Towards Automated Sleep Stage Assessment Using Ubiquitous Computing Technologies

Bing Zhai

School of Computing
Newcastle University

This dissertation is submitted for the degree of
Doctor of Philosophy

Aug 2022

Abstract

The growing popularity of ubiquitous computing devices, such as smartphones, wristbands and smartwatches, has caused an increase in the scale of collecting physiological and psychological data on their growing number of users. This availability of digital health data outside the normal confines of hospitals and other institutions provides a fundamental opportunity for researchers to infer individual behaviour and health at the scale. An application of ubiquitous computing for digital health is the development of automated systems that are robust enough to monitor sleep stages noninvasively outside the sleep laboratory. However, turning the data into actionable insights requires computational methods that can infer sleep stages from physiological time-series data related to parts of the brain's activity.

This thesis describes the novel deep-learning methods that leverage wearable sensing data for non-invasive sleep stage monitoring in large-scale populations. Firstly, the study performs a systematic evaluation of the sleep stage classification based on traditional machine learning models and neural networks using actigraphy and cardiac sensing data. The proposed deep ensemble model outperforms traditional algorithms and the deep learning baselines. However, the performance of the automated sleep stage monitoring algorithm can be affected by personal attributes such as age, BMI and sleep disorders, etc. Therefore, this work proposes a novel network based on the variational autoencoder, which can disentangle the feature space into personal attribute-specific features that are irrelevant to sleep stage classification and personal attribute-free features that only contain the sleep stage-relevant information. The proposed network can effectively reduce the effects of personal attributes on model performance. Finally, multimodal fusion strategies and methods are systematically investigated. The proposed fusion methods can significantly improve the performance of three-stage sleep classification on a large clinical sleep study dataset. The proposed methods have also experimented on a small sleep dataset collected from consumer-grade wearables. The empirical results demonstrate that wearable sensors can classify three stages of sleep with 78 % accuracy. These proposed methods generate robust predictions and may be used for long-term free-living sleep stage monitoring.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Bing Zhai
Aug 2022

Acknowledgements

I am very grateful to Dr Yu Guan for guiding me through this PhD, as well as for his insightful comments and support, especially for giving me the proper freedom and constraints to focus on digital health research. I would also like to thank Duan Haoran, who became my best research partner. April 2019-January 2021 was the most intellectually stimulating period I enjoyed in my life. I am privileged to embark on a journey that is consistent with my pursuit of research excellence. I deeply appreciate Ignacio Perez-Pozuelo, João Palotti, Luis Fernandez Luque, Gewei Zhu, Professor Mike Catt, and Professor Patrick Olivier for helping me shape my research. I would also like to thank my collaborators in Cambridge, The Alan Turning Institute, MIT, QCRI, the University of Plymouth and Weill Cornell for their support during my PhD. I would like to thank my committee members Huizhi Liang, and Huy Phan for the fruitful discussions and comments on my thesis and during my defence.

I am very grateful to Professor Thomas Plötz for his guidance and support, who introduced me to the challenges of machine learning in human activity recognition during my master's studies. This motivates me to pursue applied machine learning for health as my doctoral research direction. I would also like to thank my previous employers, James Jones & Son Ltd. who provided me with tremendous financial support during my early research phases.

Finally, the foundation of my success is my family's sacrifice and love. My parents put my dreams ahead of their needs, even though my choices were beyond their comprehension. I would also like to dedicate this thesis to my uncle Martin Li and aunt Ping Zhai who motivate me to start my research journey in the UK. I'm also grateful to my wife Lily Liu who has supported me through the good and the bad for many years now. I will not be who I am today without my best friends and my family in life.

Contents

List of Figures

List of Tables

Abbreviations

AASM American Academy of Sleep Medicine
AHI Apnoea Hypopnoea Index
ANS Autonomic Nervous System

BCG Ballistocardiography
BMI Body Mass Index
BPTT Back-Propagation Through Time)

CNN convolutional Neural Network

ECG Electrocardiogram
EEG Electroencephalogram
EMG Electromyography
EOG Electrooculogram

GAN Generative Adversarial Network
Grad-CAM Gradient-weighted Class Activation Mapping
GWAS Genome-Wide Association Study

HAR Human Activity Recognition
HF High Frequency
HR Heart Rate
HRV Heart Rate Variability

LF Low Frequency
LSTM Long Short-Term Memory

NREM Non-Rapid Eye Movement

OECD The Organisation for Economic Co-operation and Development
OSA Obstructive Sleep Apnoea

Personal Attribute

PPG Plethysmography
PSG Polysomnography

ReLU Rectified Linear Unit
REM Rapid Eye Movement
RF Radio Frequency
RNN Recurrent Neural Network
RRI R-wave to R-wave Interval

SHAP SHapley Additive exPlannations
SM Saliency Mask
SpO2 Saturation of Peripheral Oxygen
SWS Slow Wave Sleep

t-SNE t-Distributed Stochastic Neighbour Embedding

VAE Variational Autoencoder
VSG Videosomnography

WASO Wake After Sleep Onset

Chapter 1. Introduction

The latest development in Ubiquitous technology, such as wearable devices and mobile computing, along with new types of sensors, provides a new platform for long-term sleep and health monitoring. Consumer-grade and research-grade electronics such as wearable, nearable and mobile devices are now capable of supporting objective, inexpensive human activity monitoring, and can also provide clinically relevant data for scalable behaviour research on a large scale. For example, the UKBiobank accelerometer study included more than 85,670 subjects, each subject wearing an actigraph (a device equipped with accelerometers used for measuring body movement) over a recording period of 7 days [7]. Through analysis of UKBiobank data, a previous study identified 26 genetic associations related to sleep quality measurements and 10 genetic associations related to night sleep duration [8]. Another longitudinal sleep study involving more than 120,000 individuals with up to 2 years of monitored sleep duration data concluded that shorter sleep duration (hours slept at night) and greater day-to-day variability of sleep duration (standard deviation of hours slept at night) are positively associated with body mass index (BMI) [9]. These studies demonstrate that long-term sleep monitoring using wearable sensors is essential to understanding the relationship between behaviour and health, which will ultimately benefit public health research.

Physiological signal data extracted from commercial wearable devices are now used to develop sleep/wake and sleep stage classification algorithms to overcome limitations of questionnaires and laboratory measurements, with the goal of linking the sleep data to clinical observations. The studies based on wearable sensing data will eventually accelerate the advancement of preventive and predictive medicine. This chapter will first describe the implications of monitoring sleep in longitudinal free-living scenarios. Secondly, the state-of-the-art methods for sleep sensing research using ubiquitous computing techniques will be discussed, and the discussion will also illustrate how machine learning and deep learning techniques can be used for sleep monitoring outside the laboratory. Finally, this chapter will discuss the outstanding challenges in human sleep monitoring using wearable sensor data and illustrate the contributions of the original research presented in this thesis in addressing these challenges.

1.1. Sleep, Health and Society

Sleep is a reversible physiological state that is essential for life, health, and performance. The functions of sleep are not yet fully understood. It is well known that it can restore energy, promote healing, rejuvenate the body system, interact with the immune system, consolidate memory and maintain mental health [10–13]. As a result of its importance to vital human

processes and the incomplete understanding of its function, accurate sleep monitoring is of interest to the understanding of human health and is becoming an active area of research for the *ubiquitous computing* community [14–17]. Insufficient sleep or poor quality of sleep can have a negative impact on our judgment and cognitive performance. Long-term sleep deprivation or irregular sleep pattern are associated with the development of diseases [18, 19]. Therefore, the study of sleep characteristics represents a public health priority [20–22]. In addition, diet and physical activity have been shown to correlate with sleep quality [23], and they are interrelated and affect each other. In the past few decades, public health organisations in many countries have promoted healthy eating and regular exercise. Recently, poor sleep has also attracted public health attention [24].

So far, mounting evidence has shown that human sleep is regulated by circadian rhythm, homeostatic and behavioural factors [25, 26]. Regarding the behavioural factors, poor sleep quality [27] is often related to stress, smoking, consumption of sugary beverages, work and financial pressure, long and excessive work hours, insufficient physical activity and poor sleep hygiene [20, 28]. Clinical sleep studies have found other factors that interfere with sleep, such as ageing factors, chronic disease (e.g., cardiovascular disease or obesity), mood disorders (e.g., depression or anxiety) and sleep-related disorders, such as insomnia and obstructive sleep apnoea (OSA) [29]. The mental and physical damage caused by a single night of sleep deprivation may exceed the damage caused by the same lack of exercise or food [30].

Besides its ramifications for the health of individuals, sleep can also affect different areas of society and the economy. A study conducted by RAND in 2016 quantified the combined cost of insufficient sleep in five OECD (the organisation for economic co-operation and development) countries (Canada, USA, UK, Germany and Japan) is estimated at $600 billion a year [20]. The two main consequences of insufficient sleep and poor sleep quality are cognitive impairment and fatigue. Between 2005 and 2009, there was an estimated average of 83,000 car accidents related to driving when drowsy, resulting in more than 6000 deaths in the United States each year [31]. In certain professions, such as operating construction cranes or driving public transport vehicles, poor sleep can increase life-threatening risks [32]. The odds of work accidents were found to be nearly double in workers with OSA [32].

Given the fact that the role of sleep is related to well-being, personal health, disease and mortality and that it has significant social and economic impacts, researchers are increasingly interested in automated continuous sleep monitoring. The latest developments in digital consumer/commercial electronic products provide a less intrusive way to continuously monitor our physiological status. In addition, advanced deep learning tools have paved the way for the design of advanced sleep monitoring algorithms that may recognise patterns of sleep/wake or different stages of sleep based on physiological signals. The advances in computational methods have accelerated objective and non-invasive sleep monitoring for large-scale populations in a free-living fashion, which may improve our understanding of the impact of sleep on health and diseases [33].

1.2. Sleep and Well-being

The rise in life expectancy in the 21st century is continuing or even accelerating, not only in developed countries but also in developing countries. Although the increase in lifespan is indeed a great achievement in human history, it has brought some practical and economic challenges to individuals, families and society, such as the quality of later life. If it is a life of illness and disability, the benefits of a longer life cannot be realised. Neurodegenerative diseases such as Parkinson's and Alzheimer's may lead to various forms of disability and loss of independent living ability [34]. Early detection, treatment or continuous care may reduce the severity or delay disease onset.

A society with an ever-increasing population of elderly or disabled people will impair social productivity and limit the growth of the economy, and elderly care will increase the financial burden on public finances as well as families. The early onset of several neurodegenerative diseases may be accompanied by abnormal sleep behaviour. Such as, the rapid eye movement (REM) sleep behaviour disorder has been discovered as one of the early symptoms of Parkinson's [34–36]. This clinical phenotype demonstrated the potential of developing automated algorithms based on the physiological signals that may be collected from everyday wearable electronics [37].

In recent years, with the continuous development of digital health and preventive medicine, the pre-screening of diseases and monitoring of health status through consumer electronics has become prevalent in health research. One objective of preventative medicine is to delay the severity of disability through pre-screening so that the elderly can remain independent for as long as possible. Another goal is to delay the onset of disease through preventative care at an earlier age (e.g., childhood and early adulthood) [38]. Digital phenotype monitoring may provide an alternative method to pre-screening for early symptoms of neurodegenerative disease [38]. The physiological signals collected from wearable devices can provide ambulatory and real-time monitoring of our body and health status, which is essential to personalised healthcare. It can also support longitudinal studies to understand the correlation between sleep and disease.

1.3. Sleep Stage Monitoring

Sensors have been used to study sleep for decades. Traditionally, polysomnography (PSG) is the gold standard and the de-facto technique for sleep monitoring and assessment in clinical and laboratory settings as well as for diagnosing a subset of sleep disorders [39]. PSG recording can be classified into five stages, i.e., wake, REM and three types of non-rapid eye movement (NREM) sleep, including N1, N2 and N3 (More details regarding these sleep stages will be introduced in chapter 2) [40]. According to the American Academy of Sleep Medicine (AASM) rules [41], each stage lasts 30 seconds (i.e., a sleep epoch). Deep non-rapid eye movement sleep (N3) or slow wave sleep (SWS) is known to be the most "restorative" sleep stage, which controls hormonal changes that affect glucose regulation [42]. Long-term reduction in NREM sleep may adversely affect glucose homeostasis and increase the risk of type 2 diabetes [43]. REM

sleep dysregulation has played a central role in depression and Parkinson's studies [44, 45]. For instance, reduced REM sleep latency, along with increased REM sleep duration and REM sleep density, have been considered to be objective indicators of depressive disorder and inversely correlated to its severity [46, 47, 34]. The increased health research density in digital phenotypes by using inexpensive, mass-produced consumer wearables requires reliable algorithms that can classify sleep stages in longitudinal settings [48]. Beyond health and clinical applications, sleep monitoring has also been welcomed by self-trackers in the past decades [49].

Understanding the relationship between sleep stages and health can bring huge benefits to society. However, PSG study is expensive and burdensome, it is not suitable for more than two consecutive nights of sleep monitoring. Without using PSG equipment, the Actigraphy (embedded with accelerometers) provides a valid method for detecting sleep/wake and is commonly used for ambulatory monitoring of sleep time or rhythms [50, 51]. However, it is limited to monitoring sleep-wake as the actigraphy data may not contain sufficient information to discern sleep stages. The wrist movement does not reflect all the brain activities.

