, Missing, is present in rare cases where no sleep stage annotation is available. This label appears very infrequently across the dataset. Specifically, significant missingness is observed only in two participants due to a PSG re-setup during the night, resulting in approximately 15 minutes of continuous missing labels. In four other participants, only one epoch labeled as "Missing" was found.

This open-source dataset was selected not only to expand our dataset of sleep disorder participants and improve performance, but also because of its clinical richness. The availability of expert-labeled sleep stages from PSG, combined with synchronized wearable data (E4), allows for robust multimodal learning and validation. Furthermore, the diversity in sleep disorder diagnoses and the inclusion of clinical metadata make it a valuable resource for improving generalization and model interpretability across heterogeneous populations.

2.1.3 Corsano

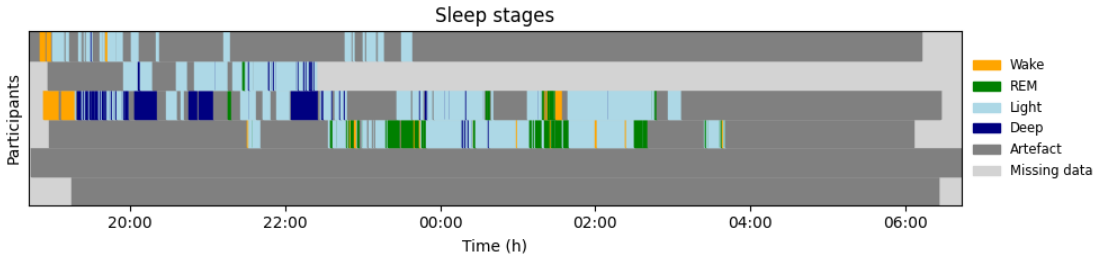
The dataset from Corsano is a work-in-progress dataset, mainly used in this work to demonstrate the generalizability of the model to different devices. The current version of the dataset provides only ACC and PPG signals, sampled at 32 Hz and 128 Hz, respectively. Although the Corsano device is capable of recording EDA, TEMP, and its own sleep stage estimations, which could be used for comparison with the model, these data are unfortunately not available in the version of the dataset currently accessible.

The first step in the data preparation involved aligning the timestamps of the Corsano recordings, which were collected in Japan, with Swiss local time. Since the recordings often spanned entire days and, in some cases, multiple days per participant, the focus is only on the nocturnal periods that corresponded to the lights off and lights on times recorded in the PSG. This restriction was necessary because PSG ground truth data was only available for the first night of recording.

Figure 2.2 shows the sleep stage labels for multiple participants across the night. Each row represents one participant, and colors indicate different sleep stages or data quality states. It is evident that the dataset is affected by a substantial amount of artifacts (in dark grey) and missing data (in light grey). These issues limit the amount of usable data for testing the model. For example, some participants have large gaps of unrecorded or corrupted data spanning several hours, particularly in the early or late portions of the night.

Due to this limitation, only the third participant from the top, who exhibits a relatively consistent and artifact-free signal throughout the night, was selected for use in the test set. This participant is the only one with more than 50% valid epochs, also displaying a representative sleep architecture, including periods of Wake, REM, Light, and Deep sleep, providing a meaningful test case for model generalization.

Figure 2.2: Sleep stages from Corsano dataset



Sleep stage annotations for all Corsano participants during the first night of recording. Each row corresponds to one participant, with colors indicating Wake (orange), REM (green), Light (light blue), Deep (blue), Artefact (dark grey), and Missing data (light grey).

2.2 Deep Learning Architecture

The deep learning architecture tested in this work is a fully CNN based on U-Net [11], [12], called U-Sleep [13], which is a State Of The Art (SOTA) algorithm for sleep stage classification from PSG-based EEG and EOG.

CNNs are a class of deep learning models particularly well-suited for processing structured signals such as images and time-series data. In the context of sleep stage classification, CNNs are widely used because of their ability to automatically extract relevant features from raw physiological signals, such as EEG and EOG. These signals often contain characteristic local patterns, such as specific brain waveforms associated with different sleep stages, that CNNs can effectively capture through their convolutional layers.

CNNs are also capable of learning both spatial and temporal dependencies in the data while maintaining a relatively low number of parameters, which makes them efficient and less prone to overfitting, particularly important when dealing with noisy physiological data and limited training samples.

Although CNN-based models were originally applied to PSG signals, recent works, including [14], have shown that similar architectures can also be effectively applied to wearable-derived signals.

The U-Sleep network architecture is composed of four main modules, each responsible for a specific stage in the signal processing and classification:

- *Conformation Module*. This module preprocesses the input signal by applying zero-padding and reshaping operations to ensure compatibility with the network architecture. It standardizes input length and format, facilitating efficient batch processing.

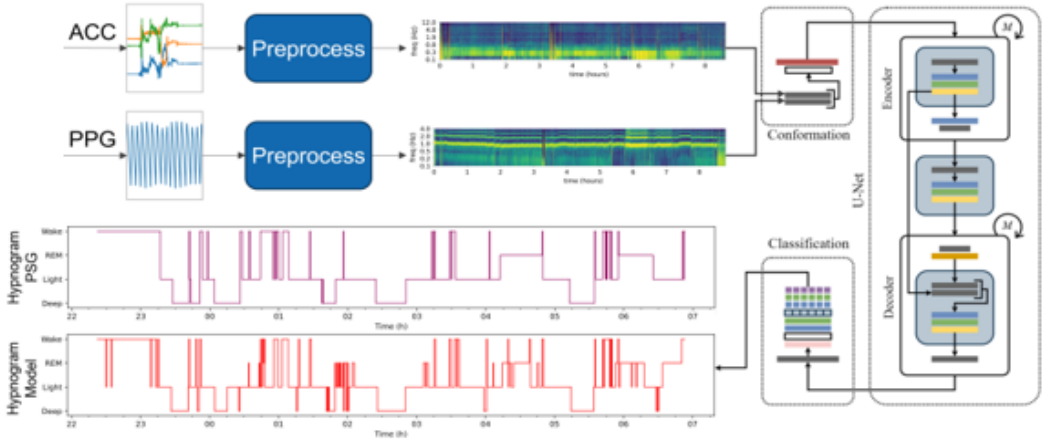
- *Encoder Module*. It is a hierarchical stack of convolutional layers that progressively extracts local and high-level features from the input while reducing the temporal resolution. It captures both short and long-range dependencies and condenses the signal into a compact latent representation.

- *Decoder Module*. It is a symmetric structure with respect to the encoder, composed of transposed convolutions or upsampling layers that restore the compressed representation back to the original temporal resolution. It enables the network to produce temporally aligned outputs by integrating encoded features with contextual information.
- *Segment Classifier Module*. Being the final stage of the network, it maps the decoded representation to discrete sleep stage predictions. It splits the output into individual epochs and assigns class labels based on learned temporal patterns.

All the modules were introduced in U-Sleep [13], except for the conformation module presented by Olsen et al. [14]. In this last study, the model was tested in another wearable device, using as input ACC and PPG signals, both present in the SMS dataset.

In Figure 2.3, the CNN architecture is illustrated, including the input signals, preprocessing step, network structure, and the resulting hypnogram, which is compared to PSG.

Figure 2.3: U-Sleep architecture



Overview of the U-Sleep CNN architecture. The diagram illustrates the input wearable signals (ACC and PPG) as in [14], preprocessing steps, the four main processing modules of the network, and the output hypnogram used for sleep stage classification, compared against PSG-based ground truth.

The U-Net architecture was selected for its effectiveness in time series segmentation tasks, such as sleep stage classification, where maintaining temporal resolution and capturing local context are essential. Its encoder-decoder structure allows the model to extract both short-term dynamics and long-range dependencies by integrating multiscale feature representations. This capability is particularly important in this domain, as stage transitions often rely on intricate temporal patterns spanning multiple physiological signals. This model was chosen for these reasons and preferred over the CNN previously tested on the E4 device by Li et al. [15], due to its flexibility to accept any type and number of input time series data and its slightly better SOTA performance.

Although recent approaches based on attention-based architectures and Large Language Models (LLMs) have demonstrated strong performance on sequential data tasks [16], they typically rely on extensive pre-training using large datasets and require substantial computational

resources, factors that are often impractical in clinical or wearable applications. In contrast, our selected U-Net-based architecture can be trained on relatively small datasets, making it a more feasible option that better fits the real-world constraints of sleep monitoring in patients with various disorders using wearable devices.

The key contributions of this work include the adaptation of the network proposed in [14], experiments with various signal combinations, identification of optimal parameters, and a comprehensive model evaluation. Additionally, this study applies the model to a clinical dataset comprising participants diagnosed with sleep disorders, highlighting its potential applicability in real-world, pathological sleep conditions.

In [14], Olsen et al. used only ACC and PPG data from a consumer wearable device. Empatica E4 also provides other signals, including EDA and TEMP, as well as derived measurements, such as BVP. In this work, different combinations of input signals were considered.

2.3 Preprocessing

The time series data were first cut between the lights-off and lights-on times according to PSG annotations. From the starting 133 participants in the SMS dataset wearing the Empatica E4, six were excluded because of a failed temperature sensor, resulting in 127 participants in the final analysis, as shown in Table 2.1. The specific preprocessing applied to each signal is presented in the following subsections.

2.3.1 cvxEDA

The EDA signal was processed following the approach proposed in [17], in which the raw signal is decomposed into its tonic and phasic components using a physiologically inspired convex optimization framework. In *cvxEDA*, the observed EDA signal $y(t)$ is modeled as the sum of three components:

$$y(t) = r(t) + p(t) + \varepsilon(t) \tag{2.1}$$

- $r(t)$ is the tonic component, modeled as a smooth, low-frequency signal using a cubic spline with knots spaced over time;
- $p(t)$ is the phasic component, obtained by convolving a sparse neural activation signal $u(t)$ with a biexponential impulse response $h(t)$;
- $\varepsilon(t)$ is Gaussian white noise.

The goal is to recover $u(t)$, $r(t)$, and consequently $p(t)$, by solving the following convex optimization problem:

$$\min_{r,u} \quad \frac{1}{2}\left\|y - r - h * u\right\|_2^2 + \lambda \left\|u\right\|_1 + \frac{\gamma}{2}\left\|D^2 r\right\|_2^2 \tag{2.2}$$

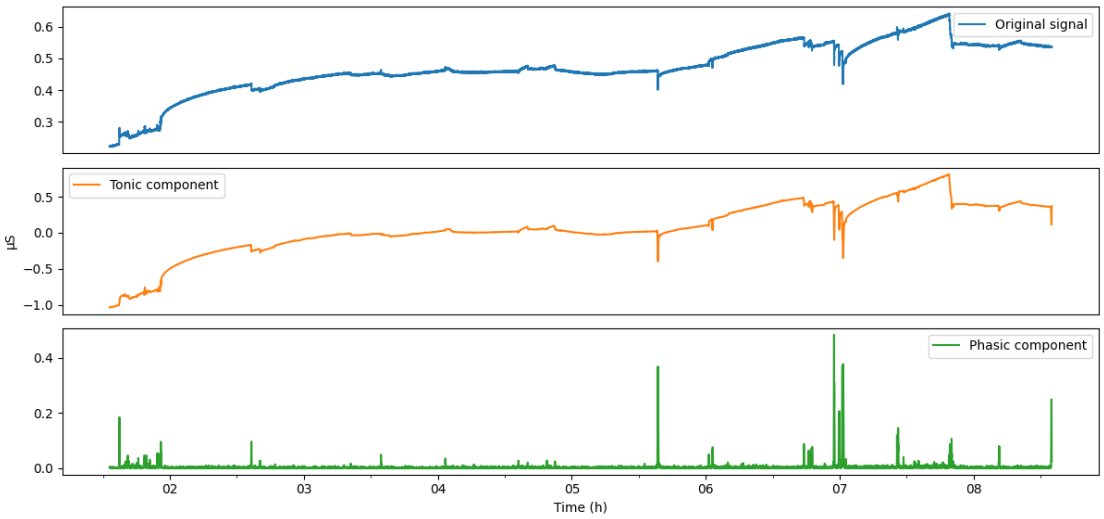
$$\text{s.t.} \quad u(t) \geq 0 \quad \forall t$$

where:

- $\|y - r - h * u\|_2^2$ ensures fidelity to the observed signal;
- $\|u\|_1$ promotes sparsity in the neural activation;
- $\|D^2 r\|_2^2$ penalizes curvature in the tonic component, enforcing smoothness;
- λ and γ are regularization parameters.