Heart motion can be monitored by the electrocardiogram (ECG), which can be used to derive heart rate (HR). HR is characterised differently in different sleep stages due to the substantial difference in the regulation of the autonomic nervous system (ANS) [52, 53]. The ANS system is regulated by sympathetic activity and parasympathetic (or vagal) activity which has 'opposite' actions where one activates a response in physiology while the other suppresses the stimulation [54]. The heart rate variability analysis is a well-established tool to characterise the cardiac autonomic activity [55]. Previous sleep physiological studies have demonstrated some characteristics vary over sleep stages [52]. This means that cardiac and movement activity can, in turn be used to separate sleep stages, which is of significant clinical relevance. Recent studies also demonstrated the feasibility of using these two modalities with machine learning and deep learning models for sleep stage classification [56]. These studies were based on non-open source data sets, which became an obstacle to developing new algorithms and reproducing results. Therefore the prospects of using these algorithms in real-world applications are further limited.

Many consumer electronics can monitor sleep stages, such as the Apple Watch, Fitbit band and Xiaomi band. Other ubiquitous sensing technologies were studied, including actigraphy [50], smart watches [6], WiFi [57], bed sensors [58] and radio signal based equipments [59], etc. Among them, in terms of reliability and usability, cardiac and movement (upper limb) sensing are considered promising modalities. The HR or heart rate variability (HRV) and limb movement data can be easily collected from lightweight research/consumer-grade devices (e.g., Apple Watch [6]). More detailed discussions regarding the emerging sleep monitoring technologies will be introduced in chapter 2.

These consumer wearables are capable of communicating with smartphones, which facilitates data collection and storage when used in large-scale research. However, due to the algorithms and data processing pipeline not being transparent to researchers, they were excluded from use in clinical sleep monitoring settings [60]. Lack of transparency can have an impact on the results, which may potentially make previous research irrelevant and incomparable, especially for open

science research. But the easy-to-collect nature of cardiac and movement sensing provided a potentially scalable method for large-scale and long-term sleep monitoring [1]. Longitudinal sleep monitoring with accurate details of sleep stages can be valuable for health and medical research. Therefore, the development of sleep stage algorithms based on wearable sensors can better serve the development of sleep science.

1.4. Problem Statement

The development of sufficiently robust sleep stage classification algorithms that can be compatible with wearable devices to discern human sleep/wake and sleep stages is an essential component of health applications in the area of ubiquitous computing and medical engineering. Long-term non-invasive sleep monitoring would provide a valuable solution for clinical applications, with the potential to improve human well-being. In practice, monitoring the sleep stage without using electroencephalogram (EEG), electrooculogram (EOG), and electromyography (EMG) signals poses a major challenge for sensor technology and algorithm development.

Recent research demonstrated that sleep stages could be discerned using handcrafted features that describe the physiology of cardiac activity and body movement activity [61]. In comparison with PSG scoring results, the performance of these methods remains low, suggesting a strong need for further improvement in terms of the model robustness and generalisation. In addition, many of these studies are not based on open-source datasets and code, the data pre-processing pipeline is also inconsistent between studies, which makes it impossible to conduct further comparable studies or reproduce the results. This is the **first challenge** for long-term free-living sleep monitoring using wearable sensing technology. In order to allow more health researchers to use sleep stage monitoring algorithms and to better serve algorithmic democracy, there is a need to develop advanced algorithms for sleep stage monitoring with adequate accuracy through open-source solutions.

Sleep stage classification based on wearable sensing data may present additional challenges. The proportion of each sleep stage duration is imbalanced and the amount of annotated sleep data is often insufficient. In addition, cardiac sensing data is known to contain personalised information such as health conditions. The inter-subject and intra-subject variability pose challenges to developing robust models. Demographic factors, including age, gender and body mass index, and breath-related sleep disorders, can be factors that cause differences in signals between subjects [62, 63]. These factors are also called personal attributes and can have a significant negative impact on the model's performance. This is the **second challenge** in algorithm development, which is to address the distribution differences among subject groups. It is necessary to develop models that can extract invariant features or make the model less affected by personal attributes.

Sleep stage classification monitoring can be realised by using multimodal sensing techniques, and how to best combine these modalities is the **third challenge** of this thesis. The third work of this thesis is to investigate whether recent advances in machine learning have shown promising performance in similar health applications, especially the use of deep learning is suitable for

sleep/wake and sleep stage classification using wearable sensors; and to explore how algorithm design can extract the most relevant information for sleep stage classification tasks. The general objective of this thesis is to achieve performance improvement with respect to the reliability and robustness of classifying sleep stages using movement and cardiac sensing data. It is important to note that the population used in this thesis consists primarily of adults with varying health conditions, while sleep in teenagers, children and infants is not covered.

1.5. Contributions and Thesis Outline

This section provides an overview and outline of the chapters in this thesis. Since most of the research in this thesis is the result of collaborations, this chapter will also introduce relevant published papers and the author's contributions to each paper.

1. **Introduction**
   This chapter focuses on how ubiquitous computing techniques can be used to monitor sleep/wake and sleep stages, and the main challenges of monitoring sleep outside the sleep lab. It is imperative that these challenges should be overcome in order to develop practical tools that can monitor the human sleep process outside of the sleep laboratory.

2. **Sensing and Analysis of Sleep/Wake and Sleep stages**
   In chapter 2, the author introduces the sleep physiology background, and the definition of each sleep stage, and summarises different commercial products and research prototypes together with their roles in sleep research. Each sensing approach was investigated with respect to usability, infrastructure requirements, performance, and limitations in different practical scenarios. Moreover, the capabilities of several traditional machine learning methods and deep learning methods to discern sleep/wake and sleep stages are investigated. Finally, the challenges of developing multi-modal sensing algorithms based on wearable sensors are summarised, especially using accelerometer and electrocardiogram (ECG). The author's contribution can be seen below:

   > Perez-Pozuelo, I., **Zhai, B.**, Palotti, J., Mall, R., Aupetit, M., Garcia-Gomez, J. M., ... & Fernandez-Luque, L. (2020). The future of sleep health: a data-driven revolution in sleep science and medicine. NPJ digital medicine, 3(1), 1-15

   The author of this thesis wrote the manuscript of data curating, data pre-processing, feature extraction, machine learning and data-driven methods. The author also contributed to the sleep-sensing section.

3. **Benchmarking Sleep/Wake and Sleep Stages Using Wearable Sensors**
   In chapter 3, the author benchmarked the capabilities of several traditional machine learning methods and deep learning methods to discern sleep/wake and sleep stages using limb movement data and ECG-derived heart rate data. It serves as a benchmark study and an example implementation of the pipeline approach using multimodal data. This

approach advocates for model transparency, alongside reproducibility by exploring these methods in the only open-access dataset, which includes participants with sleep disorders. The performance metrics of specific algorithms are presented in this work, along with guidance on algorithm selection based on classification tasks. Moreover, in this chapter, a deep ensemble model architecture was introduced that shows promising improvements in performance across different sleep stage classification tasks. The final output also includes an open-source toolkit to facilitate the reproducibility of experiments. The work in this chapter is mainly to address the **first challenge**. The author's contribution can be seen below:

> **Zhai B.***, Perez-Pozuelo, I.*, Clifton, E. A., Palotti, J., Guan, Y. (2020). Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(2), 1-33. * Equal contribution to this work

In this work, the author of this thesis co-planned the project and designed the models and experimental plans in collaboration with collaborators. The author generated the python scripts for data curating, data pre-processing, feature extraction, traditional machine learning models, deep learning models, sleep stage and sleep metrics evaluation, and conducted all the experiments. The author proposed the deep ensemble model architecture and the time deviation metric. The author also wrote the related work, methodology, and results sections, and contributed to the writing of other sections.

The results of this work suggest that three-stage sleep classification is the most promising task, based on data collected from wearable sensing devices and that the performance of sleep/wake classification can be improved. However, it is difficult to distinguish the NREM sleep stages of the four-stage and five-stage sleep classification tasks without using EEG, EOG and EMG.

4. **DisSleepNet**

   The work in chapter 4 aims to address the **second challenge**. This chapter aims to explore the use of disentanglement learning to extract personal attributes-free (e.g., age, sleep apnoea) representations for three-stage sleep classification using ubiquitous sensing. One of the common approaches is to manually extract clinically relevant descriptors and then feed them to deep learning models to learn high-level abstractions. However, cardiac and movement sensing can be affected by personal attributes such as age, BMI and sleep disorders (e.g. sleep apnoea). To mitigate the effects of these personal attributes, this chapter investigates the use of novel methods to identify representations that are less affected by personal attributes.

   In preparation to submit to International Joint Conference on Artificial Intelligence:

> **Zhai, B.**, Guan, Y., DisSleepNet: Disentanglement Learning for Personal Attribute-free Three-stage Sleep Classification Using Wearable Sensing Data

In this work, the author of this thesis conducted the planning, coding, and experimenting under the supervision of co-authors.

5. **Ubi-SleepNet**
   Although the deep learning methods achieved a higher performance compared with traditional machine learning methods, learning and summarising effective representations from multi-modal wearable sensing data, in a way that takes advantage of the complementary and redundancy of multimodal data, is a challenging task for sleep stage classification. This is because there are no established clinical rules for using HR/HRV features and activity counts for sleep stage annotation in sleep physiology research.

   To address this **third challenge**, in chapter 5, the author systematically evaluates the prevalent multimodal fusion techniques for wearable sensor fusion. The proposed fusion approaches have significantly improved the model's robustness. Furthermore, the author also adopted a gradient-based visualisation method on deep learning models to suppress less relevant information from the handcraft features. Experimental results conducted with human subjects demonstrated that the final simplified visualisation results increased the understanding of the decision-making process, based on the wearable signals, with respect to each sleep stage. The author's contribution can be seen below:

   > **Zhai, B.**, Guan, Y., Catt, M., & Plötz, T. (2021). Ubi-SleepNet: Advanced Multimodal Fusion Techniques for Three-stage Sleep Classification Using Ubiquitous Sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(4), 1-33.

   In this work, the author of this thesis carried out project planning, coding, and experimentation.

6. **Summary** In chapter 6, the author provides a summary of the challenges, insights and results of the research presented in this thesis. It highlights the current progress with respect to sleep stage monitoring using wearable sensors and explores how future work could extend such approaches to improve sleep stage monitoring performance.

Chapter 2. Background and Literature Review

Human beings spend about one-third or more of their lives sleeping. Sleep is vital to homeostasis, memory, cognitive capability and behavioural performance. In clinical sleep practice, overnight polysomnography (PSG) paired with clinical evaluation is deemed the gold standard as well as the de-facto technique for objective assessment of sleep architecture/pattern and for diagnosing sleep-related disorders such as parasomnia, sleep-disorder breathing (apnea and hypopnea), and REM sleep behaviour disorder [64, 40]. PSG-based sleep assessments are often recorded in a controlled environment, such as a sleep lab, which requires sleep clinicians to set up the PSG recording equipment. The sleep recordings also require professional annotations [40]. Over the past two decades, various portable devices have developed to monitor sleep in less obtrusive ways [65]. The motivation is to reduce the cost of monitoring and the burden of acquiring data, which can ultimately be used in the long-term, more natural environment. However, data acquisition is a profound challenge, despite the inclusion of less obtrusive and stigmatising long-term sensing mechanisms. Long-term sleep monitoring may be compromised by missing data, which may mislead pre-screening for sleep-related health problems. This chapter begins with a discussion of sensing techniques for sleep data acquisition in clinical and free-living settings, including an overview of traditional and novel approaches and their advantages and disadvantages. Secondly, data pre-processing and feature extraction methods are introduced. Finally, the strengths and limitations of emerging algorithms are discussed with a particular focus on novel data-driven technologies, including machine learning and deep learning approaches.

2.1. Traditional Sleep Monitoring in Laboratory Settings

PSG study has been adopted in clinical sleep assessment since the 1960s [66]. It consists of the various type of sensors (e.g., electrodes attached to the skin) that can measure: (1) brain activity through electroencephalogram (EEG), (2) airflow (e.g., using thermistors or nasal pressure transducers), (3) breathing effort and rate through respiratory inductive plethysmograph (RIP), (4) blood oxygen levels (e.g., using pulse oximeters), (5) body position (e.g., using accelerometers), (6) eye movement through electrooculography (EOG), (7) electrical activity of muscles through electromyography (EMG) and (8) heart rate through electrocardiogram (ECG) or pulse oximeters. Figure 2.1 and figure 2.2 show the PSG equipment used in the sleep laboratory and the collected signals.

Traditionally, PSG studies are expensive and require participants to sleep in a laboratory setting. The data are then scored by sleep experts. Due to its limitations, PSG remains impractical for long-term sleep monitoring. As a result, the scalability of this technique for large-scale

Figure 2.1 A typical polysomnography (PSG) equipment used in sleep laboratory

population-based studies is very limited, particularly when the aim is to assess typical sleep patterns in a free-living, naturalistic condition.

Ambulatory PSG is an alternative device that typically uses a smaller number of sensors, such as by reducing EEG channels. This device can be used not only in the laboratory but also for sleep monitoring at home [41, 67, 68]. However, although the Ambulatory PSG provides a simpler solution to address some issues, it remains both expensive and burdensome [69].

Another conventional method used to assess sleep in clinical settings is the videosomnography (VSG). VSG uses cameras to record sleep activities during PSG studies. These video recordings and the PSG data were then used to assess sleep disorders in clinical settings. Recent advances in telemedicine have made the use of home VSG increasingly possible [70]. However, VSGs suffer from similar scalability issues to PSGs, as they often require experts to score in a time-consuming manner.

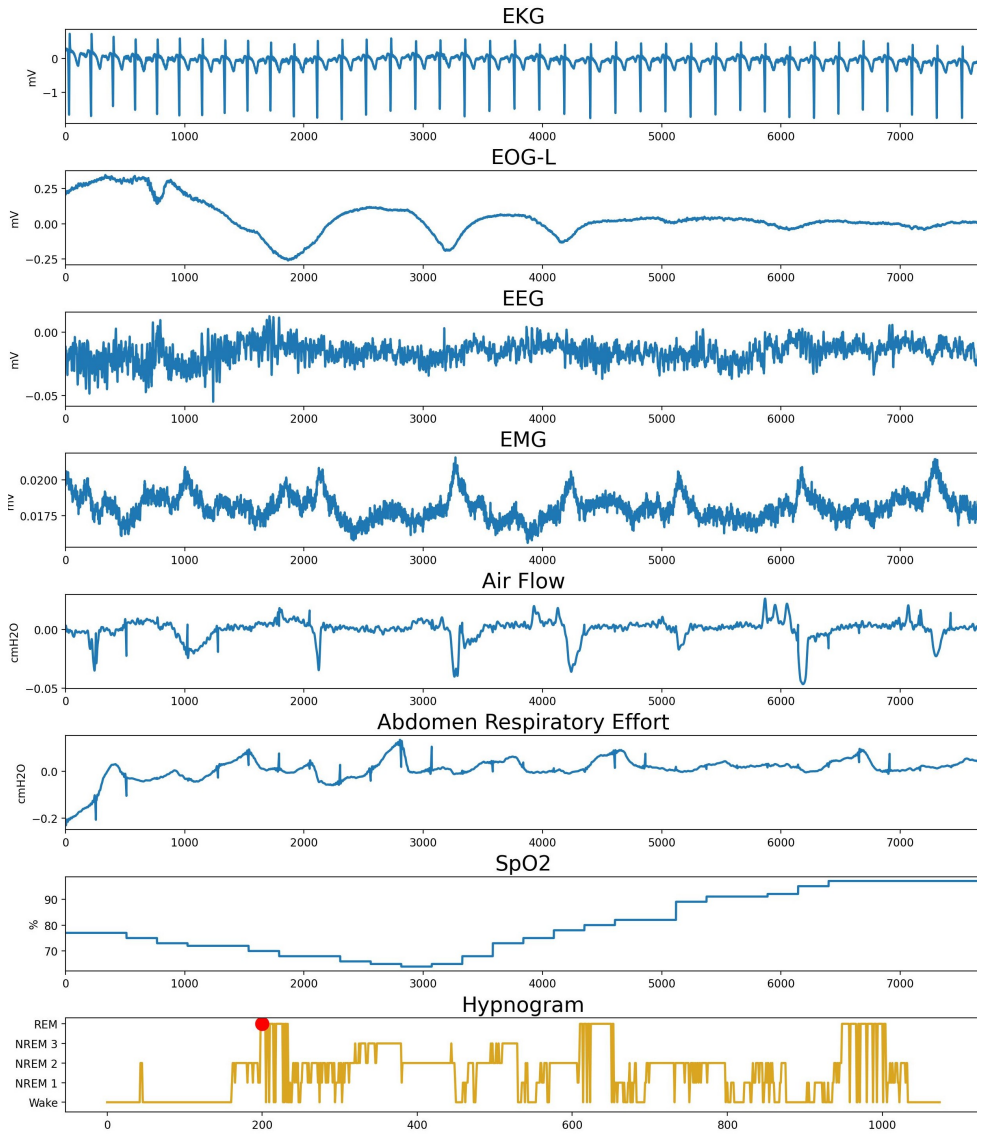


Figure 2.2 An example of PSG data for a sleep epoch (30s) base on the selected channels (Sampling Rate at 256Hz). Given these PSG signals, a red dot that appears on the hypnogram indicates the corresponding sleep stage.

2.2. Sleep Architecture

Human sleep can be divided into two categories, which include rapid eye movement sleep (REM) and non-rapid eye movement sleep (NREM). REM sleep is often associated with dream activities [71]. According to the AASM guidance, NREM sleep can be further categorised into three stages: N1, N2, and N3, each owning to its own specific characteristics and distinct EEG patterns [41].

2.2.1. NREM Sleep

The N1 stage is the state in which the brain transitions from wakefulness to sleep. It is the lightest stage of sleep, with a frequency of 4-7 Hz and an abundance of theta waves, accompanied by reduced alpha waves (8-12 Hz), often seen during wakefulness [40]. The N2 stage can last 20 minutes. It often has a unique characteristic, including bursts of waves with a frequency of 11-16 Hz [71]. It can also often be recognised by the appearance of k-complexes, which are brain waves with a duration greater than 0.5 seconds [40]. Compared to N1 and N2, N3 is the deepest stage of sleep in NREM sleep. During this time, our body relaxes further, and our heartbeat and breathing decrease further [40]. The pattern in the EEG signal is characterised by the presence of slow waves with a frequency of 0.5 - 3 Hz, also known as delta sleep or slow-wave sleep (SWS) [40]. Research has shown that this stage is critical for the recovery and growth of bodily tissues. In the early sleep cycle, stage N3 usually lasts 20 to 40 minutes [71].

2.2.2. REM Sleep

The REM sleep stage is characterised by sharp theta waves or patterns that appear in the EEG signal, like an awakening. It may account for a quarter of total sleep time [71]. It is also associated with the lowest muscle tone, and our bodies are often temporarily paralysed in addition to the muscles that control eye movement and those that control cardiorespiratory activity [40]. During this stage of sleep, although our eyes are closed, the eyeballs sometimes move rapidly, so it is often referred to as REM sleep.

The alternate appearance of NREM sleep and REM sleep constitutes a sleep cycle. A healthy person typically has 4-5 sleep cycles in one night [40]. As the sleep cycle increases, the proportion of REM sleep increases, and the proportion of NREM sleep decreases. Five sleep stages can be annotated by human experts from the recorded PSG signals, i.e., by analysing the characteristics of EEG, EMG and EOG.

2.3. Sleep Monitoring Outside the Laboratory

Accurate long-term sleep monitoring in natural environments could help researchers understand the impact of sleep on our health. Recent advances in miniaturised sensors present an excellent opportunity to develop wearable sensing solutions for the sleep monitoring of free-living individuals at the population level. Several studies investigated sleep monitoring outside the lab, including the use of actigraphy, heart rate sensors and other wearable technologies [72–

77]. Several previous works suggest that using a single modality such as heart rate may not be sufficient to accurately identify detailed sleep stages [78]. In some cases, REM sleep and wakefulness may appear to have a similar range of values on HRV features, but hands may not move during REM sleep. The availability and scope of digital products for measuring sleep have expanded significantly over the past few decades. Consumer-grade devices that can monitor sleep are becoming smaller and more affordable [1]. However, comparing the performance of consumer devices and clinical devices, using consumer products for clinical-grade sleep monitoring remains a challenge. To date, few studies have validated or systematically evaluated the reliability of these products against gold standard PSGs [79].

2.3.1. Sleep Diary and Actigraphy

In clinical sleep monitoring settings, in addition to PSG devices, sleep diaries, actigraphy and accelerometers are also widely accepted in the study of human activity and circadian cycles in free-living environments [80]. Sleep diaries are also a widely accepted tool for subjective sleep/wake and sleep quality assessment. However, it suffers from strong recall bias and may contain less accurate sleep measurement, and cannot record sleep stage duration.

Actigraphy is a method for detecting sleep/wake and is commonly used for ambulatory monitoring of sleep time or rhythms [50, 51]. It is a type of wearable wristband that consists of various sensors that can monitor the light-off time and the movement of the limbs (using an accelerometer) [50, 51]. However, it is limited to monitoring sleep-wake, as the actigraphy data may not contain sufficient information to discern sleep stages.

Recent advances in artificial intelligence and larger studies combined with PSG have led to algorithm improvements in sleep monitoring [81]. However, Two limitations of actigraphy and accelerometry are: (1) the lack of standardisation of approaches for sleep stage monitoring and (2) the lack of assessment techniques for daytime sleep. Nowadays, Wearables are often embedded with a combination of various types of sensors (such as heart rate monitors, miniaturised ECG/EEG, pulse oximetry, blood pressure monitors, galvanic skin conduction, light sensors, gyroscopes, and barometric altimeters) [1]. Sleep can also be monitored through a combination of sensing devices, such as wrist-worn accelerometers, microphones, and pressure sensors under the mattress, wireless communication systems and video cameras [82]. However, this increases usability issues such as synchronising data collected from multiple sensing devices and interpreting prediction results from multiple applications [79].

2.3.2. Emerging Sleep Sensing Technologies

The ultimate goal of studying ubiquitous computing in sleep monitoring is to achieve non-invasive sensing solutions capable of monitoring sleep-related physiological signals. Incorporating different types of sensors into the objects we interact with every day is more appealing than using multiple redundant sensors to gather homogeneous information. Over the past few decades, various sensing technologies such as under-mattress pressure sensors, microphones, image sensors and radio frequency (RF)-based sensors have been developed to track different sleep-

related information such as time spent in bed, movement, breathing, snoring, heartbeat, body and room temperature [83–85]. While these sensors have potential value for clinical and epidemiological studies as well as sleep health education, little is known about their performance compared to gold-standard PSG annotations, and more research is needed to assess their reliability and usability. Especially the sleep stage sensing technology based on audio and video signals has strong privacy concerns when used in the bedroom. Several techniques have been discovered in recent years for sleep monitoring but are still in the early stages of development (such as radio wave-based methods [72]), while others have been around for a longer time (such as smartwatches) [86, 87].

Figure 2.3 Sensing technologies can be used for sleep/wake or sleep stages monitoring [1].

Bed sensors: Bed sensors normally refer to the sensing technology that can sense pressure changes when the subject is lying on a mattress. Some researchers have developed highly sensitive pressure sensors which can be mounted on mattresses to record body movements and heart rate (also called ballistocardiography (BCG)) during sleep [88]. The small-scale body movements can reflect breathing, limb movement, and cardiac activities. A typical sensing method can monitor changes in air pressure beneath an individual while the subject is lying in bed [89–91]. In the work of Sadek et al. [92], researchers have developed micro-bend fibre optic sensors underneath the mattress to monitor pressure changes. Movement monitored by bed sensors (either on a mattress or bed frame) includes limb motion, breathing motion, and cardiac motion, which can be used to develop sleep monitoring models. Several commercially

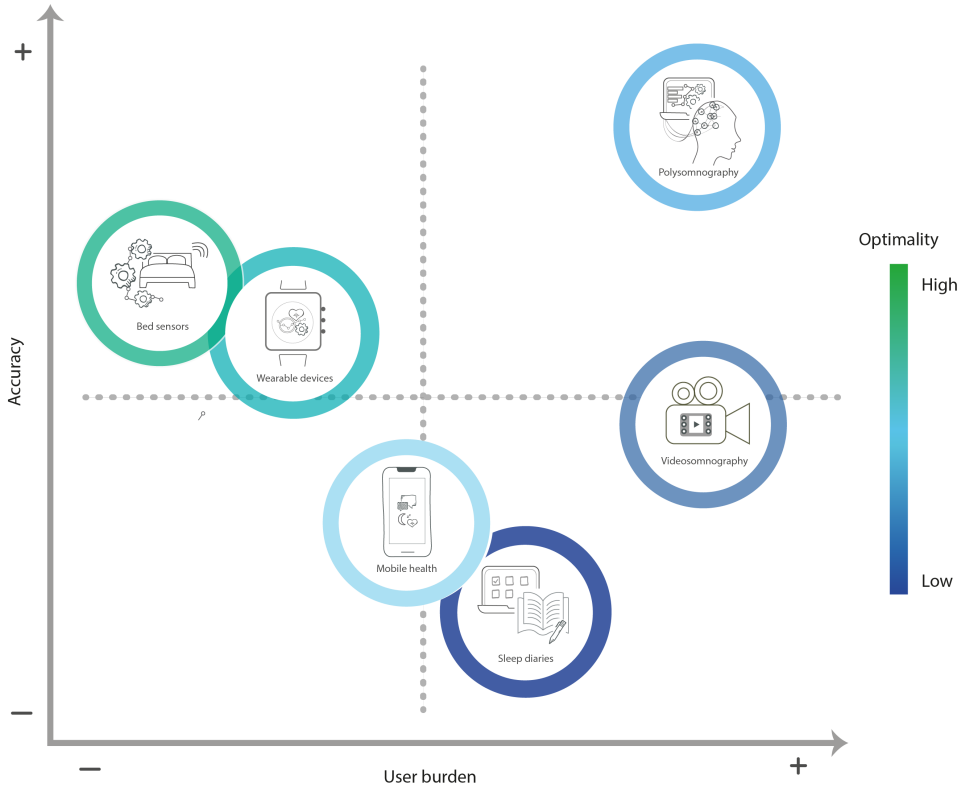


Figure 2.4 The usability and performance trade-off over different devices [1].

available products have been developed by Apple (Beddit), Nokia and Withings. However, several covariates affect the performance of these methods, ranging from differences in a posture to inter-subject variability in body mass index (BMI) and pre-existing clinical conditions [91].