This formulation allows for a robust and interpretable decomposition of the EDA signal. A visual result example from one participant in the SMS dataset is shown in Figure 2.4. The plot shows a comparison between the original EDA signal with its tonic and phasic components during a night of sleep, with time expressed in hours on the x-axis, and respective EDA values in microsiemens (μS) on the y-axis.

Figure 2.4: EDA decomposition



Plot of EDA original signal (blue), together with its tonic (orange) and phasic (green) components, obtained with the cvxEDA algorithm. The x-axis represents the time expressed in hours (h), while the y-axis is expressed in microsiemens (μS).

Only the phasic component of EDA $p(t) = h * u(t)$ was retained as input to the CNN, as it more directly reflects the fast, event-related dynamics of sympathetic nervous system activity. It has also been shown to be more strongly correlated with brain rhythms during sleep with respect to the tonic component [18], in particular being positively correlated with theta waves (dominant during light sleep) and negatively correlated with delta waves (dominant during deep sleep).

2.3.2 Delta TEMP

The TEMP signal was transformed by computing the first-order difference (delta) between consecutive samples, defined as:

$$\Delta T(t) = T(t) - T(t-1) \tag{2.3}$$

where $T(t)$ is the temperature signal at time t, and $\Delta T(t)$ represents the delta-transformed temperature [19].

This transformation was motivated by the limitations of applying spectrogram-based analysis, which is introduced in the following section, to slowly varying, low-frequency signals. Spectrograms derived from the raw temperature signal tends to exhibit highly redundant and uninformative patterns due to the signal's low temporal dynamics. In contrast, the delta-transformed signal introduces mid-range frequency components that are better aligned with the sensitivity of time–frequency representations commonly used in deep learning models.

The transformed signal also helps reduce baseline effects, that is, slow or constant signal components that do not carry meaningful physiological information, as well as intersubject variability. As a result, the model becomes more sensitive to relevant physiological fluctuations. These include thermoregulatory responses such as peripheral vasoconstriction, which are commonly associated with REM sleep and other transitions in sleep architecture [20].

2.3.3 Common preprocessing

The ACC and BVP signals were first preprocessed to reduce inter-subject variability and stabilize signal amplitude. Specifically, median centering was applied to the ACC signal to eliminate the influence of movement-related baseline shifts, while z-score normalization was used for the BVP signal to standardize its distribution (i.e., transformation to zero mean and unit variance), ensuring comparability across samples and reducing the impact of amplitude-related variance. Mathematically, z-score normalization is defined as:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{2.4}$$

where x_i is the original signal value, μ is the mean, and σ is the standard deviation.

Next, both signals underwent adaptive Interquartile Range (IQR) normalization using a 300-second sliding window. This technique dynamically scales the signal based on the local dispersion, with IQR defined as:

$$\text{IQR}_w = Q_3^w - Q_1^w \tag{2.5}$$

where Q_1^w and Q_3^w denote the first and third quartiles within the window w. The signal is then normalized with respect to the local median $\tilde{x}^w$ as:

$$x_i' = \frac{x_i - \tilde{x}^w}{\text{IQR}_w} \tag{2.6}$$

This approach is robust to non-Gaussian noise and local trends. To limit the influence of extreme values, data points beyond a threshold of 20 times the local IQR were clipped. This empirical threshold follows prior work on wearable sensor data [14].

Similarly, the transformed EDA and TEMP signals were normalized using median normalization, followed by adaptive IQR normalization and outlier clipping. The clipping thresholds were set to 20 times the IQR for EDA and 15 times the IQR for TEMP, based on their signal-specific variability profiles.

To ensure temporal alignment across modalities and datasets, all signals were resampled to a uniform frequency of 32 Hz. Specifically:

- For the SMS dataset, the PPG signal was downsampled from 64 Hz to 32 Hz, and the EDA and TEMP signals were upsampled from 4 Hz to 32 Hz.
- For the DREAMT dataset, all signals were downsampled from 64 Hz to 32 Hz.
- For the Corsano dataset, only the PPG signal was downsampled from 128 Hz to 32 Hz.

Finally, all preprocessed signals were transformed into spectrograms via the Short-Time Fourier Transform (STFT), enabling representation in the time-frequency domain. The STFT is defined as:

$$\text{STFT}\{x(t)\}(t, f) = \int_{-\infty}^{+\infty} x(\tau) w(\tau - t) e^{-j2\pi f \tau} d\tau \tag{2.7}$$

where $w(\tau - t)$ is a windowing function (e.g., Hann window) centered at time t. This representation captures how the frequency content of a signal evolves over time, highlighting transient and rhythmic patterns typical in physiological processes.

The use of time-frequency representations is particularly beneficial when working with CNNs, highly effective at learning local patterns and features in structured spatial data. Spectrograms provide a two-dimensional grid-like input, analogous to images, which aligns well with the inductive biases of CNN architectures.

2.4 Evaluation

The SMS dataset was stratified by diagnosis, as reported in Table 2.1, and split into training (65%), validation (15%), and test (20%) sets. Similarly, the combined SMS + DREAMT dataset, where the latter was added to expand the population, was stratified and split using the same proportions, as shown in Table 2.4.

To ensure consistent input dimensions, recordings were segmented into consecutive, non-overlapping windows, each including a fixed number of 30-second epochs. This segmentation strategy, commonly used in the sleep stage classification framework, has multiple purposes: it standardizes the input data to fixed-length sequences compatible with deep learning models, which typically require uniform input shapes; it reduces variability introduced by differing recording durations across subjects; and it enables parallel batch processing during training, improving computational efficiency. Moreover, this window length aligns with standard clinical practice for sleep stage annotation. According to the AASM guidelines, human experts assign sleep stages based on 30-second epochs. This duration provides a good balance between temporal resolution and physiological relevance, capturing enough signal variation to differentiate between stages.

Table 2.4: SMS + DREAMT dataset split

Diagnosis	**Full dataset (227 participants)**	**Training set (65%)**	**Validation set (15%)**	**Test set (20%)**
Breathing disorder	62.56	63.23	62.96	60.00
Hypersomnia	17.18	18.06	3.70	20.00
Healthy controls	6.17	5.81	7.41	8.89
Insomnia	3.96	4.52	7.41	-
Parasomnias	2.64	1.94	7.41	2.22
Missing diagnosis	2.64	1.92	7.41	2.22
RLS	0.88	1.29	-	-
Movement disorders	0.88	0.65	-	-
Others	3.08	2.59	-	4.44

Percentage distribution of sleep disorders in the combined SMS + DREAMT dataset and its split into training (65%), validation (15%), and test (20%) sets.

Following preliminary computational experiments, the input segment length was set to 1024 epochs, approximately 8.5 hours of sleep, which typically covers an entire night of recording. This length was selected as a compromise between capturing a complete sleep cycle and maintaining computational tractability. Recordings shorter than this length were zero-padded to preserve the fixed input shape, while longer recordings were truncated. The padding was masked during model training to minimize the impact of artificial zeros on learning dynamics.

Hyperparameter tuning was performed using HyperBand [21], a resource-efficient method that dynamically allocates training time to promising configurations while discarding underperforming ones early. This approach significantly accelerates the search process compared to traditional grid or random search, especially in high-dimensional hyperparameter spaces. The parameters to optimize were:

- M, the number of encoder–decoder blocks, which directly affects model capacity and the depth of feature abstraction; higher values allow the model to learn more complex representations, but may increase the risk of overfitting.

- K, the kernel height of the 2D convolutions, which influences the temporal and spatial resolution of learned features; adjusting this value can help the model better capture relevant patterns in signals of varying duration or structure.

- The initial filters number in the first encoder layer, which sets the starting model complexity; this parameter determines the richness of the feature maps extracted in early stages, and can impact both performance and computational cost.

- The focusing parameter α of the loss function, which adjusts the model's attention to hard-to-classify samples; tuning this value is especially useful in imbalanced classification tasks, as it allows the model to prioritize underrepresented or more difficult classes.

- The learning rate, which controls convergence speed and stability; an appropriate value ensures efficient training without divergence or getting stuck in local minima.

Optimization was performed using the ADAM optimizer [22], known for its robustness in training deep architectures.

2.4.1 Loss function

In the context of sleep stage classification, class imbalance poses a significant challenge, as some stages (e.g., REM) are underrepresented compared to others like light sleep. To address this issue, a binary focal loss was employed, which enhances the model's focus on misclassified, harder examples.

Focal loss, originally introduced for object detection tasks, modifies the standard binary cross-entropy by incorporating a modulating factor that reduce the weights of well-classified examples. In this implementation, the loss also adapts dynamically to batch-level class distributions, leveraging the positive sample ratio p_w to modulate the contribution of each class during training. This design helps to prevent the majority class, in this case light sleep, from dominating the learning process and encourages the model to learn more robust features for minority classes.

The loss for a batch of size N is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\Big[(1-p_w)^{\alpha}\cdot\mathbf{1}_{y_i=1}\cdot\log(p_i) + p_w^{\alpha}\cdot\mathbf{1}_{y_i=0}\cdot\log(1-p_i)\Big] \tag{2.8}$$

where:

- p_i: predicted probability for sample i;

- y_i: true label for sample i (0 or 1);

- p_w: positive sample ratio in the batch;

- α: focusing parameter that controls the strength of emphasis on hard examples;

- $\mathbf{1}(\cdot)$: indicator function selecting the relevant term based on the ground truth label.

This formulation ensures that the gradient contribution from well-predicted examples is reduced, preventing overfitting to the dominant classes. The focusing parameter α allows for a finer control over this effect: higher values of α place more emphasis on correcting misclassifications. By computing p_w per batch, the method remains adaptive to fluctuations in class distributions across batches, further improving robustness.

2.4.2 Performance metrics

To evaluate model performance, a combination of standard classification metrics and domain-specific sleep-related measures was employed. These metrics were chosen to ensure comparability with SOTA methods, to capture performance across all classes despite class imbalance, and to ensure clinical relevance in sleep staging applications.

Due to the imbalance in sleep stage distributions, where stages such as light sleep dominate while stages like REM are underrepresented, it is essential to report not only overall accuracy but also per-class metrics that highlight model behavior on minority classes.

Per-stage (i.e., single-class) performance was evaluated using the following metrics:

- Accuracy measures the overall correctness of the classification model, i.e., the proportion of total predictions (both positive and negative) that are correct:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.9}$$

  where TP (true positives) and TN (true negatives) are correctly predicted instances, while FP (false positives) and FN (false negatives) represent misclassifications. This metric gives a general sense of how well the model performs across all classes.

- Sensitivity (also known as recall or true positive rate) indicates the model's ability to correctly identify instances of a specific sleep stage when it is actually present:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.10}$$

  A higher sensitivity means the model rarely misses true occurrences of the target class (i.e., few false negatives).

- Specificity (true negative rate) reflects the model's ability to correctly recognize when a specific sleep stage is not present:

$$Specificity = \frac{TN}{TN + FP} \tag{2.11}$$

  High specificity implies that the model makes few false positive predictions, thus avoiding incorrect detections of the sleep stage when it is absent.

- F1-score is the harmonic mean of precision and recall, and it provides a balanced measure that takes into account both false positives and false negatives:

$$F1\text{-}score = 2 \cdot \frac{Pr \cdot Re}{Pr + Re} \tag{2.12}$$

  where $Re = \frac{TP}{TP+FN}$ is the recall (sensitivity), and $Pr = \frac{TP}{TP+FP}$ is the precision. The F1-score is particularly useful when the class distribution is imbalanced, as it penalizes extreme values in either precision or recall.