Consumer-graded wireless EEG and reduced-array EEG: EEG is an important tool for sensing brain activity during sleep, although the number of sensors mounted on the scalp makes it a cumbersome device. Compared to PSG studies, the recent development of reduced-array EEG (wireless recording) has been used for various neuropsychiatric tests and applications due to its ease of use [93, 94]. The main disadvantage of these devices is that the electrodes embedded in the cap are visible, and the form factor limitations prevent comfortable, continuous, and long-term sleep monitoring. Furthermore, once the electrode gel dries, the signal quality decreases, and the gel leaves residues [95], which may impact the wearing experience. The lightweight headband EEG (e.g., Sleep Profiler^{TM} and Dreem^{TM}) can monitor sleep in a natural environment. But it takes extra effort each time of wearing to adjust the equipment position to reduce the skin impedance to an acceptable level [96]

Furthermore, a recent study has shown that the ensemble model trained on data collected from a single-channel EEG can achieve a concordance rate of 0.87 when compared to the expert ratings on the same signal [97]. However, the research was conducted in clinical settings, the usability and signal quality remains unknown when deploying this equipment in a free-living environment. The positive impact of this study demonstrates the feasibility of automated scoring in long-term sleep monitoring applications.

In-ear EEG is unconventional EEG equipment that has shown promise in recent years. For example, in the work of Mikkelsen et al. [98], sleep stages were predicted by a machine learning model based on in-ear mobile EEG data, showing promising results compared to manually scored PSGs, although experiments were limited to laboratory setting.

Wireless EEGs have gained attention in recent years, with several established companies, as well as start-ups, launching products [99]. Their performance for sleep monitoring has been compared to conventional EEG, which is a part of PSG and has demonstrated strong results [93, 94]. Further, Koley and colleagues showed that automatic scoring using ensemble models on a single channel EEG could yield agreement rates of 0.87 when compared to expert scoring of the same signal [97]. Whilst this study was conducted in a clinical environment and hence lacked the validation of free-living conditions, together, these investigations show that the results of conventional EEG can be approximated by simpler devices and automated scoring algorithms. Likewise, in 2017, a prototype based on an in-ear EEG sensor showed 74% agreement with hypnogram annotated from PSG data [100].

While the performance of these wireless, miniaturised, and in-ear EEG devices is promising, more extensive studies are needed to determine their feasibility in sleep epidemiology and free-living settings and in applied sleep research. In addition, these ear EEG devices commonly adopt around-ear or/and in-ear style and are made of silicone materials, which offer a bearable wearing experience, making them less popular than the mass-produced wearables [101].

Smartwatches and fitness trackers:

Smartwatches and smart wristbands have demonstrated that consumer-grade sleep monitoring products could be accepted in everyday life. These devices estimate sleep metrics and sleep stages through the use of motion signals (accelerometers, as described in previous sections) and heartbeat sensing data. Although several studies have evaluated the sleep stage prediction performance for these consumer products (Fitbit, Garmin, Misfit, Apple, Polar, Samsung, Withings and Mio) with respect to expert PSG annotations, the data processing pipeline and the algorithms used in the study remain largely undisclosed [102].

Mobile phone sensing: In addition to wearable electronics, smartphone-based sleep monitoring applications have also attracted a lot of attention from sleep researchers. Many of these applications estimate sleep parameters and sleep stages using onboard sensors, including gyroscopes, microphones, accelerometers and light sensors. For example, in the work of Hao et al. [103], the researcher developed an application called iSleep which adopts the smartphone's built-in microphone to detect events that occur during sleep, such as body movement, coughing, and snoring. The software classifies events (snoring, coughing, sleep) with over 90% accuracy under different environmental conditions. However, processing high sampling rate data poses a big issue for battery life due to the significant energy consumption. In recent years, applications using onboard sensors on smartphones have also become very popular in the consumer market, such as the Sleep Time applications. However, the accuracy of these applications is very low. In the work of Bhat et al. [104], compared with PSG's epoch-wise annotations, the app only achieved 45.9% accuracy.

Ultrasound sensors, WiFi and radio signal approaches: Ultrasonic sensors can be used to detect object motion and surface changes, which can partly represent body movement and breathing patterns during sleep [91, 105, 106]. These Doppler ultrasound sensors measure sound waves that are reflected from moving objects or the changes on the surface of the object (e.g., chest movement during breathing). The biggest benefit of this technology is that it avoids the radio interference problem. This technique mirrors that used in conventional radar systems and uses signal processing methods to extract the information related to breathing rate, heart rate and body motions. The method has been shown to be able to detect physical motion with 86% recall and less than 10% error [107].

However, wireless-based approaches face some challenges when deploying them in clinical research owing to a) the non-standardised measurement methods; b) the lack of precise understanding of the physiological origins that influences the signal waveform; c) comparatively low reliability and specificity of these signals to the existing clinical methods (for example, WiFi signals may be scattered by multiple subjects), which may hamper its wide applications in health and medical research [108].

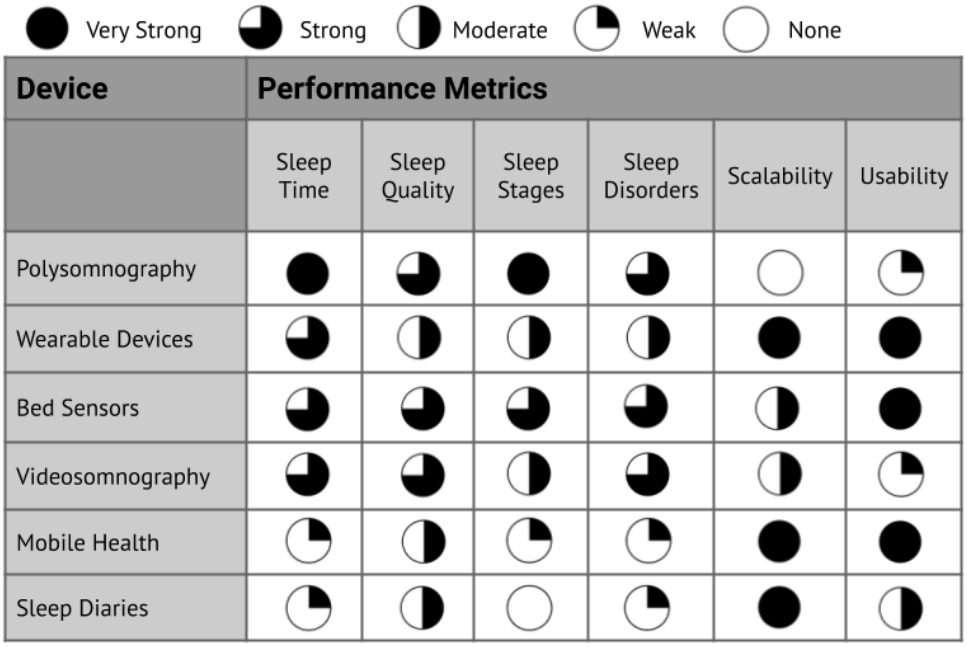


Figure 2.5 The evaluation of each sensing technology with respect to their performance metrics [1].

Likewise, high-frequency and submillimeter-wave radio technology has been shown to capture physiological activities. Similar to the ultrasound, the bouncing back radio waves could be used to extract the breathing patterns, heart rate and body movement [109–111, 107]. These radio signals can be used to determine sleep stages, as shown by Zhao and colleagues [112], as well as to monitor insomnia [113]. The technology encounters the same challenges as ultrasound, which are sensitive to environmental changes. Moreover, it is subject to electromagnetic interference. In addition, measurement performance is highly dependent on the individual being monitored, sleeping position, objects in the bedroom and radio interference [114].

Figure 2.3 provided an overview of the latest technologies that can be used for sensing sleep/wake or sleep stages. To summarise the usability and performance of each sensing technology for sleep monitoring, Figure 2.4 and Figure 2.5 demonstrated the user acceptance, usability versus the performance of each technology.

2.4. Data Pre-processing and Feature Extraction

Once the sleep sensing data has been collected, it should be pre-processed before the modelling stage. In recent years, integrating data collected from various types of sensors has become a trend in order to develop non-invasive sleep monitoring solutions [1]. However, several challenges should be considered in data preprocessing; for example, the sampling rate may vary depending on the quality of the sensing system, sensors, amplifiers, electrode materials used, and engineering and manufacturing processes may also lead to differences in noise topology. In addition, data measurement, processing and storage may also differ between sensing systems. For example, depending on the application and device, it might store r-wave to r-wave interval (RRI) instead of raw ECG. Therefore, before any feature extraction or modelling, the data needs to be cleaned and filtered to remove artefacts that are specific to the modality employed.

2.4.1. Data Pre-processing

Several aspects should be considered in terms of pre-processing the data before it can be used for analysis. The preprocessing depends largely on the type of application being built and the characteristics of the data itself (e.g., data quality, standardisation of data formats). Pre-processing signal data includes two main operations: (1) fusing signals provided by different sensor types; (2) missed data detection and imputation. The missing sensor data may be caused by one or more of the following reasons: 1) The user did not wear it, 2) A battery power supply issue, 3) System function errors (for example, program bugs, running out of memory space, or communication problems). Missing data can be detected by various algorithms, such as threshold-based methods and smoothness detection.

Artefacts in physiological data are another common phenomenon. Depending on the nature of the data, several pre-processing approaches can be applied. Smoothing and de-noising-based tools can remove unwanted spikes, trends and outliers from the signal [115]. For example, polynomial detrending methods can remove continuous quadratic or linear trends that may be caused by changes in skin impedance which often appear in ECG signals [116]. Similarly, the rolling median filter can also remove unwanted spikes from inter-beat signals. To remove the influence of external electromagnetic interference, band-pass filters are commonly used for removing such artefacts. The ultimate goal of denoising is to ensure that the noise follows a specific distribution, such as a Gaussian distribution [116].

Besides denoising and smoothing, re-sampling is also an essential technique in preprocessing stage to ensure the consistency of the data collected from different types of sensors in the temporal dimension. Linear or polynomial interpolation can be used to fill missing or corrupted

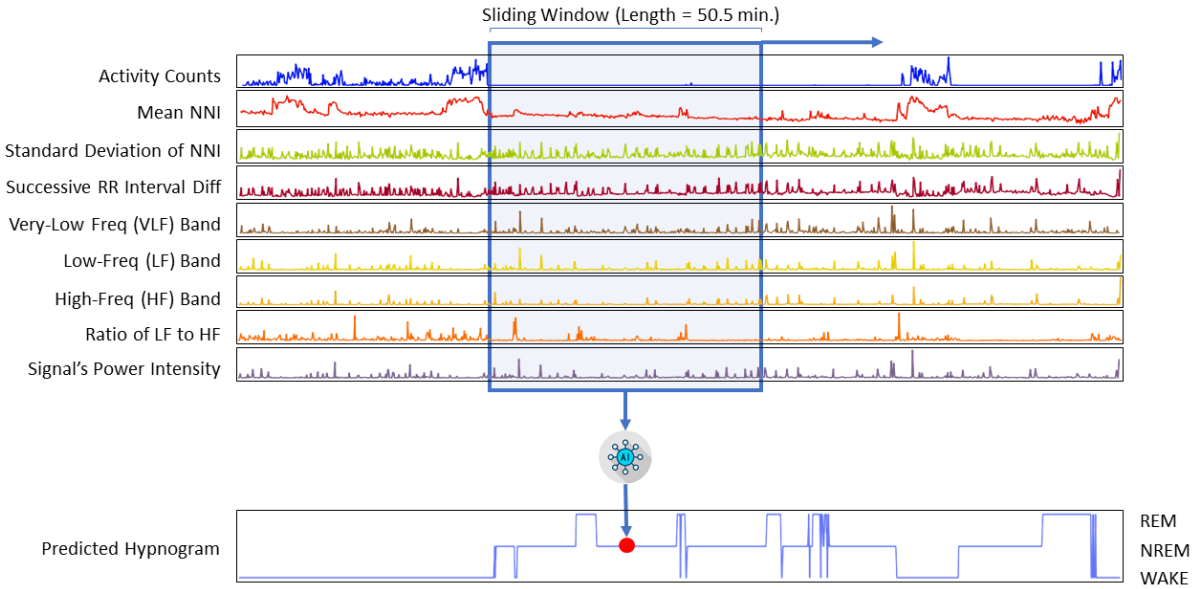


Figure 2.6 An example of using sliding window method on activity counts and HRV features for sleep stage classification

data [116]. These methods can suppress the noise level and variability in the signal and transform the data into a predefined range without changing its distribution.

2.4.2. Sliding Window Method

Sleep sensing data usually consists of multimodal time series data. Sliding window methods are commonly used to segment time series data into a finite set of data frames which can be used to extract heuristic features that can be used for machine learning (ML) models and/or deep learning (DL) models. Figure 2.6 demonstrated an example of a sliding window method, taking a period of night sleep data that corresponded to NREM sleep. The sliding window method consists of two hyperparameters: a window length and a stride. Window length refers to how long the time series data should be split. A stride parameter indicates how far the window should move at each step, and this usually determines the redundant information contained in the window. The hyperparameters of the sliding window are usually determined by domain knowledge or convention. For example, in sleep monitoring, EEG is the gold-standard signal for sleep stage classification. In sleep monitoring, EEG is the gold standard signal for sleep stage classification. The window length of processing EEG signals for sleep stage classification is usually set to 30 seconds, and the stride is set to the same value as the window length, which means there is no overlap [41].

2.4.3. Heuristic Method Based Feature Engineering and Sleep Physiology

Once the data has been preprocessd, feature extraction takes place to deal with unstructured data. Feature extraction from signals can be performed through a variety of methodologies that may fall under heuristic-based approaches or statistic feature-based approaches. For time series data, sliding window methods are commonly used to segment the data with finite lengths into

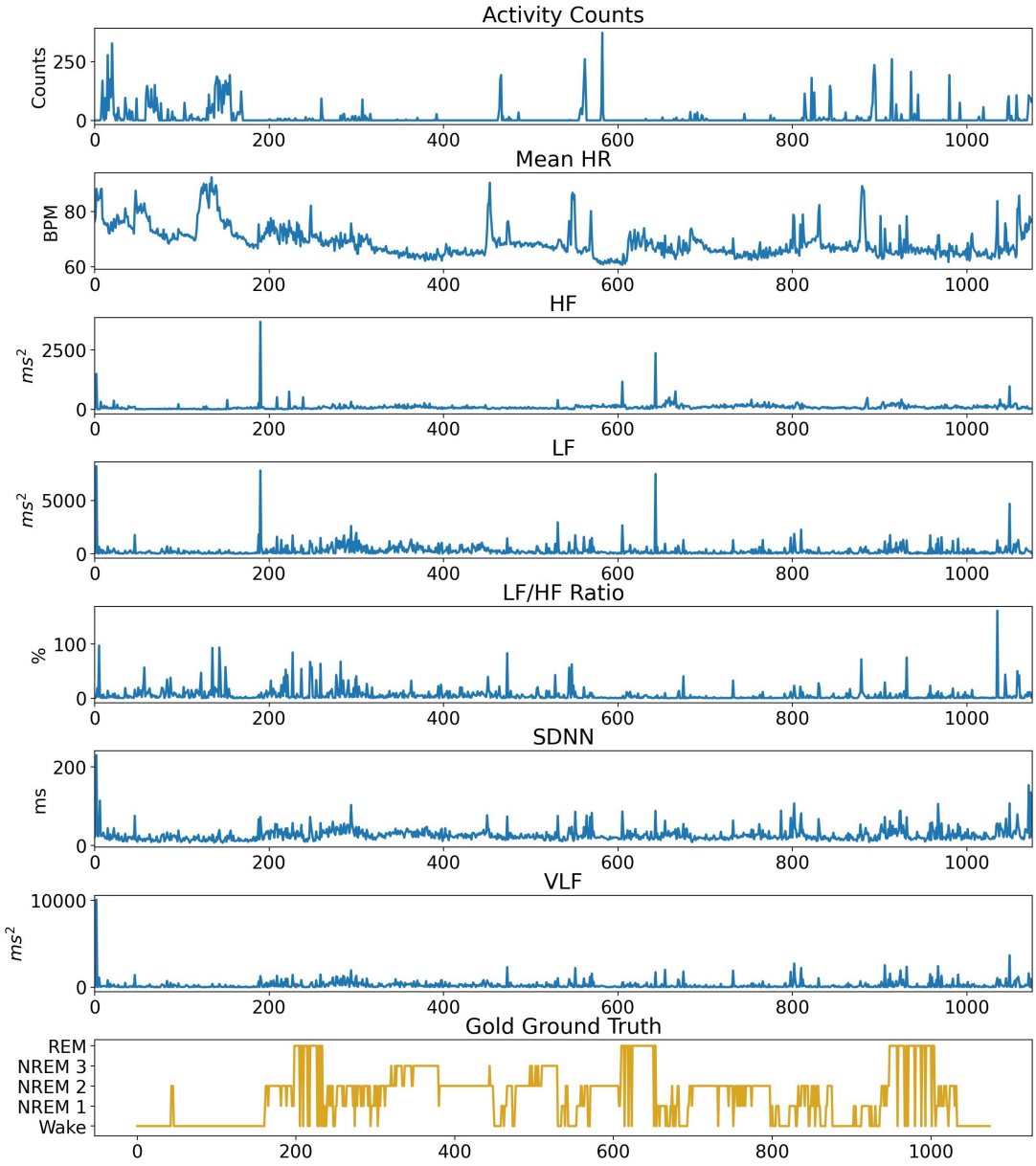


Figure 2.7 Example of HRV data for the entire night sleep based on the selected features

frames. The statistic feature approach could extract features such as mean, standard deviation, energy, quantile, and entropy from the data segmented by the sliding window method. Heuristic methods usually extract features based on domain knowledge. For example, during sleep, cardiorespiratory activity is regulated by the autonomic nervous system (ANS) [54, 117–121]. Due to the differences in ANS manifestations, including sympathetic and parasympathetic (or vagal) tone, heart rate variability (e.g., low frequency and high frequency) is characterised by different sleep stages [122]. Usually, the parasympathetic and sympathetic actions have "opposite" effects where one activates a response in physiology while the other suppresses it [123].

Cardiovascular autonomic control plays a vital role in sleep, varying among the transition to different sleep stages. The modulation of the ANS regulates cardiovascular functions during sleep onset and sleep stages [125, 126]. Heart rate variability (HRV) analysis is a classic tool for ANS analysis. Table 2.1 listed the most frequently used heart rate variability features.

Time Domain Features	
Mean HR ♥	*Mean heart rate for that window*
Maximum HR ♥	*Maximum heart rate for that window*
Minimum HR ♥	*Minimum heart rate for that window*
Std HR ♥	*Standard deviation for the heart rate for that window*
SDNN ♥	*Standard deviation of Normal-to-Normal interval (NNi)*
SDSD ♥	*Standard deviation of NNi differences*
NN50 ♥	*Number of NNi differences greater 50ms*
pNN50 ♥	*Ratio between NN50 and total number of NNi*
NN20 ♥	*Number of NNi differences greater 20ms*
pNN20 ♥	*Ratio between NN20 and total number of NNi*
RMSSD ♥	*Root mean of squared NNi differences*
Median NNi ♥	*Median of NNis*
Range NNi ♥	*Range between smallest NN intervals to largest NN intervals*
CVSD ♥	The coefficient of variation of successive differences , the RMSSD divided by mean NNi
Coeff. of Variation of NNI ♥	*The Coefficient of Variation of NNi, i.e. the ratio of sdNN divided by mean NNi*
Geometrical Domain Features	
Triangular Index ♥	*The HRV triangular index measurement is the integral of the density distribution (that is, the number of all NN intervals) divided by the maximum of the density distribution (class width of 8ms)*
Frequency Domain Features	
Low Frequency ♥	*Low Frequency is the variance (i.e., power) in HRV in the Low Frequency (.04 to .15 Hz). Reflects a mixture of sympathetic and parasympathetic activity*
High Frequency ♥	*High Frequency is the variance (i.e., power) in HRV in the High Frequency (.15 to .40 Hz). Reflects fast changes in beat-to-beat variability due to parasympathetic (vagal) activity*
Variance in Low Freq. ♥	*VLF is the variance (i.e., power) in HRV in the Very Low Frequency (.003 to .04 Hz). Reflect an intrinsic rhythm produced by the heart which is modulated by primarily by sympathetic activity*
Low/High Freq. Ratio ♥	*The LF/HF ratio is sometimes used by some investigators as a quantitative mirror of the sympathy/vagal balance*
Norm. Low Freq. Ratio ♥	*Normalized low frequency ratio calculated from the raw values of low frequency band (LF or HF) divided by the total spectral power*
Norm. High Freq. Ratio ♥	*Normalized high frequency ratio calculated from the raw values of high frequency band (LF or HF) divided by the total spectral power*
Mean NNi ♥	*Mean over the NN intervals*
Total Power ♥	*Total power of the density spectral*
Non-linear Domain Features	
Cardiac Sympathetic Index ♥	*Cardiac Sympathetic Index [124]*
Mod. Cardiac Symp. Index ♥	*A modified cardiac sympathetic index calculated by* $\frac{SD2^2}{SD1}$
Cardiac Vagal Index ♥	*Cardiac Vagal Index [124]*
SD1 ♥	*Poincaré plot standard deviation perpendicular to the line of identity*
SD2 ♥	*Poincaré plot standard deviation along the line of identity*
SD1/SD2 Ratio ♥	*Ratio of SD1 to SD2*

Table 2.1 Full set of cardiovascular related features grouped by cardiovascular domain [5].

HRV is typically higher during the night, reflecting the fact that sleep is a state in which vagal activity, characterised by rapid fluctuations in activity controlling coronary artery tone, HR and systolic blood pressure, is dominant [127–129]. Thus, HRV shows a nocturnal increase in the deviation of mean RR intervals. These deviations also differ between sleep stages. During the NREM sleep, the HRV analysis illustrated a higher parasympathetic tone compared to the REM sleep, which is characterised by likely sympathetic hyperactivity associated with a vagal withdrawal [130–132]. The sympathetic activity can be reflected in heart rate (HR) and standard deviation of normal-to-normal heartbeat/interbeat intervals (SDNN). The spectral power in

the high-frequency (HF) band of 0.15-0.4 Hz indicates parasympathetic modulation, which is activated by the stretch receptors via respiratory stimulation [133]. In contrast, the spectral power in the low frequency (LF) band of 0.04-0.15 Hz is assumed to reflect the sympathetic tone [134, 135, 133, 123].

The wakefulness-sleep transition is accompanied by about 15% decrease in blood pressure, HR and accompanied by increased HF power and decreased LF power and LF/HF ratio [130, 52, 131]. HRV analysis has also demonstrated that the HF band doubles in relative power when going from quiet wakefulness to non-REM sleep [136]. The REM sleep is accompanied by increased HR, LF power, and LF/HF ratio and reduced HF power, rising toward wakefulness, showing the increased and predominant sympathetic heart modulation [130, 52, 131]. Not all sleep stages are associated with brain activity. A study performed by Desseilles et al. found that HRV analysis combined with brain imaging has identified close connectivity between autonomic cardiac modulation and activity in brain areas such as the amygdala and insular cortex during REMS, but no connectivity between the brain and cardiac activity during different non-REMS stages [52]. Figure 2.7 demonstrated an example of a typical night's sleep of the same adult in terms of using HRV features and activity counts.

2.4.4. Machine Learning Methods

Once the features have been extracted, the classification model can learn the underlying difference of each sleep stage from a given corpus and infer the new cases/instances. This is a typical task in the field of ML and DL, and a significant number of models have been developed. ML and DL models have been broadly used in real-world applications. For example, in computer vision, deep learning-based models achieved state-of-the-art performance on object recognition, segmentation, and generation, while in the field of natural language processing, machine translation and sentiment analysis have been embedded into commercial services in our daily lives. Especially the DL models are used to mimic human cognitive functions, reasoning, and problem-solving abilities and have brought about a paradigm shift in digital medicine research too. Increasingly, ML is changing research methodology and facilitating the personalisation of sleep medicine through its advancements [137].

Regarding sleep science, the use of ML and DL is multifaceted. First, it makes it possible for us to understand our sleep habits and sleep health. For example, it can suggest an appropriate bedtime according to the user's preferences and improve sleep hygiene [138, 139]. Secondly, it can speed up the pre-screening of sleep-related health problems, expand population coverage, and enable the automation of analysis with lower computing resource costs. For example, converting sensor data into predefined knowledge (e.g., categorical labels), thereby providing an inexpensive and objective alternative to manual sleep stage scoring [140].

Since the main objective of this thesis is to develop ML solutions for sleep stage classification, discriminative-based models are the main focus in terms of modelling. The rule-based methods (e.g., threshold methods) require the programming of pre-conceived rule sets and exhibit limited flexibility. By contrast, ML methods provide a more flexible alternative to data modelling,

especially when applied to structured tabular data. This section will first introduce two ML models generally recognised in the sleep and health research community, which include support vector machines and random forest models.

Support Vector Machine

The support vector machine (SVM) extends the linear model by introducing kernel tricks. The use of kernel methods with linear models can form non-linear decision boundaries while using convex optimisation for computational convenience [141, 142]. It projects the input data implicitly from a low-dimensional feature space into a higher-dimensional kernel space (often infinite) to reproduce a kernel Hilbert space (RKHS), where a linearly separable hyperplane is estimated [143]. The support vector has several advantages: it is a convex optimisation problem that can be solved by using existing optimisation tools; for small datasets, it is relatively fast to train; the sparseness is reserved in solution representation. However, the performance is highly dependent on the kernel function, and the noise distribution in the training dataset impacts the choice of hyperparameters. SVMs have been used in a wide range of settings in sleep stage classification based on PSG data [144].

Random Forest

Using a series of simple classifiers to divide the feature space in a sequence is an alternative approach to explicitly constraining the decision boundary. Examples of such methods include boosting, which partitions the data in the process of sequence optimisation, which itself weights the data based on the predictions from the previous step before classifying it. The ensemble of multiple classifiers of this type can make the solution more generalisable, an example is the random forest which consists of a bunch of trees and each tree in the forest is trained on a random subset of the training dataset. As discussed in [145], each tree may exhibit low bias and high variance. In the forest, each tree is derived from a subset of the training data to decorrelate their predictions. The variance of the entire forest is much lower than the variance of a single tree, even if there is increased bias, leading to better predictions of unseen data [145]. The advantages of the random forest lie in two aspects. First, given the limited training data, the bagging process randomly splits the training dataset into subsets. Many hypotheses may be equally applicable to the training data, which increases the diversity of the hypotheses. By combining predictions from multiple good and diverse predictors, an ensemble reduces the risk of making wrong assumptions. Secondly, from the computation perspective, each tree in the forest tends to be one of the local optimal solutions. Combining the results of multiple random searches may provide a better approximation of the real unknown function. Random Forest and their extensions have been applied to a wide range of problem settings in sleep medicine research [146, 147].