These metrics were computed per class to better understand model sensitivity to each individual sleep stage, particularly important for clinical stages (e.g., REM or N1) whose accurate detection is crucial for diagnosis and treatment.

Overall multi-class performance was quantified using:

- Balanced Accuracy provides an average of recall (sensitivity) values computed for each sleep stage, thus offering a performance measure that is robust to class imbalance:

$$Balanced\ Accuracy = \frac{1}{4} \cdot \sum_{i=1}^{4} Sensitivity_i \tag{2.13}$$

  where $Sensitivity_i$ is the sensitivity for the i-th sleep stage. By giving equal weight to each class regardless of their frequency, balanced accuracy ensures that the performance on underrepresented stages is not overshadowed by dominant ones.

- Macro-F1 is the arithmetic mean of the F1-scores computed independently for each sleep stage. Unlike the standard (Micro) F1-score, it does not take class frequency into account, making it especially suitable when evaluating models on imbalanced datasets:

$$Macro\text{-}F1 = \frac{1}{4} \cdot \sum_{i=1}^{4} F1 - score_i \tag{2.14}$$

  where $F1 - score_i$ is the F1-score for the i-th class. This metric provides a class-balanced view of performance.

- Cohen's Kappa (κ) measures the agreement between the model's predictions and the PSG-based annotations, while adjusting for the agreement that could occur by chance:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{2.15}$$

  Here, p_0 denotes the observed agreement, i.e., the proportion of sleep epochs where the predicted and true labels match. The expected agreement p_e estimates the probability of random agreement based on the marginal distributions of predicted and actual labels. A κ value of 1 indicates perfect agreement, whereas 0 implies no better agreement than chance.

Metrics such as Cohen's κ are particularly informative in sleep staging, as they align with inter-rater reliability measures used in clinical settings. High κ values indicate strong agreement with expert annotations, which is essential for model adoption in diagnostic contexts. Furthermore, macro-averaged metrics ensure that underrepresented but clinically important stages (e.g., REM) are not hidden by the majority class (N2), providing a fair and interpretable evaluation.

2.4.3 Sleep measures

In addition to classification metrics, clinically relevant sleep measures were computed to evaluate the utility of the model in practical, real-world settings such as sleep medicine and home monitoring. While classification metrics provide a detailed account of the model's per-epoch performance, they do not directly reflect the overall sleep architecture, which is an essential aspect for both diagnostic and research purposes. Therefore, sleep summary metrics were derived to assess how well the model can reconstruct key features of a typical night of sleep, including timing, duration, and structure of various stages. These included:

- Total Sleep Time (TST): the total duration of all epochs classified as any sleep stage (excluding wake), providing a measure of overall sleep quantity.
- Sleep Onset Latency (SOL): the time elapsed from lights-off (the start of the recording) to the first epoch classified as sleep, indicating the subject's ability to enter sleep.
- REM Latency (RL): the interval between sleep onset and the first epoch of REM sleep.
- Sleep Efficiency (SE): the ratio between total sleep time and the total recording time, typically expressed as a percentage; it reflects how efficiently the time in bed is spent sleeping.
- Wake After Sleep Onset duration (WASO_d): the cumulative duration of wake epochs occurring after the first sleep epoch, indicating the degree of sleep fragmentation during the night.
- REM sleep duration (REM_d): the total duration of epochs classified as REM sleep, reflecting the extent of REM sleep achieved across the night.
- Light sleep duration (Light_d): the total time spent in light sleep stages (N1 + N2), which represent the transitional and lighter phases of non-REM sleep.
- Deep sleep duration (Deep_d): the cumulative time spent in deep sleep (N3 stage), associated with restorative physiological processes and slow-wave activity.

These measures were derived from the predicted hypnograms and compared to those computed from the PSG annotations, allowing for an assessment of the model's ability to recover meaningful sleep architecture. Beyond their descriptive value, such measures serve to evaluate sleep quality and are in clinical routine reports. Therefore, accurate estimation of these quantities is a critical step toward validating the model's practical relevance and its potential for integration into tools for remote or automated sleep monitoring.

Chapter 3

Results

Hyperparameter tuning identified the optimal configuration as $M = 10$, $K = 16$, initial filters = 32, $\alpha = 0.15$, and learning rate = 10^{-3}.

Performance analysis for this configuration using the SMS datasets is presented in the following section, following the standardized framework introduced in [23]. This approach includes a sleep measure analysis, with a summarizing table to compare the proposed model with the ground truth represented by PSG and Bland-Altman plots; an epoch-by-epoch analysis to present the performance metrics, both at a single stage and global level, commonly used in the classification framework; and finally the distribution of the sleep stages comparing that from PSG and the model.

Subsequently, additional results are reported for the combined SMS + DREAMT dataset, including an epoch-by-epoch analysis and a learning curve to show the effectiveness of adding training data. Finally, epoch-by-epoch results for the Corsano test participant are shown.

3.1 SMS dataset

The first experiments with the SMS dataset were about finding the best signal combination in terms of global performance metrics. The reason is that the model has already been tested with ACC and PPG only, but the E4 also provides other signals, namely EDA and TEMP. These signals are proven to be correlated to the sleep pattern, so the first goal of this work is to add these signals as input to the model. An overview of the results in terms of global metrics for the most meaningful signal combinations is shown in Table 3.1. It includes the values expressed as mean $\pm$ SD of Accuracy, Balanced Accuracy, Macro-F1, and Cohen's κ.

In Figure 3.1, four confusion matrices are shown, comparing the model to the ground truth provided by PSG. Each confusion matrix corresponds to the model results for a different combination of input signals, starting from the model with ACC only, then adding BVP, EDA, and TEMP once at a time. The figure highlights how adding each signal changes each class's performance, including the values for precision, recall, and F1-score. The main matrix displays the counts and normalized values for each class. The right column shows recall for each class, the bottom row shows precision, and the bottom-right cell shows the macro-averaged F1-score.

As it can be seen from both Table 3.1 and Figure 3.1, the best results are obtained for the

Table 3.1: Results for different signal combinations

ACC	BVP	EDA	TEMP	Accuracy	Bal. Acc.	Macro-F1	Cohen's κ
x				0.57 ± 0.09	0.50 ± 0.08	0.46 ± 0.08	0.31 ± 0.11
x	x			0.67 ± 0.07	0.64 ± 0.09	0.61 ± 0.09	0.48 ± 0.12
x	x		x	0.66 ± 0.08	0.62 ± 0.09	0.59 ± 0.11	0.46 ± 0.14
x	x	x		0.68 ± 0.08	0.63 ± 0.11	0.60 ± 0.12	0.49 ± 0.14
x	x	x	x	**0.69 ± 0.08**	**0.64 ± 0.10**	**0.62 ± 0.11**	**0.50 ± 0.14**

Global metrics results express as mean ± SD for different signal combinations from SMS dataset.

model with all signals combined, including ACC, BVP, EDA, and TEMP, for which a more detailed analysis is shown in the following subsections.

3.1.1 Sleep measures analysis

In Table 3.2, a detailed comparison is presented between sleep measures obtained from PSG and those estimated by the proposed model. For each sleep measure, the table reports the mean and SD, the 95% Confidence Interval (CI) of the mean, and p-values from statistical tests. When the distribution of paired differences was normal, a paired t-test was applied; otherwise, the Wilcoxon signed-rank test was used. The difference is considered statistically significant if p-value < 0.05. The evaluated measures include the overall sleep measures: TST, SOL, RL, SE, WASO duration, and the duration of individual sleep stages (REM, light, and deep). These numerical results provide an overview of the agreement between the model's estimates and the PSG reference, highlighting potential differences or variability across individuals.

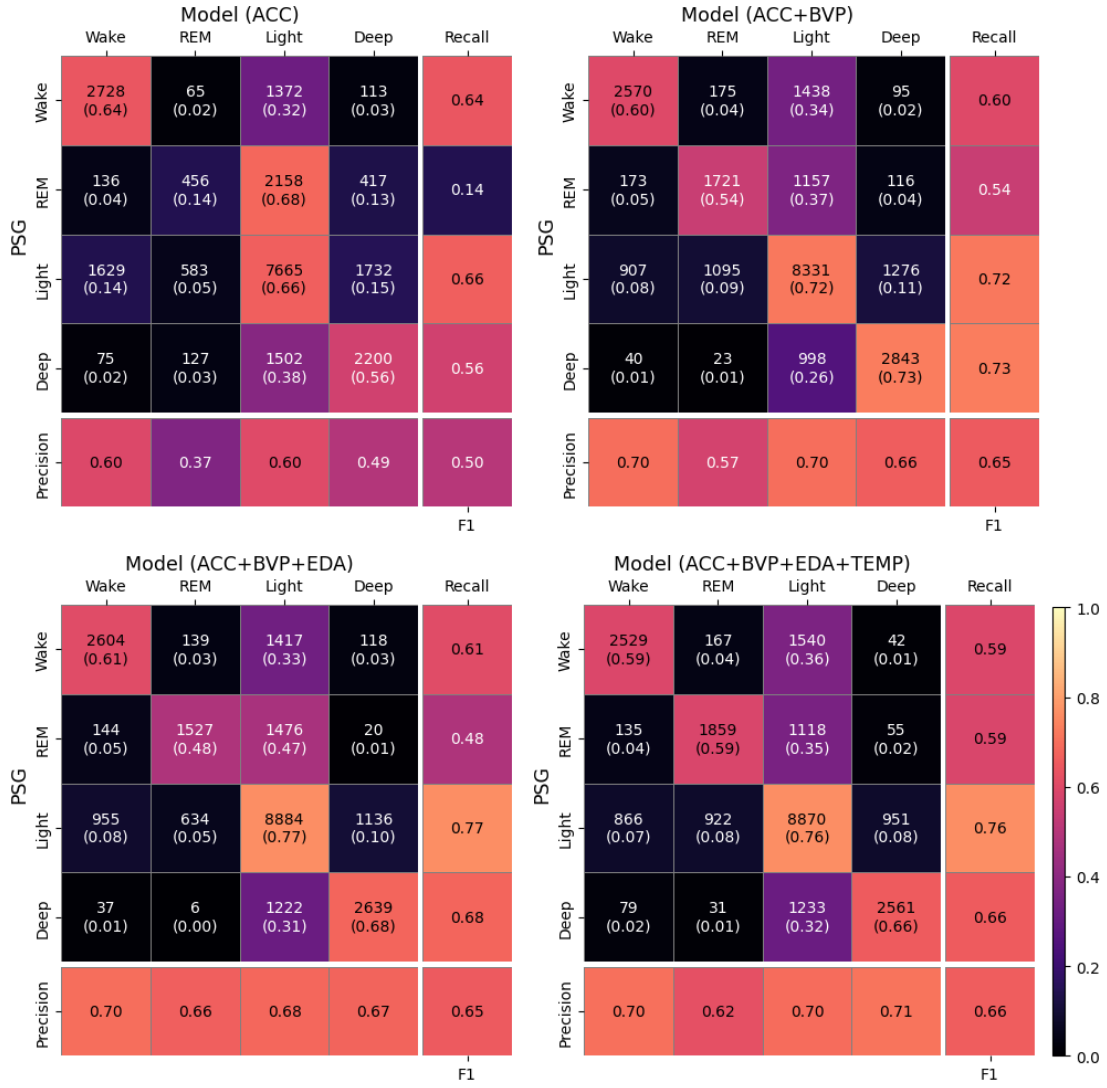
In Figure 3.2, Bland–Altman plots are shown for each sleep measure to visually assess the agreement between the model and PSG. A Bland–Altman plot is a method commonly used to compare two quantitative measurement techniques by plotting the difference between the two measurements against their mean. This allows for the identification of systematic biases and the assessment of the extent of agreement across the measurement range. In each plot, the x-axis represents the mean of the two values (model and PSG), while the y-axis shows their difference. Differences are expressed in minutes for all sleep measures, except for sleep efficiency (SE), which is reported in percentage on both axes. Each plot includes horizontal lines indicating the mean difference (bias) and the limits of agreement (mean ± 1.96·SD), which define the range within which 95% of the differences between the two methods are expected to lie. These visualizations help detect potential proportional biases or increased variability at specific ranges of the measurements, offering a complementary perspective to the statistical comparisons reported in Table 3.2.