2.4.5. Deep Learning Based Feature Extraction and Discriminative Methods

The rapid development of applications in the field of computer vision and natural language processing in recent years mainly benefited from the powerful representation learning ability of deep learning models [148, 149]. Traditional handcrafted features may not contain the best discriminative information that is applicable to every task, as the task-specific information may be selected during the feature extraction process. Deep neural networks can learn optimised feature extractors based on objective functions. For example, convolutional neural network (CNN) can extract latent features from images by training a set of kernel filters [150]. For time-series data, the mainstream method is to first use the sliding window method to segment the data into pieces, then use CNNs to learn representations based on the segmented raw data [151, 152]. For example, in sleep stage classification, some researchers have demonstrated the effectiveness of using deep neural network on EEG data [153–156]. While the learned representations can achieve higher performance than the handcrafted features, the disadvantage is that this representation is dataset and task-dependent. In the past decade, deep learning-based methods have attracted a lot of attention in digital medicine research.

Convolutional Neural Network

A neural network normally consists of many small interconnected computing units (neurons) to form a large network that can learn a function to map the inputs and the labels [150]. Among all neural networks, the CNN usually consists of convolutional layers, which consist of a set of convolutional kernels that can extract abstract information related to the task whilst filtering irrelevant information [151]. The convolutional layers are usually inter-weaved with pooling layers to reduce the dimension of data [157]. Depending on the kernel shape, the data passed through layers are normally in multi-dimensional tensors or specifically called feature maps [157].

CNN was broadly adopted in multimedia and time series analysis, such as image recognition [158] and speech recognition [149]. For image recognition, the input data is usually an entire image, and convolutional layers can extract hierarchical abstract representations in the form of feature maps, benefited by the increased feature maps in the first few layers [150]. A poling layer divides each input feature map into regions by strides (e.g., a $m \times n$ shape of stride for 2D feature maps) and calculates a statistical value for each region as the output. A pooling layer is applied after one or more convolutional layers to reduce the spatial size of the feature maps. The final output of a set of convolution and pooling operations is a sufficient small feature vector which is often flattened or pooled into one single vector. One or more fully connected layers take feature vectors as input to perform the classification or regression tasks.

Convolutional Neural Network For Time Series Data Analysis

CNN can be used for physiological signals in two ways. The first way is to input the raw signals directly to convolutional networks, as the network will extract hierarchical representations that

are relevant to the tasks. The benefits of CNNs used on time series data include shift-invariance and local correlation (or permutation invariant) properties [159, 150]. For image recognition, the shift-invariant or equivariant means the network is somewhat resistant to the location of the object of interest in an image [157]. Such property is also useful when using EEG signals for sleep stage classification. Certain signatures of the sleep stage (e.g., sleep spindle) could be recognised in the EEG signal, and a limited shift along the time axis should not significantly undermine the prediction performance [157]. As the same convolutional kernel scans all sections of the signal. Local correlation enforces a sparse local connectivity pattern between neurons of adjacent layers, so the network can discover the concept of topology via learning and the models that are trained to parameterise this spatial relationships [151]. However, this method normally requires abundant samples to learn robust and generalised representations.

The second method is more effective when the training dataset is significantly smaller, where the CNN may not be able to extract robust latent features. The raw signal data can be segmented by the sliding window method into multiple pieces. For each segment (a data frame), the first way is to extract the spectrogram, filter banks or handcraft features as the inputs of CNN. In this case, the input is treated either as an image (e.g., a time and frequency feature map) or a summary of time series physiological signals. For the feature map representations, a 2D convolutional neural network can be used to learn the latent features. For the handcraft features, the order of features can be combined in various ways, a 1-D convolution can be used for each feature to avoid the combination dependency. This section will only explain the basic 1-dimensional convolutional functions; extra operations such as striding and padding will not be covered here.

Suppose a time series feature vector $\mathbf{v} \in \mathbb{R}^T$ where T denotes the temporal steps. We then conduct convolutions on it using linear filters. A filter vector can be denoted as the weight vector $\mathbf{w} \in \mathbb{R}^L$ with the length L. The adjacent temporal feature from i to j can be denoted $\mathbf{v}_{i:j}$. The convolution operation $*$ between $\mathbf{v}$ and $\mathbf{w}$ results in the output vector $\mathbf{o} \in \mathbb{R}^{T-L+1}$ where

$$o_i = (\mathbf{v} * \mathbf{w})_i = \sum_{l=1}^{L} (\mathbf{v}_{i:(i+L-1)} \odot \mathbf{w}) \tag{2.1}$$

In Equation 2.1, $\odot$ denotes the element-wise multiplication. After the convolution operation, an activation function σ will normally apply to each o_i to get the new feature a_i which can be denoted as:

$$a_i = \sigma(o_i + b) \tag{2.2}$$

Here, $b \in R$ is a bias term. The activation function can be a variety forms, such as ReLU [160] or sigmoid [150].

Recurrent Neural Network and Long-short Term Memory

Although the convolutional neural network has been successfully applied in several time series tasks [161, 153, 162], the lack of inferring long-term temporal dependencies is a dispensable disadvantage in time series analysis. recurrent neural network (RNN) is specifically designed to

model such a prior, in which the hidden layers not only depend on the inputs of the current time step but also depend on the outputs of the hidden layers of a previous time step [163]. Recurrent neural networks using nonlinear functions in the hidden layers often suffer the gradient vanishing or gradient explosion phenomenon during the training process, especially when the time steps become too long [164]. This is due to the gradient of the RNN's loss function with respect to the model parameters calculated using back-propagation through time (BPTT) [164]. The error is propagated backwards through time, it is repeatedly multiplied by the hidden layer's weight matrix. During this process, the spectral radius (e.g., the maximum absolute value of its eigenvalues) determines the multiplication outputs. If the spectral radius is below one, then the error will varnish [164]. If it is greater than one then the error will "explode". This made the training process difficult and unstable.

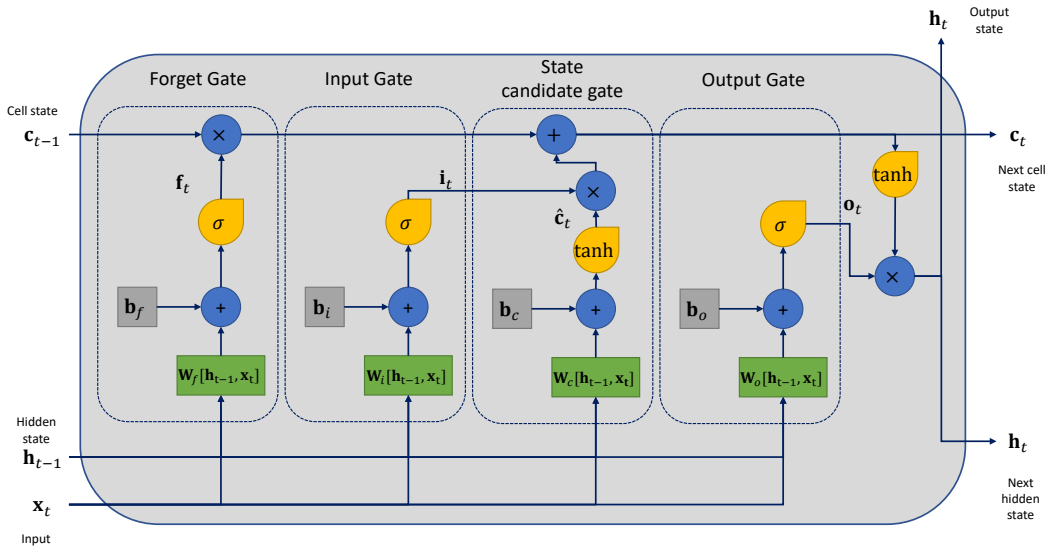


Figure 2.8 The structure of the LSTM neural network [2].

The gradient explosion problem can be solved by applying gradient clipping, a threshold method that forces the gradient to be within a certain range [165]. However, the gradient explosion can be elevated by designing "memory" cells to preserve information over a long time. Among all the derivations of recurrent neural networks, the long short-term memory (LSTM) is the one that could effectively reduce the vanishing gradient problem by designing an advanced neural network structure, which has been successfully used in natural language processing tasks [166]. The network is capable of selectively reserving information via the input gate, forget gate and output gate as described in Fig. 2.8. The cell state takes the information calculated from the current inputs $\hat{\mathbf{c}}_t^{(i)}$ and the previous cell state $\mathbf{c}_{t-1}^{(i)}$. The input gate and forget gate learn to select the information based on the current inputs and previously hidden outputs.

The detailed structure of the LSTM cell is described by the following equations:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \tag{2.3}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \tag{2.4}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \tag{2.5}$$

$$\hat{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \tag{2.6}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t \tag{2.7}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{2.8}$$

where σ is the sigmoid activation function, the $\odot$ stands for the element-wise multiplication. The input gate $\mathbf{i}$, forget gate $\mathbf{f}$, and output gate $\mathbf{o}$ along with the sigmoid function to produce values between 0 and 1. $\hat{\mathbf{c}}_t$ and $\mathbf{h}_t$ stand for the candidate cell state and the hidden state respectively. $\mathbf{W}$ and $\mathbf{b}$ are the weight matrices and the bias vectors in a gate. The LSTM is successfully applied in the PSG signal-based sleep stage classification. Huy et al. [153] demonstrated that using multimodal data with LSTM models can improve five sleep stage classification performance.

2.4.6. Variational Autoencoder Model And Disentangled Representation Learning

Many existing end-to-end deep learning models can learn a highly discriminated representation in a supervised manner, as the deep neural networks generate such representations at every layer to maximum a posterior (MAP) of $p(y|x)$. Typically, no additional restrictions are placed on the latent feature learning process, so the learnt features are usually more useful for tasks specified on the training data set but may be performing poorly on different datasets or tasks, that is, poor generalisation. A practical approach is to employ disentanglement learning methods to separate the feature space and only retain more generalised features for downstream tasks [167]. To achieve this objective, models based on variational autoencoder (VAE) have attracted the attention of researchers. Unlike the CNN and LSTM, which learn a direct mapping of $y = f(x)$, these variational framework-based models sample features from known distributions, which helps encourage richer representations through unsupervised learning [168]. They often assume that there is a "bottleneck" representation layer. Autoencoders are one of the prevalent models in unsupervised learning. They are deep neural networks which consist of two components, an encoder and a decoder. The encoder learns a function f to map the input data into a latent feature $\mathbf{z}$ in the bottleneck representation layer $\mathbf{z} = f(\mathbf{x})$, while the decoder learns a function g to reverse the representation $\mathbf{z}$ to the input space $\hat{\mathbf{x}} = g(\mathbf{z})$. The autoencoder learns a mapping that can encode the input data into a low dimensional manifold [169]. Such representation could retain high-level abstraction while ignoring the nuisance factors [170]. By extending autoencoders to variational autoencoders (VAE), instead of using the deterministic autoencoder that encodes the input data into a latent instance (a vector or a matrix), the VAE could map it as a distribution over the latent space to incorporate variations of the input data. VAE framework consists of two networks, a probabilistic encoder and a probabilistic decoder. The encoder maps samples from the data distribution to the latent variables, $\mathbf{z} \sim p(\mathbf{z}|\mathbf{x})$, while the decoder maps the prior latent distribution

$\mathbf{z} \sim p(\mathbf{z})$ to samples of the data distribution $\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})$. VAEs are probabilistic model which assumes the data are generated by sampling a likelihood $\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})$ from unknown distribution of latent factors $\mathbf{z} \sim p(\mathbf{z})$ [171]. The input data has been embedded into a smooth manifold of latent variables. The VAE framework is capable of learning to disentangle representation through its inherent properties. However, the posterior $p(\mathbf{z}|\mathbf{x})$ is intractable because of high-dimensional data or because of complex forms of distributions. Therefore, the variational methods are employed using the stochastic encoder $q(\mathbf{z}|\mathbf{x})$ to approximate the true posterior, which could converge to a Gaussian prior distribution $\mathbf{z} \sim N(0,I)$ with a diagonal covariance matrix that each dimension is factorised or independent [170]. The factorisation is achieved when the factors of variation are aligned with the axes of the Gaussian posterior [171]. The framework also provides the opportunity to add additional constraints on latent representation space. A disentangled representation can be considered as the changes of a specific latent dimension which will only influence the changes of a single factor in the generated data [171]. The recent works demonstrate that using the disentanglement learning methods with the adversarial training, the distribution-specific (or domain specific) factors or attributes can be removed to some extent during the training process [172–174]. It has been suggested that generative learning factors which were disentangled can be useful for a large variety of tasks and domains [173]. A disentangled representation could boost the performance of state-of-the-art (SOTA) machine learning approaches in which the models are still struggling but where humans excel [175]. These scenarios may require knowledge transfer, where reasoning new data can be facilitated by recombining learned factors. Many applications in computer vision tasks (e.g., human pose representation learning [176], and face recognition [177], etc.) have demonstrated that the disentangled representation can achieve higher performance.

2.5. Multimodal Fusion

Data collected from different modalities to represent diverse physiological information may have varying predictive power and noise topology [1]. The way these signals are integrated could have significant implications for downstream tasks. Fuster-Garcia et al., tested two actigraphy raw data fusion methods to perform non-linear regression of signals using artificial neural networks [178]. The first fusion method proposed in their study is a centralised architecture assuming all signals may be present in a common system, the second fusion method allows a distributed fusion of the signals [178]. In both methods, the signal with the highest quality (e.g., its completeness, sensitivity to patient motion or signal-to-noise ratio) is used as the reference signal, and the rest of the signals from the other devices are transformed to the representation space before linearly combining them with the reference model. As a result, a single-modality signal with quasi-linear decrements of error with respect to the number of input signals is generated [178]. In practice, multi-sensor fusion methods at the raw data level can compensate for acquisition errors and are more tolerant of errors and missing data

From an information-sharing perspective, complementary fusion means that the modalities do not directly depend on each other but instead combine the outputs, leading to collective

measurement estimates. Some examples of this would be combining pressure sensor data, tri-axial accelerometer data and plethysmography (PPG) data for monitoring sleep quality [56]. These complementary fusion protocols can significantly improve the classification performance of sleep stages [74, 56].

Multimodal fusion in machine learning has been extensively studied in pattern recognition applications, such as in image and video captioning [179], visual question answering [180], audio-visual speech recognition [149] and emotion recognition [181]. In the field of ubiquitous computing, multimodal fusion has also been adopted for human activity recognition [182], sleep stage classification [5], fatigue assessment [183] and person identification [184].

These studies demonstrated that multimodal representation learning could extract valuable information from complementary data sources for classification tasks. Traditional fusion strategies include feature level fusion (e.g., [185]), score-level fusion (e.g., [186]) and decision-level fusion (e.g., [187]). In the end-to-end DL era, the boundary between multimodal representation and fusion has been blurred. Representation learning is interlaced with classification (or regression) objectives. Nevertheless, the fusion strategy for DL models may still be carried out in three stages, such as early fusion, late fusion and hybrid fusion [188].

Fusion at different stages may influence the results of representation learning. For example, early and late fusion may inhibit intra-modal or inter-modal interaction [188]. Neverova et al. noted that highly correlated modalities should be fused together [189]. Hazirbas et al. demonstrated that the performance of fusion is highly affected by choice of which layer to fuse [190].

Based on the complexity of fusion methods, the operation can be divided into three types: simple operations, attention-based methods and tensor-based methods [191]. For feature vectors from different modalities, concatenation and addition are two commonly used simple operations [188]. The simple concatenation method was also commonly adopted to combine the raw inputs or combine the representations obtained from the pre-trained model of each modality [192]. Other researchers have explored more advanced fusion methods, such as the attention-based fusion scheme for human activity recognition [182]; the attention mechanism is widely used for multimodal fusion. This usually refers to dynamically calculating a weight vector for each time step (or spatial position) and weighting a set of feature vectors [162].

In the case of tensor-based methods, more advanced tensor-based fusion only demonstrated the usefulness of fusing image and text-based tasks, such as bilinear pooling, which is a method of fusing two unimodal representations to a joint presentation by calculating their outer product. This method can capture the multiplicative interaction between all elements in two vectors [193]. But this method may result in a large number of model parameters, and the inference time is often questionable when they are deployed on wearable and mobile devices.

In the case of sleep monitoring using the modalities that may be derived from wearable sensors, several previous works (e.g., [194]) have achieved promising results for sleep stage classification by concatenating multimodal intermediate features and feeding them into DL models. However, these studies focus on the choice of modalities rather than the fusion techniques.

Different modalities may contain complementary information. The way to fuse heterogeneous intermediate features is worthy of exploration. In terms of movement sensing and cardiac sensing, they are different in signal-to-noise ratio, data generation process and measurement frequency. Moreover, the activity count is better in sleep/wakefulness classification, but it is difficult to discern different sleep stages [50]. For healthy adults, the difference in heart rate variability between REM sleep and wake is less than the difference in NREM and REM sleep [52].

2.6. Data Visualisation and Explainability

One of the challenges of using deep learning models on health or life-critical systems is the lack of transparency or incomprehension by humans regarding the model decision processes. When intelligent systems fail, they often give incorrect results or stop working without warning or explanation, leaving users staring at the incoherent output and wondering why the system is doing it [3]. Visualisation and interpretation of model decision-making processes are critical to building intelligent systems that humans can trust and integrate meaningfully into their daily lives. It is clear that one must build "transparent" models that explain the reasons behind mode prediction [3].

Figure 2.9 The three images to the right are heat maps generated by Grad-CAM based on the dense captioning model. [3].

Humans have varying levels of understanding regarding different types of data. For example, images and text represent the most common and natural forms of communication that people use in daily life [195]. Abstract data, such as large data tables or Excel tables of handcraft features (for example, heart rate variability features) organised in time series format, are less intuitive for users to understand [195].

Visualisations can use visual objects to represent abstract data in point, line, and bar formats that are easier for humans to understand and interpret. Graphical representation relies on human high-throughput visual perception channels and the ability to connect data representations to human knowledge and expertise. This makes the time series data more intuitive to understand, such as identifying repeating patterns and trends [196]. Sleep data is usually presented in a time series format, usually visualised using a line chart. The horizontal x-axis represents time. Raw signal visualisation is mainly meaningful for domain experts to interpret complex patterns. For example, PSG recordings mostly contain the time series data that recorded the electrical changes by electrodes, which reflects certain physiological activities (e.g., saturation of peripheral oxygen

(SpO2), breath, cardiac activities, eye movements, muscle movements and neuron activities in the brain). Specific patterns can be automatically detected by algorithms or highlighted by experts on the chart, such as the sleep spindles appearing in the EEG waves [99].

CNN has achieved unprecedented breakthroughs in various computer vision tasks, so the transparency of the decision-making process has attracted the attention of many researchers. A saliency mask (SM) based method can unveil where a CNN looks into a time series data for recognising their predictions. A representative method is gradient-weighted class activation mapping (Grad-CAM) which uses the gradients of any target class in a classification neural network flowing into the final convolutional layer to produce a coarse heat map that highlights the important regions in the input data for predicting the class [3]. An example of Grad-CAM based visualisation can be seen in Figure 2.9. The highlighted area is highly consistent with the area marked by the bounding box, even though the captioning model and Grad-CAM technology do not use any bounding box annotations. A CAM for a specific label is first used to calculate the gradients with respect to the final convolutional layer's feature map activations. Then the method calculates the average value for each activation unit over time steps to build up a weighted vector which is considered as the "importance" score. Afterwards, the algorithm interpolates these weights into a full scale of the input data dimension to generate a heatmap [3]. For time series data with a CNN, the method is defined as follows:

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial \mathbf{y}^c}{\partial \mathbf{A}_{ij}^k} \tag{2.9}$$

where Z is the number of time steps times the number of features; $\mathbf{y}^c$ is a one-hot vector representing the c-th class; α_k^c is a weighted scalar with respect to each activation unit of the last convolutional layer; $\mathbf{A}_{ij}^k$ represents the i-th time index, the j-th input signal (or channel) and the k-th activation of feature map. After obtaining the weighted vector, an activation function is applied to obtain the forward activation maps, followed by a rectified linear unit (ReLU) function, which is denoted as follows:

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c \mathbf{A}^k) \tag{2.10}$$

A ReLU function is applied to the linear combination of maps, as this chapter is only interested in the features that have a positive impact on the class of interest, that is, on certain time steps, the inputs of handcraft features whose value should be increased in order to increase $\mathbf{y}^c$. The negative values are likely to belong to other classes. For example, the inputs of physiological signals (e.g. Heart Rate, LF/HF ratio) on time steps should be increased in order to increase the probability of the predicted sleep stage. Grad-CAM can provide a local explanation at the instance level by invoking the backpropagation with respect to classes and/or the activation, which generates aesthetically pleasing and heuristic explanations of a time series saliency map.

In addition to Grad-CAM-based decision process visualisation tools, many traditional machine learning methods can also produce feature importance scores. The main disadvantage of these methods is that these scores usually work at the task level rather than the class level. In [197], they proposed the shapley additive explannations (SHAP) model that produces class-specific feature importance scores in classification tasks. This is of significant importance in health machine learning applications. In addition, it could also provide instance-level explanations. SHAP is a unified framework for value estimation of additive feature attributes that can be generalised to many models by producing SHAP values. SHAP values make it possible to explain the output of a function f as the sum of the effects ϕ_i of each individual feature that is introduced to a conditional expectation [197]. The SHAP function is designed to assess the impact of missing features on model predictions. Suppose we have a function f and a mapping function h_x that maps the binary pattern of missing features represented by $\mathbf{z}'$ to the input space of the original function. Given such a mapping, we can evaluate $f(h_x(\mathbf{z}'))$ to calculate the effect of when a specific feature is presented or not. To obtain SHAP values, the function of $f_x(\mathbf{z}') = f(h_x(\mathbf{z}')) = \mathbf{E}([f(\mathbf{z})|\mathbf{z}_S])$ is the expected value calculated based on the missing values for features not in the set S [198], where S is the set of non-zero indexes in $\mathbf{z}'$. And $\mathbf{z}' \subseteq \mathbf{x}'$ represents all $\mathbf{z}'$ vectors where non-zero entries are a subset of the non-zero entries in $\mathbf{x}'$, which is a simplified input [198]. SHAP values are calculated based on combining these conditional expectations with Shapley values from game theory to attribute ϕ_i values to each feature [198]. The Shapley values are denoted as:

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!\,(|M|-|S|-1)!}{|M|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{2.11}$$

where M is the number of simplified input features [198]. By using Shapley, SHAP returns individual contributions to the full feature set, which allows us to understand individual feature contributions not only for a particular task but also for each class. Figure 2.10 shows an example use of SHAP on a tree-based model to investigate the top ten most important features with respect to the behavioural changes induced by seasonal flu. The SHAP method can calculate feature importance scores based on handcrafted features and traditional machine learning models with acceptable computational time. For time series data, it has to calculate the perturbation of the feature combination to summarise the feature importance with respect to time steps. Therefore, it could be very slow when applied to time series data with deep learning models.

Visualising the signal segments that matter to decision-making is imperative for researchers and clinicians. In this thesis, sleep monitoring uses wearable sensing data, including the ECG data that reflects the ANS activities which are partially influenced by the brain activities during sleep. So cardiac sensing and movement sensing may contain useful information to discern all the sleep stages and there are no established rules for using these signals to distinguish sleep stages. This poses significant challenges for visualisations involving the model decisions.

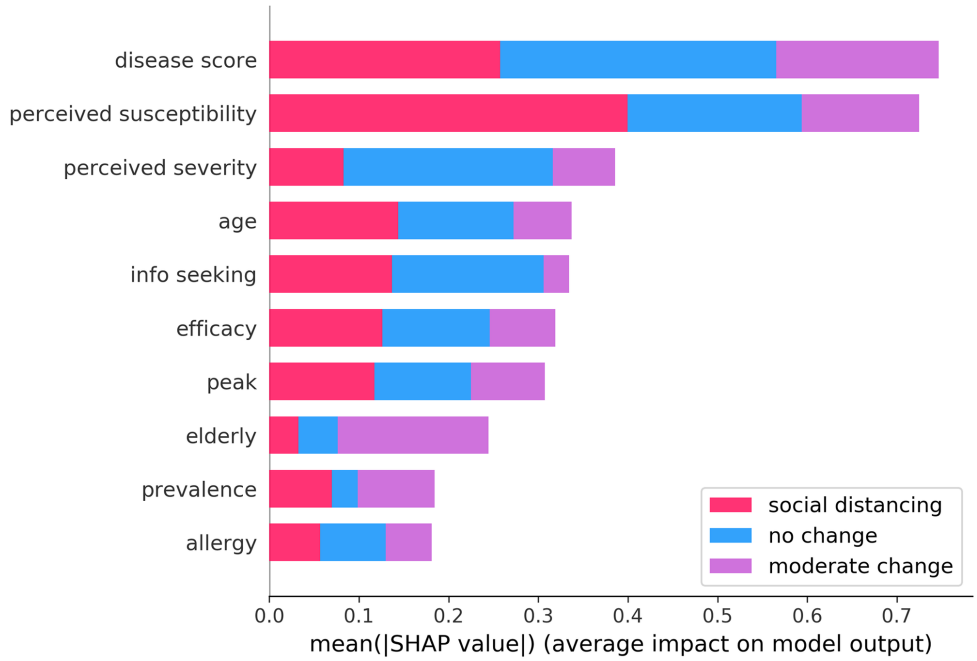


Figure 2.10 An example of using SHAP to interpret feature importance with respect to the behavioral changes induced by the seasonal flu [4].

2.7. Evaluation Metrics

Evaluating the performance of the sleep stage classification pipeline is crucial to its design, as the different components are selected from a large set of possible combinations guided by performance. The way of modelling sleep stage classification tasks is therefore tightly linked to the choice of evaluation metric. Most of the performance metrics listed here have been adopted in sleep stage classification using PSG signals. The most common performance metric in sleep stage classification is the overall accuracy, i.e. the fraction of correctly classified instances.

To assess class imbalance and evaluate performance, several popular metrics based on True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) classifications should also be adopted. These evaluation metrics can be summarised as follows:

- **Accuracy** counts the number of correctly classified sleep epochs, normalised over the total number of sleep epochs ($\text{Acc} = \frac{TP+TN}{TP+TN+FP+FN}$).
- **Recall** measures the proportion of positives that are correctly identified as the given stage ($R = \frac{TP}{TP+FN}$).
- **Specificity**, also known as true negative rate, measures the proportion of negatives that are correctly identified as the given stage ($S = \frac{TN}{FP+TN}$).
- **Precision** is the fraction of correctly classified instances among the overall positive predictions ($P = \frac{TP}{TP+FP}$).
- **$\mathbf{F_1}$ score ($\mathbf{F_1}$)** conveys the balance, with the harmonic mean, between precision and recall ($F_1 = 2 \times \frac{\text{P} \times \text{R}}{P+R}$).