Table 3.2: Sleep measures mean difference between PSG and the model

Sleep Measure	Metric	PSG	Model	p-value
TST (min)	Mean ± SD	359.23 ± 78.41	372.15 ± 86.50	0.24
	95% CI mean	[327.56, 390.90]	[337.22, 407.09]	
SOL (min)	Mean ± SD	18.92 ± 15.15	12.81 ± 12.29	**0.02***
	95% CI mean	[12.80, 25.04]	[7.85, 17.77]	
RL (min)	Mean ± SD	132.78 ± 76.46	107.75 ± 72.49	0.12
	95% CI mean	[101.90, 163.66]	[78.47, 137.03]	
SE (%)	Mean ± SD	74.22 ± 13.82	76.25 ± 12.78	0.20
	95% CI mean	[68.64, 79.80]	[71.09, 81.41]	
$WASO_d$ (min)	Mean ± SD	60.83 ± 41.40	54.81 ± 42.53	0.55
	95% CI mean	[44.10, 77.55]	[37.63, 71.98]	
REM_d (min)	Mean ± SD	60.90 ± 33.09	57.15 ± 30.52	0.55
	95% CI mean	[47.54, 74.27]	[44.82, 69.48]	
$Light_d$ (min)	Mean ± SD	223.25 ± 60.86	245.60 ± 61.84	**0.04***
	95% CI mean	[198.67, 247.83]	[220.62, 270.58]	
$Deep_d$ (min)	Mean ± SD	75.08 ± 31.79	69.40 ± 32.37	0.39
	95% CI mean	[62.24, 87.92]	[56.33, 82.48]	

Comparison of sleep measures between PSG and the proposed model on the test set. It includes mean $\pm$ SD and 95% CI mean for each sleep measure. All values are expressed in minutes, except for SE, which is given as a percentage. The p-value of the difference between model and PSG is also included, where * denotes statistically significant difference ($p < 0.05$).

Figure 3.1: Confusion matrices for different signal combinations



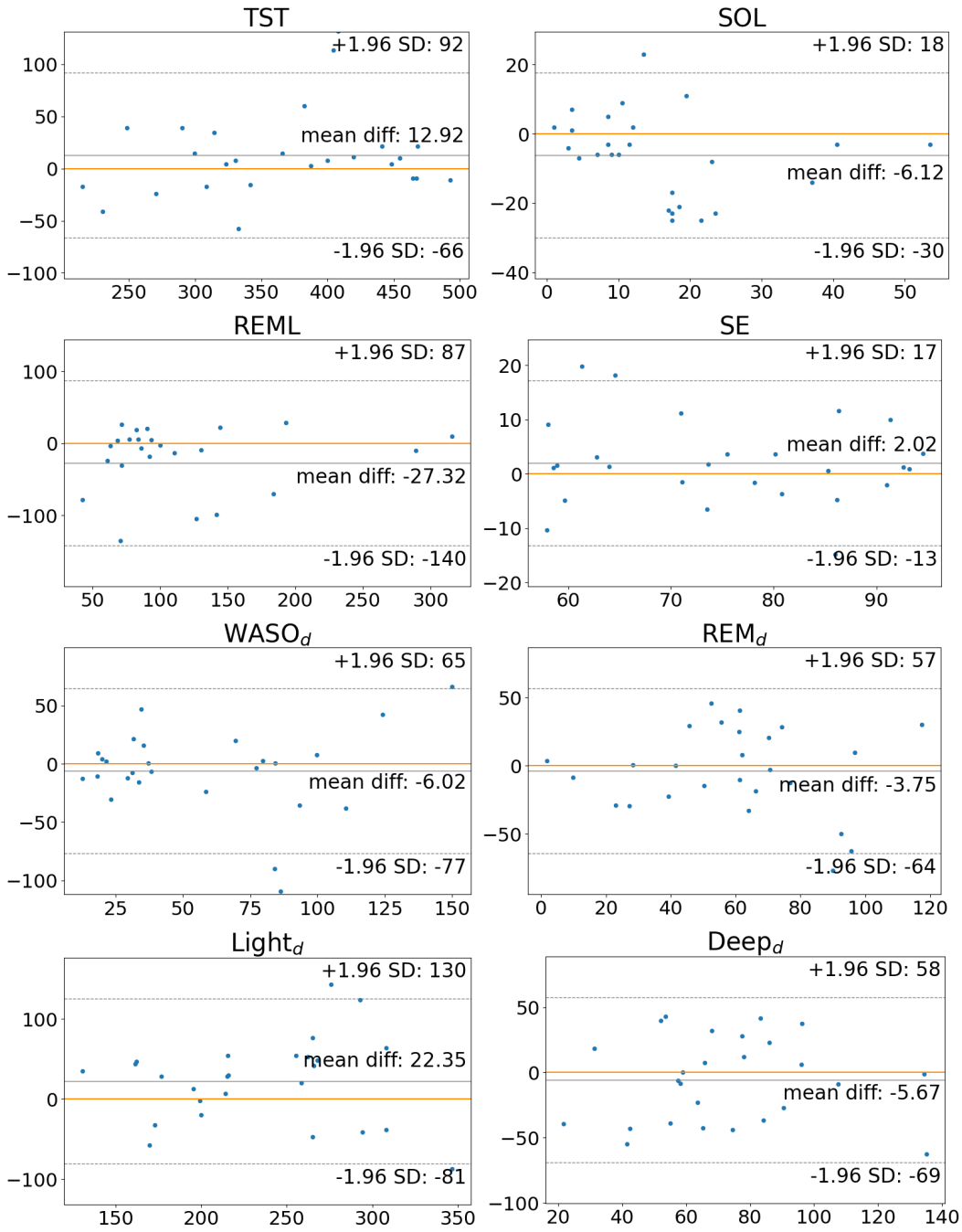
Confusion matrices illustrate the progressive improvement in classification performance as additional signals are incorporated from a single signal to the complete set of four modalities.

3.1.2 Epoch-by-epoch analysis

The epoch-by-epoch performance of the model is summarized in Table 3.3, which includes both single-class (per-stage) and multi-class evaluation metrics. All results are computed on the test set and are reported as mean ± SD across participants.

For the per-stage evaluation, performance metrics are provided separately for Wake, REM, Light, and Deep sleep stages. The reported metrics include Sensitivity, Specificity, Accuracy,

Figure 3.2: Bland-Altman plots of sleep measures



Bland-Altman plots of sleep measures of the model compared to PSG on the test set. Blue dots represent individual samples, orange horizontal lines indicate zero difference (i.e., perfect agreement), grey lines show the mean difference between the two measures.

Table 3.3: Performance metrics (SMS dataset)

Sleep stage	Sensitivity	Specificity	Accuracy	F1-score
Wake	0.59 ± 0.20	0.94 ± 0.07	0.87 ± 0.07	0.60 ± 0.17
REM	0.55 ± 0.25	0.95 ± 0.04	0.90 ± 0.04	0.53 ± 0.22
Light sleep	0.76 ± 0.11	0.66 ± 0.12	0.71 ± 0.07	0.72 ± 0.09
Deep sleep	0.66 ± 0.23	0.95 ± 0.05	0.89 ± 0.04	0.64 ± 0.15
Global	**Accuracy**	**Bal. Acc.**	**Macro-F1**	**Cohen's κ**
	0.69 ± 0.08	0.64 ± 0.10	0.62 ± 0.11	0.50 ± 0.14

Performance metrics for single-class and multi-class four-sleep stage classification on the test set of the SMS dataset. All values are expressed as mean $\pm$ SD.

and F1-score, offering a detailed assessment of the model's ability to detect each sleep stage on an epoch level. These metrics reflect the stage-wise classification performance and highlight the model's strengths and limitations in distinguishing between individual sleep stages.

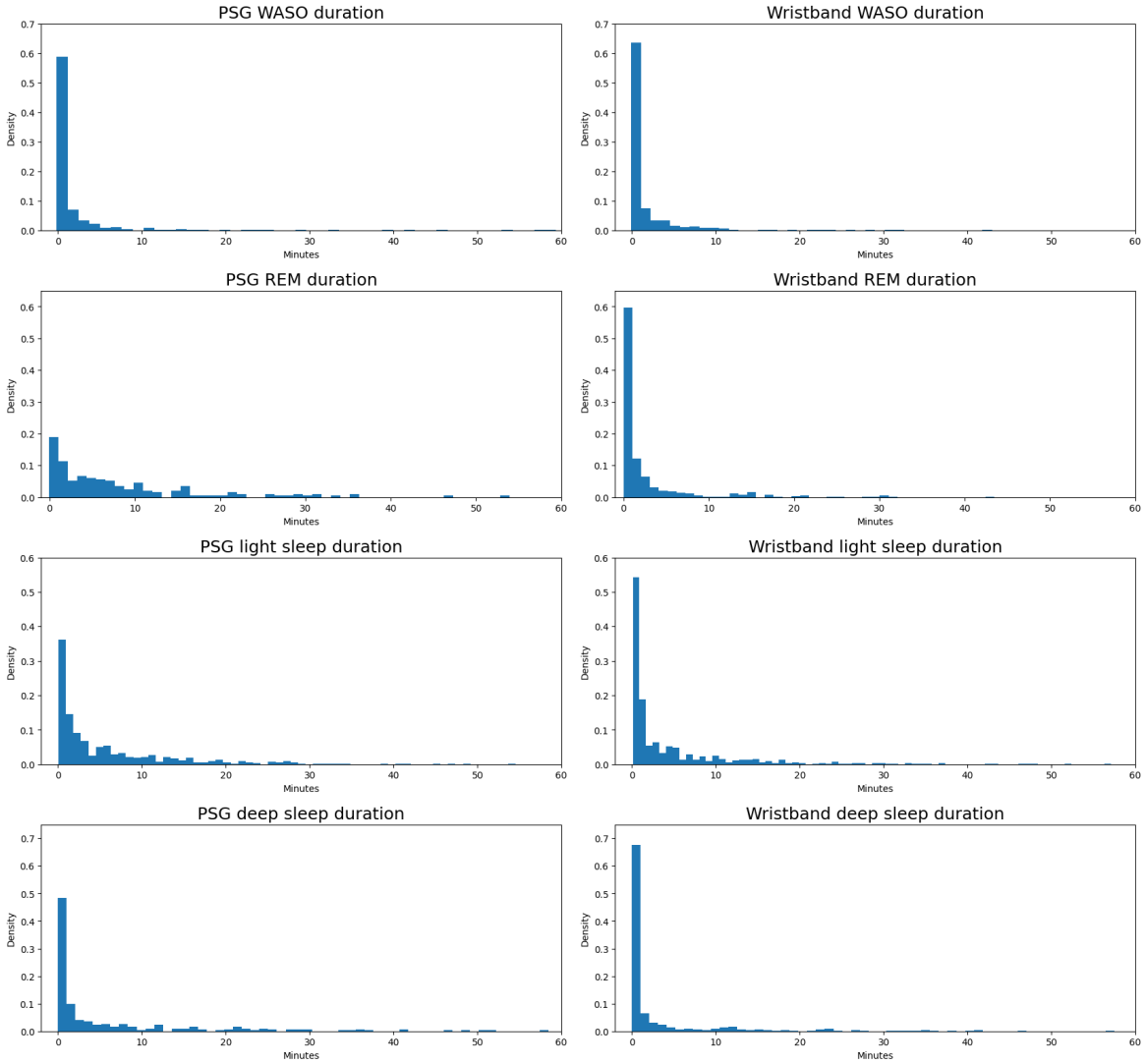
In addition to the per-stage metrics, global evaluation measures are presented to capture the overall classification performance across all sleep stages. These include overall Accuracy, Balanced Accuracy, Macro-F1 score, and Cohen's κ. While Accuracy provides a general measure of correctness, Balanced Accuracy and Macro-F1 account for class imbalance, ensuring that performance on minority classes is adequately reflected. Cohen's κ further quantifies the agreement between the model's predictions and the reference annotations, correcting for chance agreement. Together, these global metrics provide a comprehensive view of the model's effectiveness in sleep stage classification.

3.1.3 Distribution of sleep stages

The sleep stage distributions of consecutive durations are shown in Figure 3.3 for the participants in the SMS test set, comparing PSG values to the model results. Each plot shows a histogram, including the time expressed in minutes on the x-axis and the density, i.e., the normalized values, on the y-axis, so that the area integrates to one.

A clear similarity can be observed between the sleep stage distributions obtained from PSG and the E4, particularly in the case of WASO and deep sleep, where the two modalities show comparable patterns. However, some discrepancies emerge when analyzing REM sleep: the proposed model tends to produce more concentrated distributions toward shorter durations, suggesting an overestimation of REM sleep fragmentation. A similar, but less pronounced, trend is also present for light sleep, indicating that the model may slightly underestimate the continuity of the light sleep stage.

Figure 3.3: Sleep stage distributions of consecutive durations



Comparison of PSG and wristband's model distributions of consecutive durations for WASO, REM, light and deep sleep. Each histogram shows the time expressed in minutes on the x-axis and the density, i.e., the normalized values, on the y-axis, with the resulting area integrating to one.

3.1.4 Results per diagnosis

The classification performance divided per diagnosis is reported in Table 3.4. The table summarizes global performance metrics—Accuracy, Balanced Accuracy, Macro-F1, and Cohen's κ—for each primary diagnosis included in the SMS dataset. These metrics offer a comprehensive overview of the model's ability to generalize across clinically heterogeneous subgroups.

As highlighted in Table 3.4, the best performance is obtained for the healthy control participants, reaching an Accuracy of 77%, Balanced Accuracy of 73%, Macro-F1 of 74%, and Cohen's κ of 64%. Good results, comparable or even higher than overall results on the test set, are obtained for hypersomnia and breathing disorders participants, followed by movement disorder, resulting in a good Accuracy, but lower performance in terms of the other global metrics. The worst results, substantially lower than the overall model results, are obtained for insomnia and parasomnias participants.

Table 3.4: Results per diagnosis

Diagnosis	**Accuracy**	**Bal. Acc.**	**Macro-F1**	**Cohen's κ**
Breathing disorders	0.68	0.64	0.62	0.49
Hypersomnolence	0.72	0.67	0.62	0.52
Healthy controls	**0.77**	**0.73**	**0.74**	**0.64**
Movement disorders	0.69	0.52	0.52	0.53
Insomnia	0.52	0.51	0.50	0.29
Parasomnias	0.51	0.49	0.47	0.29

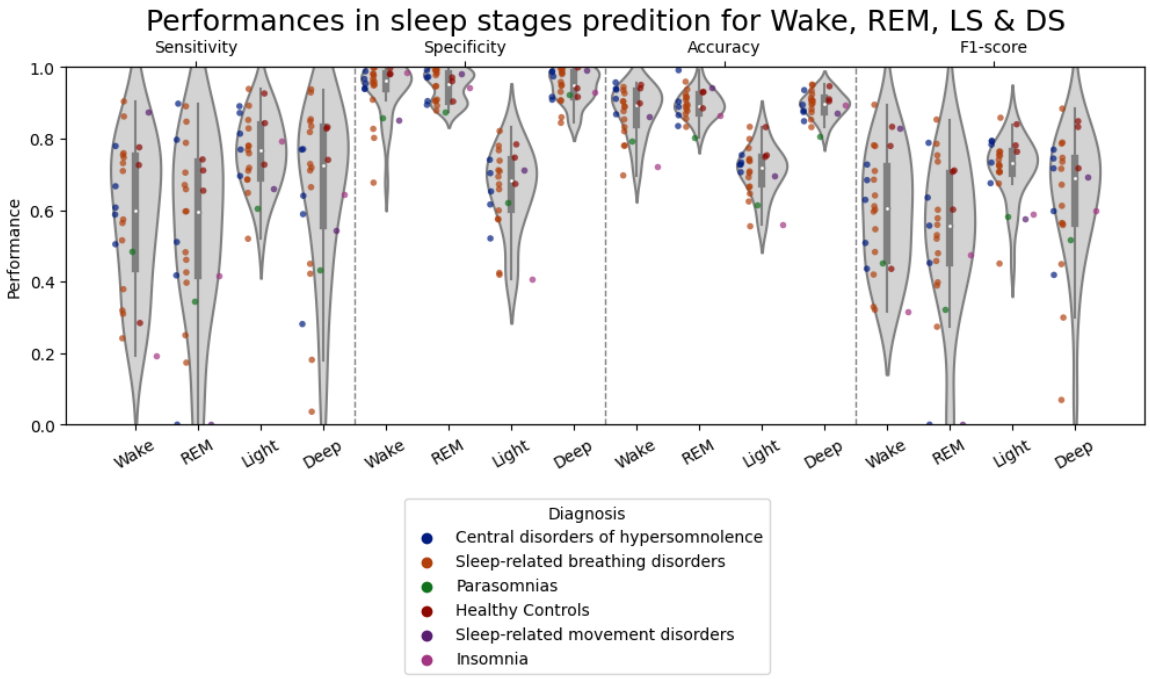
Mean classification performance for each diagnosis class, computed across all participants in the test set.

Further insight into class-wise model behavior is provided in Figure 3.4, which displays violin plots for various performance metrics, specifically the per-class metrics across individual sleep stages: Sensitivity, Specificity, Accuracy, and F1-score. A violin plot combines aspects of a boxplot with a kernel density estimate, offering a richer visualization of the distribution of data. The shape of each violin reflects the distribution of the corresponding metric, where the width at any given point represents the density of the data; wider sections indicate a higher concentration of values.

In this figure, each violin includes the SMS test samples, each one color-coded according to the participant's diagnosis, enabling a direct visual comparison of metric distributions across diagnostic groups. This representation not only shows central tendencies such as the median and interquartile range but also reveals the overall shape of the distribution, including skewness and tails, characteristics that are often hidden in simpler visualizations.

The plots also highlight the performance trade-offs that the model faces in more challenging diagnostic categories, where greater variability among individuals is frequently observed. Such visualizations are instrumental in identifying patterns of model behavior that may be diagnosis-dependent, and they help identify areas where the model may benefit from additional refinement.

Figure 3.4: Violin plots with diagnoses

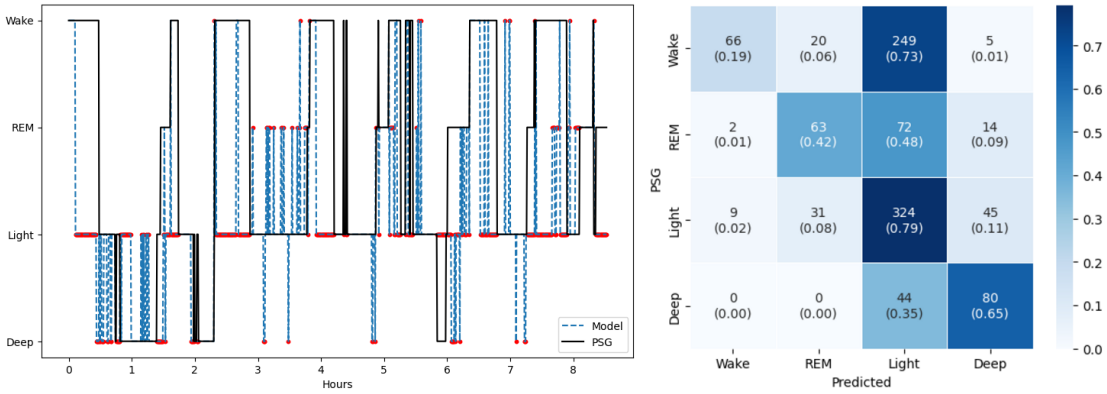


Per-class violin plot performance metrics, including test set samples, represented as dots coloured according to diagnoses. Here LS = Light Sleep, DS = Deep Sleep.

To further illustrate the model's limitations and strengths, Figure 3.5 presents a detailed example of a participant diagnosed with insomnia. The figure includes a hypnogram that compares predicted (dashed blue) and actual (black) sleep stages derived from PSG, with red dots highlighting where the model is misclassifying a stage. On the right, the figure shows a confusion matrix with both absolute and normalized values. This example clearly highlights the difficulty of the model in capturing transitions and accurately distinguishing between sleep stages in patients with fragmented or atypical sleep architecture. In this specific case, the model struggles to identify the numerous wake epochs of the insomnia participant, mostly classifying them as light sleep.

Overall, these results indicate that while the model achieves promising performance for some diagnostic categories, particularly healthy individuals and those with hypersomnolence, there remains significant room for improvement in detecting and accurately characterizing more complex and heterogeneous sleep disorders such as insomnia and parasomnias.

Figure 3.5: Hypnogram and confusion matrix from an insomnia participant



Hypnogram and confusion matrix including a comparison between PSG and model sleep stages on an insomnia test participant. On the left, the hypnogram shows predicted (dashed blue) and actual (black) sleep stages from PSG, with red dots indicating the model's misclassified epochs. On the right, a confusion matrix compares the model's predictions to PSG, with both absolute and normalized values.

3.2 SMS + DREAMT datasets

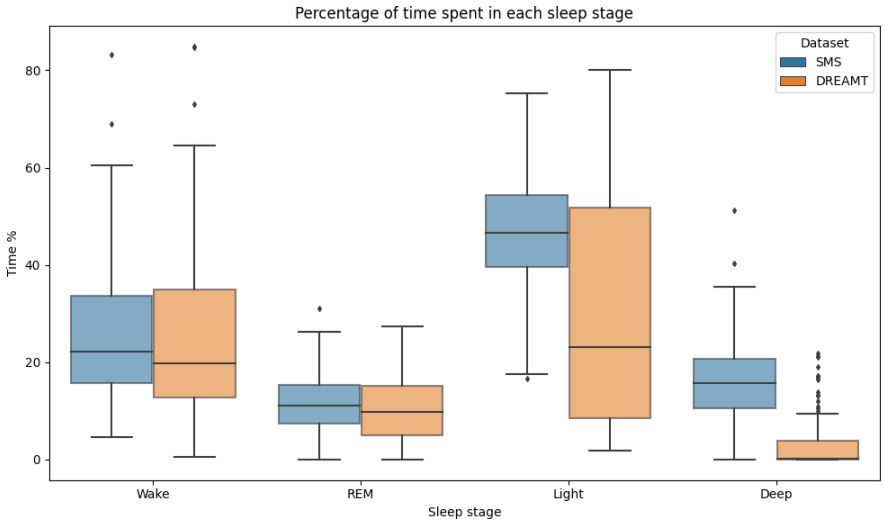
In this section, the SMS and DREAMT datasets are merged to leverage the benefits of a larger and more diverse dataset, with the goal of enhancing model performance and generalizability.

To provide insight into the characteristics of the two datasets, the distribution of sleep stages is examined and visualized using box plots, as illustrated in Figure 3.6. This analysis reveals substantial differences in the representation of sleep stages between the SMS and DREAMT datasets. Specifically, Figure 3.6 depicts the proportion of total sleep time spent in each stage—wake, REM, light, and deep sleep—aggregated across participants. Each box plot includes the median, interquartile range (IQR), and outliers, offering a clear depiction of both central tendency and variability.

The SMS dataset (blue) demonstrates relatively stable stage proportions among individuals, particularly in the light and REM sleep stages, indicating a more homogeneous distribution. In contrast, the DREAMT dataset (orange) shows a higher degree of variability across participants, most notably in the light sleep stage. Furthermore, deep sleep is frequently underrepresented in the DREAMT recordings, suggesting a potential limitation in capturing that stage correctly. Conversely, light sleep constitutes a substantial portion of sleep time in both datasets, often dominating the overall distribution.

The epoch-by-epoch analysis for the combined dataset, similar to the one proposed for the SMS dataset only, and the learning curve with increasing training size are shown in the following subsections. Both the datasets include PSG sleep stages used as ground truth, and the E4 signals, ACC, BVP, EDA, and TEMP, already used in the SMS dataset section. The results are shown using the same model, input signals combination, and parameter configurations of the previous analysis.

Figure 3.6: Box plot to compare sleep stage distribution between datasets



Box plot representing the distribution of sleep stages in the SMS (blue) and DREAMT (orange) datasets.

3.2.1 Epoch-by-epoch analysis

In Table 3.5 the epoch-by-epoch performance metrics are shown for the combined SMS + DREAMT dataset. The results show the limitation of the model in detecting REM and deep sleep, mostly in terms of Sensitivity and F1-score. This most probably influences the global metrics, specifically Balanced Accuracy, Macro-F1, and Cohen's κ, showing lower results than the model with SMS dataset only. However, the Accuracy is still higher due to the good performance in classifying wake and light stages.

Table 3.5: Performance metrics (SMS + DREAMT dataset)

Sleep stage	**Sensitivity**	**Specificity**	**Accuracy**	**F1-score**
Wake	0.66 ± 0.23	0.90 ± 0.13	0.86 ± 0.09	0.64 ± 0.15
REM	0.45 ± 0.31	0.96 ± 0.04	0.91 ± 0.04	0.43 ± 0.29
Light sleep	0.72 ± 0.15	0.66 ± 0.19	0.71 ± 0.08	0.71 ± 0.12
Deep sleep	0.41 ± 0.38	0.94 ± 0.07	0.91 ± 0.06	0.39 ± 0.34

Global	**Accuracy**	**Bal. Acc.**	**Macro-F1**	**Cohen's κ**
	0.70 ± 0.08	0.60 ± 0.13	0.55 ± 0.14	0.46 ± 0.15

Performances for single and multi-class sleep stage classification on the SMS + DREAMT test set.

3.2.2 Learning curve

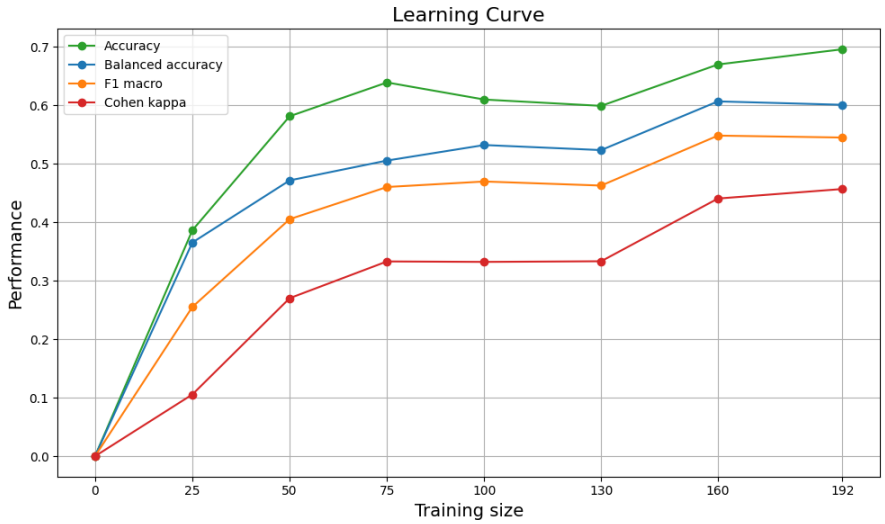
Figure 3.7 displays the learning curve for the model trained on the combined SMS and DREAMT datasets. The x-axis represents the proportion of the training set used during training, while the y-axis shows model performance in terms of four global metrics: Accuracy, Balanced Accuracy, Macro-F1, and Cohen's κ.

A learning curve illustrates how the model's performance evolves as the amount of training data increases. It provides insight into the model's learning capacity and helps assess whether the performance has stabilized or could benefit from additional data.

To generate the curve, the model was trained on progressively larger subsets of the training data, ranging from a small fraction up to the full dataset. For each subset, performance was evaluated on a fixed hold-out validation set to ensure comparability between results. This procedure allows us to observe whether the model is underfitting, overfitting, or approaching optimal generalization. As the training size increases from 0 to 192 samples, all metrics show overall improvement:

- Accuracy increases up to around 75 samples and then stabilizes around 0.65–0.7.
- Balanced Accuracy follows a similar trend, reaching values close to 0.6.
- Macro-F1 improves significantly until around 100 training samples, after which it stabilize around 0.55.
- Cohen's κ, starting from zero, increases more gradually, stabilizing around 0.45.

Figure 3.7: Learning curve SMS + DREAMT dataset



The plot shows the learning curve for the model trained on the combined SMS and DREAMT datasets. The x-axis represents the training set size, increasing from 0 to 192, while the y-axis shows model performance for Accuracy (green), Balanced Accuracy (blue), Macro-F1 (orange), and Cohen's Kappa (red).

3.3 Corsano dataset

The results are shown for a single test participant, using as input only the ACC and PPG signals available from the Corsano device. The model was pretrained on the SMS and DREAMT datasets and was not fine-tuned on Corsano data, due to the few data available, but also in order to evaluate its generalizability to unseen devices.

Table 3.6 shows the single and multi-class metrics for the test participant.

Table 3.6: Performance metrics (Corsano dataset)

Sleep stage	**Sensitivity**	**Specificity**	**Accuracy**	**F1-score**
Wake	0.82	0.98	0.96	0.79
REM	0.11	0.92	0.89	0.07
Light sleep	0.73	0.67	0.70	0.76
Deep sleep	0.59	0.87	0.80	0.59
Global	**Accuracy**	**Bal. Acc.**	**Macro-F1**	**Cohen's κ**
	0.68	0.56	0.55	0.43

Performance metrics for single-class and multi-class four-sleep stage classification on the test set of the Corsano dataset. All values are reported as point estimates (standard deviations not available).

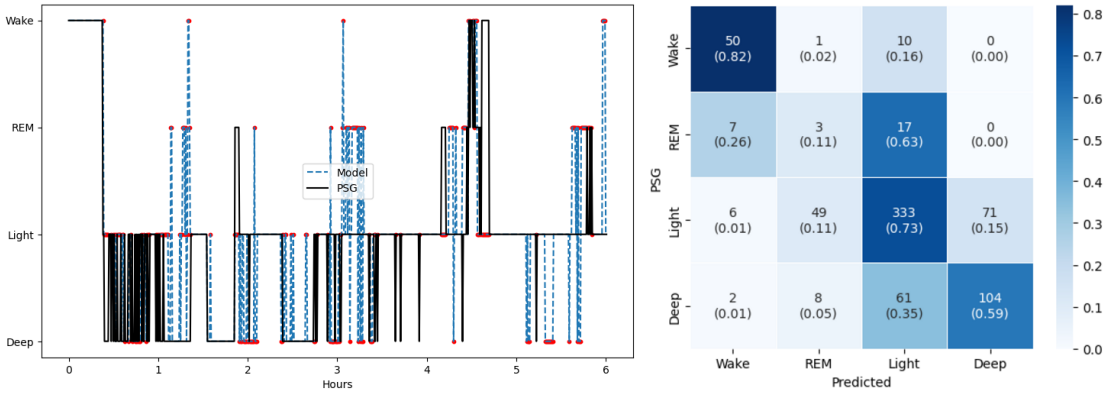
The model demonstrated varying performance across the different sleep stages. High sensitivity (82%) and specificity (98%) were observed for Wake, indicating strong detection ability in this class. Light sleep showed balanced performance, with sensitivity and specificity values of 73% and 67%, respectively, and the highest F1-score among all stages (76%).

Deep sleep achieved moderate results, with a sensitivity of 59% and specificity of 87%. In contrast, REM sleep detection was notably limited, with a low sensitivity (11%) and F1-score (7%), despite relatively high specificity (92%).

Global metrics reflect the imbalanced per-class performance. The overall accuracy was 68%, while balanced accuracy and macro-F1 were lower (56% and 55%, respectively), suggesting variable model effectiveness between classes. Cohen's κ of 43% indicates moderate agreement with the reference labels.

A visual results analysis is shown in Figure 3.8, including the hypnogram and confusion matrix comparing the model performance to the PSG. While the model shows reasonable performance overall, mostly in detecting Wake and Light epochs, it is evident that it struggles to accurately identify REM sleep, likely due to the low number of REM epochs in the test data relative to other stages.

Figure 3.8: Hypnogram and confusion matrix from Corsano



Hypnogram and confusion matrix including a comparison between PSG and model sleep stages on an insomnia test participant. On the left, the hypnogram shows predicted (dashed blue) and actual (black) sleep stages from PSG, with red dots indicating the model's misclassified epochs. On the right, a confusion matrix compares the model's predictions to psg, with both absolute and normalized values.

3.4 Comparison to literature

A comparison of the performance of the proposed model on all three datasets with previous studies on four-class sleep stage classification using wearable signals is presented in Table 3.7. The studies included in this literature review were selected based on the following criteria:

- Results for four-sleep stage classification, including Wake, REM, Light (N1+N2), and Deep (N3) sleep, compared to with PSG annotations;
- Use of signals acquired exclusively from wrist-worn wearable devices;
- Reporting of at least a subset of the global evaluation metrics adopted in this work, to allow for models comparison;
- SOTA performance reported for the corresponding experimental settings.

The table reports, for each work, the model architecture, the input signals used, the number of participants, any relevant diagnoses, and the evaluation metrics available in the respective studies, including accuracy, balanced accuracy, macro F1-score, and Cohen's kappa.

Olsen et al. [14] employed U-Net using as input ACC and PPG. A larger dataset of 231 participants, made up of mostly Breathing Disorders (BD) subjects, included wrist-worn tri-axial ACC at 25 Hz and PSG-derived PPG at 100 Hz. The hold-out set was composed of 35 healthy participants with both ACC and PPG derived from another wearable device, at 25 and 50 Hz, respectively. Another dataset of 35 participants with ACC and PPG both recorded at 25 Hz was also included, reaching a total of 301 participants. The reported metrics include an accuracy of 69% and Cohen's κ of 58%.

Li et al. [15] employed a transfer learning approach, adapting a model which combined a CNN with a Support Vector Machine (SVM) classifier pretrained on ECG signals. The dataset

Table 3.7: Comparison to related works

	Model	Signals	Participants	Diagnosis	Accuracy / BA / F1 / κ
SMS	U-Net	ACC + BVP + EDA + TEMP	127	Sleep disorders	69 / 64 / 62 / 50
SMS + DREAMT	U-Net	ACC + BVP + EDA + TEMP	227	Sleep disorders	70 / 60 / 55 / 46
Corsano	U-Net	ACC + PPG	1	Sleep disorders	68 / 56 / 55 / 43
Olsen et al. [14]	U-Net	ACC + PPG	301	Healthy/BD	69 / - / - / 58
Li et al. [15]	CNN + SVM	ACC + PPG	105	Healthy/PTSD	69 / - / - / 44
Silva et al. [24]	Features RNN	ACC + PPG	1522	Healthy/SA	71 / 72 / - / 56
Song et al. [25]	SLAMSS	ACT + HRM + HRSD	808	Healthy	72 / - / 73 / -
Zhai et al. [26]	LSTM	ACT + HR	1743	Healthy*	70 / - / 52 / 54

Comparison of recent sleep stage classification works to the proposed model. The overview includes the model used, the input signals, the number of participants, the diagnosis, and the global metrics for all three datasets of this study, and for some SOTA works. *16% with sleep disorders.

includes 105 participants, some of whom with Post-Traumatic Stress Disorder (PTSD). The model with ACC and PPG collected from the E4 achieves an accuracy of 69% and Cohen's κ of 44%.

Silva et al. [24] proposed a Recurrent Neural Network (RNN) algorithm based on hand-crafted features, using ACC and PPG signals from commercial smartwatches, evaluated on a population of 1522 participants, including both healthy subjects and a minor percentage of individuals with diagnosed Sleep Apnea (SA). The results include accuracy (71%), balanced accuracy (72%), and Cohen's κ (56%).

Song et al. [25] used a Long Short-Term Memory (LSTM) model called Sequence-to-sequence LSTM for Automated Mobile Sleep Staging (SLAMSS) on Actigraphy (ACT), Heart Rate Mean (HRM), and Heart Rate Standard Deviation (HRSD). The data, derived from an actiwatch, includes ACT, HRM, and HRSD from 808 healthy participants. The model results in 72% of accuracy and 73% of F1-score.

Finally, Zhai et al. [26] designed a multimodal LSTM sleep staging model using both motion (from ACT) and cardiac features (from PPG-derived HR), trained and evaluated on a large-scale dataset of 1743 participants. The data included mainly healthy individuals and 16% suffering from sleep disorders. The model achieved an accuracy of 70%, a macro F1-score of 52%, and a Cohen's κ of 54%.

Chapter 4

Discussion

This study, which involves deep learning for four-sleep stages classification, targets heterogeneous datasets comprising participants with a range of diagnosed sleep disorders and markedly disrupted sleep. This population is both clinically relevant and presents great challenges for modeling. Furthermore, the datasets used in this work include up to 227 participants, which increases classification difficulty due to the limited amount of training data and high inter-individual variability. Certain diagnostic categories are also underrepresented, posing an additional challenge to the model's generalizability across diverse sleep disorders.

In contrast, as shown in Table 3.7, most existing studies involve mainly healthy participants or include only a small proportion of individuals with sleep disorders. Moreover, these previous works use as input signals those with higher resolution, in some cases PSG-derived, or large datasets, contrary to the proposed approach, which considers only wearable-derived signals in a realistic clinical population.

Despite these constraints, the proposed approach achieves performance comparable to that reported in the literature, underscoring the robustness of the multimodal model used, which integrates ACC, BVP, EDA, and TEMP signals. These findings highlight the potential of this method for application in real-world, diagnostically diverse sleep populations.

The following sections provide a more detailed discussion of each objective introduced at the beginning of this study, including the integration of EDA and TEMP, results per diagnosis, the augmentation of the dataset, and finally, the cross-device generalization.

4.1 Integration of EDA and TEMP

The proposed model tested on the SMS dataset with ACC, BVP, EDA, and TEMP for four-sleep stages classification achieves an overall accuracy of 69%, balanced accuracy of 64%, F1-score of 62%, and κ of 50% on the test set. As shown in Figure 2.1, both EDA and TEMP exhibit correlations with sleep stages; however, preliminary experiments integrating these signals into the model resulted in worse classification performance. This observation motivated the development of alternative preprocessing steps specifically adapted for EDA and TEMP, enhancing their contribution to the overall model accuracy.

To provide a comprehensive interpretation of the results, the following subsections address

the sleep measure and epoch-by-epoch concordance analysis.

4.1.1 Sleep measures analysis

In addition to the performance for sleep stage classification, the agreement between sleep measures derived by the proposed model and the ground truth values obtained from PSG was also evaluated. As shown in Table 3.2, and in the visual analysis provided by the Bland-Altman plots in Figure 3.2, no statistically significant differences were found for most metrics, with the exception of SOL and the duration of light sleep (Light_d). These results suggest that the model is generally capable of providing reliable estimates of key sleep architecture parameters.

This difference in SOL can be linked to a known limitation of wrist-worn wearables: their inability to reliably detect wakefulness in the absence of movement, especially near sleep onset. This is a well-documented limitation in wearable sleep monitoring, as actigraphy-based systems often struggle to distinguish low-movement wakefulness from actual sleep. Consequently, the model may underestimate SOL, mistaking motionless wake periods for light sleep.

Similarly, the overestimation of light sleep duration may be attributed to the model's tendency to misclassify wake or REM epochs as light sleep. This trend is in line with the class distribution imbalance commonly found in sleep datasets, where light sleep represents the majority class. As a result, the model may be biased to overpredict this stage, particularly in ambiguous or transitional epochs.

Together, these findings highlight specific areas where wearable-based models still face challenges, particularly in the accurate detection of sleep transitions and minority sleep stages. Despite this, the overall alignment of estimated sleep measures with PSG reinforces the potential utility of the proposed approach for large-scale or at-home sleep monitoring in heterogeneous populations.

4.1.2 Epoch-by-epoch analysis

The detailed epoch-by-epoch evaluation presented in Table 3.3 reveals both the abilities and limitations of the model in classifying sleep stages. Notably, the model demonstrates high specificity across most classes, particularly for wake, REM, and deep sleep. This reflects a strength in avoiding false positives, especially for stages that are often confused with others in wearable-based systems.

In contrast, sensitivity values are more differentiated between stages. The model performs worst in detecting REM (0.55 ± 0.25) and Wake (0.59 ± 0.20), highlighting a tendency to miss true instances of these stages. This suggests challenges in capturing the physiological signals associated with these sleep stages, which may be explained by their lower occurrence and their high transitional nature. Consequently, the F1-score for REM sleep is the lowest among the stages (0.53 ± 0.22), reflecting both the limited recall and the complexity of accurately identifying this stage based on wrist-worn data. Light sleep, on the other hand, achieves the highest sensitivity (0.76 ± 0.11) and a relatively stable F1-score (0.72 ± 0.09), with a trade-off in specificity (0.66 ± 0.12), indicating that the model tends to assign many epochs to this dominant stage.

Global classification metrics (Accuracy: 0.69 ± 0.08; Balanced Accuracy: 0.64 ± 0.10; Macro F1-score: 0.62 ± 0.11) reflect moderate concordance with the gold-standard PSG, as

also supported by a Cohen's κ of 0.50 ± 0.14. These results are encouraging, especially given the complexity and diagnostic heterogeneity of the dataset, and are in line with or superior to other prior works using wearables in mostly healthy subjects.

In addition, the results presented in Table 3.1 and the visual analysis given by the confusion matrices in Figure 3.1 illustrate the incremental benefit of incorporating multiple signal modalities. Starting from a single-sensor setup using only ACC, the model shows limited ability to identify deep and REM sleep stages. The inclusion of BVP provides a noticeable performance improvement, particularly for REM classification. The highest overall accuracy is reached when all available modalities (ACC, BVP, EDA, and TEMP) are combined, demonstrating improved detection of each class. These findings highlight the value of a multimodal strategy in enhancing sleep stage classification, especially for capturing less frequent and physiologically complex stages like REM sleep.

4.2 Diagnosis-specific performance analysis

Most validation studies of wearable-based sleep monitoring have been conducted on healthy participants, whose sleep patterns tend to be continuous, stable, and undisturbed. Consequently, these studies may not properly reflect the performance of wearable devices in clinical settings where sleep is often fragmented, irregular, and affected by participants or context-specific conditions. This work addresses this limitation by including a clinically heterogeneous population and by stratifying the evaluation of model performance according to specific sleep disorder diagnoses. This approach enables a more comprehensive understanding of how wearable-based sleep staging models generalize across various pathological profiles.

As shown in Table 3.4, the highest classification performance was observed in healthy control participants, with consistently strong metrics, reaching an Accuracy of 77%, Balanced Accuracy of 73%, Macro-F1 of 74%, and Cohen's κ of 64%. This result is in line with the literature, given that healthy participants typically exhibit more structured and easily distinguishable sleep patterns, which are easier for the model to learn and classify accurately.

Performance was more heterogeneous across clinical subgroups. Participants diagnosed with hypersomnolence and breathing disorders achieved relatively high scores, suggesting that despite their underlying conditions, these groups maintain discernible physiological patterns that the model can capture effectively. In contrast, the model performed least effectively for individuals with insomnia and parasomnias, both of which produced the lowest metrics across the diagnoses. In particular, Balanced Accuracy values of 51% for insomnia and 49% for parasomnias indicate significant challenges in accurately identifying sleep stages in these groups. These difficulties are likely caused by highly irregular sleep architecture and increased intra-subject variability that makes classification more challenging.

This issue is further illustrated in the example shown in Figure 3.5, where the model misclassifies extended wake periods as light sleep in an insomnia case. This misclassification likely arises because the participant remains motionless in bed with reduced heart rate, physiological conditions that resemble light sleep despite actual wakefulness, highlighting a common limitation of wearable devices that rely heavily on movement and heart rate signals.

Similarly, classification performance for individuals with movement disorders was relatively weak, with a Balanced Accuracy of 52%. This may reflect the heterogeneity of this diagnostic

group and the complex nature of their sleep disorders, which may deviate significantly from patterns the model was trained on.

To improve diagnostic sensitivity, especially in underrepresented or pathologically complex groups, future work should prioritize the development of larger and more balanced datasets, ideally including sufficient representation from each disorder type. However, the results obtained in this study demonstrate that the inclusion of multiple physiological signals, particularly those with strong associations to sleep patterns, can yield performance comparable to SOTA models [24], even under challenging clinical conditions.

4.3 Augmented dataset

To increase the diversity and size of the training data, two existing datasets, SMS and DREAMT, were combined. These datasets share a common set of physiological signals collected using the E4 device (ACC, BVP, EDA, and TEMP) together with gold-standard PSG annotations. By merging these datasets, the aim was to create a more heterogeneous sample set with a wider range of sleep behaviors and disorders, improving the model's generalizability to real-world applications.

However, as shown in Figure 3.6, the two datasets have different class distributions, increasing the inter-subject variability. The most challenging characteristic of the additional DREAMT dataset is that 49% of the participants have no deep sleep, 69% of whom suffer from OSA, which can explain this unusual sleep pattern. For this reason, the results are not very promising, with a slight improvement only in accuracy with respect to the model trained on the SMS dataset. A more detailed analysis of the results with the combined dataset is shown below.

4.3.1 Epoch-by-epoch analysis

As shown in Table 3.5, overall classification metrics, including relatively high values of Accuracy (70%), indicate that the model generalizes well across the dataset. However, the consistent gap between the other metrics (Balanced Accuracy, Macro-F1, and Cohen's κ) highlights the presence of class imbalance, likely driven by the overrepresentation of light sleep and the relative underrepresentation of REM and deep stages.

This imbalance is further illustrated in Figure 3.6, which shows the distribution of sleep stages across the two datasets. Notably, participants from the DREAMT dataset, many of whom suffer from OSA, exhibit a marked absence or reduction of deep sleep. This pathological sleep architecture significantly alters the class distribution in the combined dataset and likely impacts the model's ability to learn representative features for the deep sleep class. As a result, the model may be biased toward stages that are more consistently represented across both datasets, such as light sleep, while underperforming on those that are rare or dataset-specific.

Additionally, the relatively lower Macro-F1 and Cohen's κ scores reinforce this issue, as these metrics are particularly sensitive to the model's ability to detect less frequent classes. Poor detection of REM and deep sleep epochs, both of which are commonly affected in clinical populations, reduces agreement beyond chance and signals limited generalization for these clinically relevant states. This is particularly concerning in diagnostic contexts, where accurate

staging of REM and deep sleep plays a critical role in evaluating sleep quality and identifying specific disorders.

4.3.2 Learning curve

The learning curve presented in Figure 3.7 confirms that model performance improves as more training samples are introduced, with the most substantial gains occurring up to approximately 100 instances. Beyond this threshold, performance stabilizes, suggesting that the model is nearing its capacity to extract useful patterns given the current feature set and architecture. This leveling effect is typical when the model becomes saturated with the available information, indicating no further improvement from simply increasing data volume without modifying the learning approach.

In sum, while dataset augmentation improves generalization, the underlying class imbalance, partly due to the physiological effects of OSA in the DREAMT dataset, remains a limiting factor for model robustness. Addressing this through targeted data augmentation, stratified sampling, or class-aware loss functions may be necessary to enhance performance, especially for clinically underrepresented sleep stages.

To address these limitations, future work could explore various strategies, including:

- Class balancing methods, such as resampling or cost-sensitive learning, to mitigate skewed class distributions;
- Feature augmentation, including temporal context windows or signal-derived frequency features;
- Advanced architectures, such as RNNs or LLMs, which might better capture the temporal dynamics of sleep;
- Transfer learning approaches, leveraging pre-trained models on larger wearable datasets.

By combining datasets and expanding the training size, this study represents a step toward more robust and clinically applicable wearable sleep stage classification. However, further improvements in both data diversity and model complexity are likely necessary to reach performance levels suitable for diagnostic use in more challenging populations.

4.4 Cross-device generalization

To assess the generalizability of the proposed model to different wearable devices, a preliminary evaluation was conducted using data from a participant wearing the Corsano device, used as test set. The model had been pretrained exclusively on the combined SMS and DREAMT datasets, which use the E4 wearable. Despite this, the model was able to process and classify Corsano data without additional fine-tuning, demonstrating encouraging results.

As shown in Table 3.6, the classification performance on Corsano data was comparable to the results obtained from the E4-based datasets, achieving an Accuracy of 68%, Balanced Accuracy of 56%, Macro-F1 of 55%, and Cohen's κ of 43%. These values suggest that the

model can generalize reasonably well to data acquired from a different device, provided that the signal modalities and preprocessing steps remain consistent.

The model shows the worst per-class performance on REM sleep, a very common result in this context since it is the rarest class, as evident in Figure 3.8.

It is important to note that this evaluation is based on a single subject from an ongoing and still-growing Corsano dataset. For this reason, the results should be considered preliminary and interpreted with caution. However, they offer promising evidence that the proposed multimodal architecture is robust enough to operate with different wrist-worn technologies, opening the door to wider applicability.

In future work, the model can be applied to the full Corsano dataset as it becomes available. This will allow for a more systematic validation of cross-device performance and may inform further adaptations or fine-tuning strategies to improve the model performance across different wearable devices.

Chapter 5

Conclusions

This work showed the applicability and clinical relevance of deep learning techniques for sleep stage classification using wrist-worn sensors in a heterogeneous population with diagnosed sleep disorders. In contrast to many previous studies limited to healthy subjects with consolidated sleep, this model effectively handles the complexity and variability due to pathological sleep patterns. The achieved performance, comparable to SOTA wearable-based methods, highlights the potential of multimodal physiological signals from consumer-grade devices as scalable and non-invasive alternatives to the gold standard PSG.

Importantly, by combining datasets from different sources and evaluating the model on a clinically diverse population, the generalizability and robustness of wearable sleep staging in real-world settings were demonstrated. The preliminary cross-device testing on data collected from the Corsano wearable further suggests that the developed approach can adapt to different devices, supporting its applicability in different clinical and research contexts.

This study underscores the importance of personalized Artificial Intelligence (AI) models that not only leverage multimodal sensor data but also incorporate diagnostic information and population heterogeneity to achieve personalized sleep monitoring. Such approaches pave the way toward patient-centered care in sleep medicine, enabling continuous, at-home sleep assessment and more personalized therapeutic interventions.

5.1 Limitations and Future Work

Despite these promising results, several limitations must be acknowledged. This study was conducted in a clinical environment where manual PSG scoring, although performed by experts, may suffer from inter-scorer variability and less standardized protocols due to practical constraints. The relatively small sample size and the heterogeneous representation of diagnostic categories limit the model's ability to generalize uniformly, especially for less frequent or more complex disorders such as insomnia and parasomnias.

Furthermore, the current U-Net architecture processes fixed-length input segments, which may truncate longer sleep recordings. Extending the model to accept variable-length inputs or leveraging architectures designed for sequential data, such as recurrent or attention-based networks, could enhance temporal modeling of sleep dynamics.

Incorporating diagnosis labels into the training process through weighted loss functions or multitask learning frameworks could improve sensitivity for underrepresented classes and enable the model to adapt to specific pathological profiles. Addressing class imbalance via oversampling, stratified batch sampling, or synthetic data generation will likely enhance robustness across diverse patient populations.

Finally, while initial cross-device generalization results are encouraging, comprehensive validation on larger, heterogeneous datasets from multiple wearable platforms is necessary. Future work will focus on fine-tuning the model on such datasets, improving preprocessing standardization, and exploring transfer learning techniques to further boost cross-device compatibility.

In conclusion, this study lays a strong foundation for the development of clinically applicable, wearable-based sleep staging tools. Continued efforts to integrate richer clinical data, improve model architectures, and validate in different populations will be critical steps toward guaranteeing reliable, personalized sleep health monitoring outside of laboratory settings.

Bibliography

[1] Neil S. Zheng, Jeffrey Annis, Hiral Master, Lide Han, Karla Gleichauf, Jack H. Ching, Melody Nasser, Peyton Coleman, Stacy Desine, Douglas M. Ruderfer, John Hernandez, Logan D. Schneider, and Evan L. Brittain. Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the all of us research program. Nat Med, 30:2648–2656, 2024. doi: 10.1038/s41591-024-03155-8. URL https://doi.org/10.1038/s41591-024-03155-8.

[2] Gawon Cho, Adam P. Mecca, Orfeu M. Buxton, Xiao Liu, and Brienne Miner. Lower slow wave sleep and rapid eye-movement sleep are associated with brain atrophy of advulnerable regions. bioRxiv, 2025. doi: 10.1101/2025.01.12.632386. URL https://www.biorxiv.org/content/early/2025/01/14/2025.01.12.632386.

[3] Sami Nikkonen, Pranavan Somaskandhan, Henri Korkalainen, Samu Kainulainen, Philip I. Terrill, Heidur Gretarsdottir, Sigridur Sigurdardottir, Kristin Anna Olafsdottir, Anna Sigridur Islind, María Óskarsdóttir, Erna Sif Arnardóttir, and Timo Leppänen. Multicentre sleep-stage scoring agreement in the sleep revolution project. Journal of Sleep Research, 33(1):e13956, 2024. doi: https://doi.org/10.1111/jsr.13956. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.13956.

[4] Oriella Gnarra, Marie-Angela Wulf, Carolin Schäfer, Tobias Nef, and Claudio L A Bassetti. Rapid eye movement sleep behavior disorder: a narrative review from a technological perspective. Sleep, 46(6):zsad030, 02 2023. ISSN 0161-8105. doi: 10.1093/sleep/zsad030. URL https://doi.org/10.1093/sleep/zsad030.

[5] MedTechNews. Polysomnography: Advancements, challenges, and emerging applications in sleep medicine and neuroscience, 2023. URL https://medtechnews.uk/research-reports/polysomnography-advancements-challenges-and-emerging-\applications-in-sleep-medicine-and-neuroscience/.

[6] Oriella Gnarra, Carmen Calvello, Tommaso Schirinzi, Francesca Beozzo, Claudia De Masi, Matteo Spanetta, Mariana Fernandes, Piergiorgio Grillo, Rocco Cerroni, Mariangela Pierantozzi, Claudio L. A. Bassetti, Nicola Biagio Mercuri, Alessandro Stefani, and Claudio Liguori. Exploring the association linking head position and sleep architecture to motor impairment in parkinson's disease: An exploratory study. Journal of Personalized Medicine, 13(11), 2023. ISSN 2075-4426. doi: 10.3390/jpm13111591. URL https://www.mdpi.com/2075-4426/13/11/1591.

[7] Taeyoung Lee, Younghoon Cho, Kwang Su Cha, Jinhwan Jung, Jungim Cho, Hyunggug Kim, Daewoo Kim, Joonki Hong, Dongheon Lee, Moonsik Keum, Clete A Kushida, In-Young Yoon, and Jeong-Whun Kim. Accuracy of 11 wearable, nearable, and airable consumer sleep trackers: Prospective multicenter validation study. JMIR mHealth and uHealth, 2023. doi: 10.2196/50983. URL https://doi.org/10.2196/50983.

[8] Massimiliano De Zambotti, Nicola Cellini, Aimée Goldstone, Ian M. Colrain, and Fiona C. Baker. Wearable sleep technology in clinical and research settings. Medicine & Science in Sports & Exercise, 2019. doi: 10.1249/MSS.0000000000001947. URL https://doi.org/10.1249/MSS.0000000000001947.

[9] Oriella Gnarra, Alexander Breuss, Lorenzo Rossi, Manuel Fujs, Samuel E.J. Knobel, Jan D. Warncke, Stephan M. Gerber, Claudio L.A. Bassetti, Robert Riener, Tobias Nef, and Markus H. Schmidt. Transferability of a sensing mattress for posture classification from research into clinics. In 2023 International Conference on Rehabilitation Robotics (ICORR), pages 1–6, 2023. doi: 10.1109/ICORR58425.2023.10304684.

[10] Ke Wang, Jiamu Yang, Ayush Shetty, and Jessilyn Dunn. DREAMT: Dataset for real-time sleep stage estimation using multisensor wearable technology. PhysioNet, version 2.1.0, April 2025. RRID:SCR_007345.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 2015.

[12] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, Alexander Dovzhenko, Olaf Tietz, Cristina Dal Bosco, Sean Walsh, Deniz Saltukoglu, Tuan Leng Tay, Marco Prinz, Klaus Palme, Matias Simons, Ilka Diester, Thomas Brox, and Olaf Ronneberger. U-net: deep learning for cell counting, detection, and morphometry. Nature Methods, 16(1):67–70, January 2019. ISSN 1548-7105. doi: 10.1038/s41592-018-0261-2. URL https://doi.org/10.1038/s41592-018-0261-2.

[13] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-sleep: resilient high-frequency sleep staging. npj Digital Medicine, 4 (1):72, April 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00440-5. URL https://doi.org/10.1038/s41746-021-00440-5.

[14] 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.

[15] 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 2021. ISSN 13616579. doi: 10.1088/1361-6579/abf1b0. URL https://pubmed.ncbi.nlm.nih.gov/33761477/.

[16] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data. 2024. doi: 10.48550/arXiv.2401.06866. URL https://arxiv.org/abs/2401.06866.

[17] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxeda: A convex optimization approach to electrodermal activity processing. IEEE Transactions on Biomedical Engineering, 2016:797–804, 04 2016. doi: 10.1109/TBME.2015.2474131. URL https://doi.org/10.1109/TBME.2015.2474131.

[18] Hugo F. Posada-Quintero and Ki H. Chon. Phasic component of electrodermal activity is more correlated to brain activity than tonic component. In 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pages 1–4, 2019. doi: 10.1109/BHI.2019.8834567.

[19] Andrea Di Credico, David Perpetuini, Pascal Izzicupo, Giulia Gaggi, Nicola Mammarella, Alberto Di Domenico, Rocco Palumbo, Pasquale La Malva, Daniela Cardone, Arcangelo Merla, Barbara Ghinassi, and Angela Di Baldassarre. Predicting sleep quality through biofeedback: A machine learning approach using heart rate variability and skin temperature. Clocks&Sleep, 6:322–337, 2024. doi: 10.3390/clockssleep6030023. URL https://doi.org/10.3390/clockssleep6030023.

[20] Peretz Lavie, Richard P Schnall, Jonathan Sheffy, and Amichai Shlitner. Peripheral vasoconstriction during rem sleep detected by a new plethysmographic method. Nature Medicine, 6(6):606, June 2000. doi: 10.1038/76135.

[21] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. Journal of Machine Learning Research, 2018. doi: 10.48550/arXiv.1603.06560. URL https://doi.org/10.48550/arXiv.1603.06560.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[23] Luca Menghini, Nicola Cellini, Aimee Goldstone, Fiona C. Baker, and Massimiliano de Zambotti. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. Sleep Research Society, 2020, 2020. doi: 10.1093/sleep/zsaa170. URL https://doi.org/10.1093/sleep/zsaa170.

[24] 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, 2023. doi: 10.1016/j.sleep.2024.05.033. URL https://doi.org/10.1016/j.sleep.2024.05.033.

[25] Tzu-An Song, Samadrita Roy Chowdhury, Masoud Malekzadeh, Stephanie Harrison, Terri Blackwell Hoge, Susan Redline, Katie L. Stone, Richa Saxena, Shaun M. Purcell,

and Joyita Dutta. Ai-driven sleep staging from actigraphy and heart rate. PLOS ONE, 18(5):1–29, 05 2023. doi: 10.1371/journal.pone.0285703. URL https://doi.org/10.1371/journal.pone.0285703.

[26] Bing Zhai, Ignacio Perez-Pozuelo, Emma Clifton, João Palotti, and Yu Guan. Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4:1–33, 06 2020. doi: 10.1145/3397325.