- **Cohen's Kappa** (κ) measures inter-rater reliability/agreement, comparing observed accuracy with an expected accuracy ($\kappa = \frac{P_o - P_e}{1 - P_e}$, where P_o the observed proportional agreement and and P_e the expected proportion of agreement). In this context, Cohen's κ factors out the agreement by chance arising from the class imbalance of different sleep stages throughout the night.

Due to the sleep stage class imbalance issues, these performance metrics should be calculated in a class-wise manner and reported as the mean values. And the two-tailed t-tests can be used to calculate statistical significance with respect to model performance improvement.

Many sleep classification studies used accuracy or F_1 to measure their model's performance, yet these are high-level metrics which do not consider class-wise performance. Confusion matrices, on the other hand, provide class-wise predictions and corresponding error types. However, for clinicians and other health practitioners, these matrices are not the most obvious way to represent the time deviation of sleep stages, as they include too many low-level details.

2.8. Summary

This chapter investigated different sensing approaches and their advantages and limitations when used for sleep stage monitoring outside the laboratory. Several classical machine learning methods and deep learning models were introduced for the supervised learning tasks regarding the different granularity of the sleep stage classification tasks. The sensing approach that is most suitable for a free-living environment is body-worn sensors, especially wearable bands and smartwatches as they are inexpensive compared to research-grade wearables and are generally available to large-scale studies.

The data collected from wearable sensors are typically processed in a pipeline approach, with the components tuned to extract the most discriminative information that benefits the sleep stage classification tasks. The design of feature extraction and classification methods both rely on a large amount of available wearable data and gold ground truth annotations, which remains a major challenge for building a robust machine learning model. The availability level of data granularity determines the signal processing pipeline. For instance, many consumer-grade wearable devices do not provide raw signal access to all their sensors, so the use of clinical features (e.g., heart rate) would be realistic solution in this case. The empirical studies in chapter 3 demonstrate the feasibility of designing efficient data processing pipelines to classify sleep stages at different levels of granularity. To alleviate the model generalisation issues when using these handcrafted features, in chapter 4, a VAE-based disentangled representation learning model is proposed to learn invariant features with respect to personal attributes, which can achieve better performance on unseen populations for three-stage sleep classification. Moreover, the multimodal data poses inherent challenges for learning effective representation. In chapter 5, a set of proposed multimodal fusion approaches demonstrate that advanced fusion techniques can result in significant improvement in three-stage sleep classification. Moreover, this chapter will explore the use of Grad-CAM on cardiac and movement-sensing data to make the decision-making process transparent for deep learning.

Chapter 3. Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing

3.1. Introduction

Traditionally, human sleep has been monitored in laboratory settings using polysomnography (PSG) which has been described in chapter 2. Whilst PSG is considered the gold-standard for sleep monitoring, as a result of the need for sensing equipment, its use is limited to laboratory settings and typically to just one or two nights. These single nights of observed sleep in an unfamiliar environment may not reflect normal sleep. Further, it is impractical to measure sleep using this method for more than two consecutive nights as it is burdensome to patients or study participants. PSG is also expensive and requires expert set-up and analysis. For these reasons, efforts to monitor individuals' typical sleep duration and quality longitudinally in large, free-living populations have generally relied upon sleep diaries or self-reported questionnaire data. Whilst sleep diaries are cost-effective, scalable and able to collect information regarding typical sleep patterns, there are concerns as to the validity and reliability of participant responses [199]. Wearable sensors offer a potential solution. Such sensors provide valuable, unobtrusive tools through which to objectively monitor physical activity in large population studies, with potential applications for sleep monitoring.

Conventional approaches to monitoring sleep using wearable devices are primarily based on actigraphy (count-based movement information) and accelerometry (raw, high frequency data which is often in tri-axial) [200–203]. However, recent technological and battery life advances increasingly facilitate multimodal sensing (e.g., combining accelerometry with HR sensing). Multimodal sensing facilitates more intricate human activity recognition (HAR) tasks and has shown promise for sleep-stage classification [204]. The validity of actigraphy for the classification of sleep-wake transitions has been demonstrated over the past three decades [200, 203, 201, 202]. Algorithms applied to actigraphy for this purpose exploit differences in body movement between wakefulness and sleep. Recent work has demonstrated how different methods for binary sleep-wake classification using actigraphy compare when applied to the same, standardised dataset [205]. Furthermore, HRV metrics could be valuable for multistage classification as autonomic function fluctuations occur between non-REM sleep and Wake/REM sleep, whilst these same functions are consistent when comparing Wake to REM [206, 207, 136, 208].

Understanding time spent in different sleep stages (beyond binary sleep-wake classification) in free-living environments has important implications for commercial applications, as well as for research. For example, accurate sleep architecture inferences may provide better information to guide sleep-related behavioural changes and recommendations [209]. PSG is the gold standard

for sleep stage assessment, but it is not scalable to large, population-based studies of free-living individuals with the power to make inferences regarding the implications of sleep for health and illness. Wearable devices are a potentially inexpensive and scalable solution to monitor sleep in large populations. Limited literature exists regarding the performance of multimodal sensing using wearable technologies in sleep-wake and sleep-stage classification [61]. Moreover, the code and dataset for most of these studies are not publicly available, and these studies are usually conducted on small datasets. This becomes an obstacle to reproducing and improving sleep monitoring using wearable computing devices. The development of a set of benchmarks to evaluate the performance of sleep-wake and sleep-stage classification methods on multimodal data using movement and cardiac sensing would address a major gap in the existing literature.

In order to address this gap, this chapter focused on five major contributions:

1. This chapter introduces a framework for pre-processing and analysing multimodal sensor data from movement (actigraphy) and cardiac (RR intervals from ECG) sensors. These sensor data are derivable from research-grade ECG or photoplethysmogram (PPG) devices.

2. This chapter systematically compares single modality to combined sensing (actigraphy + HR/HRV) approaches for classifying sleep-wake using different machine learning models.

3. This chapter extends this systematic comparison to explore the performance of single modality approaches and combined sensors across three different multistage classification tasks: (A) Conventional three-stage classification (NREM, REM, Wake), (B) Four-stage classification (light sleep, deep sleep, REM and Wake) and (C) Five-stage classification (AASM-standard), also using traditional machine learning and deep learning models.

4. This chapter introduces an easy-to-interpret evaluation metric, namely, time deviation, which aims to be accessible to sleep practitioners. This chapter also studies the modality/feature importance by using Random Forest with SHAP, yielding some interesting findings (e.g., high frequency HRV is the most important feature in recognising REM sleep).

5. This chapter introduces an ensemble structure for multistage classification of sleep based on multi-timescale and multimodal DL ensembles. This architecture aims to exploit the individual contributions and strengths of different classifiers. This approach can significantly improve the performance of the three-stage sleep classification task.

This chapter presents a systematic multimodal and multistage evaluation of sleep-wake cycles and the sleep stages in a large, diverse population. Each individual method and modality is experimentally explored, and approaches to improve classification performance through modality fusion are explored. Additionally, feature importance for different classification tasks is investigated, leading to a deeper understanding of the physiological underpinnings of each model.

3.2. Related work

Since the 1980s, a vast number of studies have explored new methods and techniques to infer sleep-wake cycles using actigraphy with either single-axial [202, 200, 203] or, more recently, tri-axial accelerometry [77]. While these methods have proven valuable, they were often derived in small cohorts or by using non-clinical grade equipment or sleep diaries. The recent availability of large datasets, provided by initiatives such as the National Sleep Research Resource [210, 211][1], makes it possible for researchers to create large standardised benchmarks. For example, Palotti et al. leveraged one of the available datasets, the Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Study[2], to compare the performance of the most relevant heuristic approaches and ML methods for binary sleep-wake classification [205]. Whilst novel, their work was limited by: (1) exclusively comparing methods for sleep-wake classification rather than multistage classification; (2) only using actigraphy data. This chapter addresses these limitations in the MESA Sleep Study dataset, the largest dataset suitable for such experiments until the year 2022.

Beyond actigraphy data, this chapter also investigated the use of HR and HRV data. HR can be defined as the average number of heartbeats per minute, while HRV is a measure of the variability in beat-to-beat intervals, known as RR intervals. The denoised and filtered RR interval includes the normal R peak, often referred to as the Normal-to-Normal interval (NNI). These measurements are powerful biomarkers that have been used to understand training and recovery, address chronic disease and monitor stress and sleep [127–129]. Several studies demonstrated the HRV features are different during sleep stages in nocturnal sleep as previously discussed in chapter 2 [127–132]. Conversely, several studies have shown that HR does not change significantly between sleep stages, although some work has suggested a rise during REM sleep [212, 213]. Hence, it is worth investigating whether sleep stage classification can be performed on the comprehensive HRV features [214]. Recently, Radha et al. have reported that HRV has great potential to classify sleep stages [215]. However, their work was performed on a private dataset and conducted using some features that are often not present on wearable devices. In this chapter, HR/HRV features are extracted from research-grade wearable sensors and their performance is evaluated on the largest public dataset.

Aside from ML and DL models, ensemble architectures are becoming increasingly prevalent for HAR tasks. For instance, in 2015, Single et al. adopted an approach consisting of three variants of long-short term memory (LSTM) networks that worked in parallel to tackle a biological sequence analysis task and then used *majority voting* to decide upon the final classification prediction [216]. In [186], Guan and Ploetz developed an LSTM ensemble model via epoch-wise bagging for efficient training. They injected several random factors to increase the diversity of the classifiers and improve performance in several HAR tasks. Recent work has explored the application of ensemble models for automatic sleep stage classification using PSG/EEG signals. These studies have shown promising results, improving the performance of shallow ML and even DL approaches [217–219]. Koley et al. used an ML ensemble architecture approach

[1] https://sleepdata.org
[2] https://sleepdata.org/datasets/mesa

consisting of five binary support vector machine (SVM) classifiers to classify different sleep stages [217]. Using a *"winner-takes-all"* ensemble method [220] the researchers managed to extract more discriminant patterns from EEG. Recently, [153] applied an ensemble method to multimodal PSG data (EOG, EEG and EMG) by fusing classifiers. All these previously reported models are based on sleep epoch (30 seconds) level feature extraction protocols and use classifier ensembles derived in sensors which often exceeded 100Hz sampling rates (EEG data, etc). These methods demand high specifications with regard to computing power. Currently, the limited data storage and processing capabilities of wearable devices mean that using methods based on a high sampling rate is unlikely to be possible in long-term free-living environments.

In summary, the development of sleep stage monitoring methods based on wearable technology is still stagnant in traditional ML methods due to a) the lack of large-scale open-source multimodal datasets, b) closed-source pipelines for data preprocessing and modelling, and c) non-standard evaluation criteria. Moreover, we don't know which sleep stages can be realistically detected from the cardiac and movement-sensing data. These problems hindered the development of sleep monitoring using ubiquitous computing techniques. The work in this chapter is considered the first systematic study of sleep stage classification using large-scale sensing data that may be available from wearable devices

3.3. Methods

The MESA Sleep dataset is introduced and described in the first part of this section [210, 211]. All experiments reported here were conducted based on this dataset. Section 3.3.2 provides an overview of the data pre-processing and feature extraction method for modalities which consist of cardiac sensing (HR and HRV) and movement sensing (actigraphy). All tasks explored, including the *ensemble method*, are introduced in Section 3.3.3. In Section 3.3.4, the models used for the benchmark study are presented. Section 3.3.5 describes how these experiments were designed. Finally, in Section 3.3.6, the metrics used to evaluate the classification models are discussed.

3.3.1. Dataset Description

The Multi-Ethnic Study of Atherosclerosis (MESA) dataset is a multi-centre longitudinal study designed to investigate the characteristics of sub-clinical cardiac disease. The study comprises 6814 asymptomatic men and women of black, white, Hispanic and Chinese-American ethnicity, of which 2,237 were also enrolled in the MESA Sleep Study. As part of the MESA Sleep Study, all participants wore an actigraphy device for one week and underwent concurrent PSG for one night. Data for this study was acquired in six different centres across the US and followed the appropriate Institutional Review Board approvals and written informed consent for participant data acquisition [210, 211].

The MESA Sleep Study was conducted using a Compumedics Somte System for PSG, which includes the ECG signals here used to derive HR and HRV and their associated features,

Dataset	Total	Female	Male	Black	Chinese-American	White	Hispanic	Age ($\mu \pm \sigma$)	Min Age	Max Age
Training	1395	752 (54%)	643 (46%)	383 (28%)	153 (11%)	511 (37%)	348 (25%)	69.29 ± 8.73	54	94
Test	348	198 (57%)	150 (43%)	103 (30%)	39 (11%)	128 (37%)	78 (22%)	68.52 ± 9.21	55	89

*Numbers are N, N(%) or mean (SD). Age is given in years.

Table 3.1 Breakdown of population based on sex, age and demographic characteristics, by dataset (training or test).

Dataset	Total Sleep Time (TST)	Total Time in Bed (TIB)	Sleep Efficiency (%)	Wake After Sleep Onset (WASO)	N1	N2	N3	REM
All	359.0±80.7	475.0±85.3	76±13.1	90.3±62.7	49.42±30.9	207.0±60.1	40.0±33.4	66.9±28.9
Training	357.5±80.2	473.5±84.9	75.9±13.1	90.5±62.7	49.3±31.1	206.6±60.4	39.6±33.3	66.6±29.3
Test	365.1±82.5	480.9±87.0	76.4±12.8	89.2±62.6	49.8±30.4	208.5±58.9	41.68±33.7	68.3±27.4

*Numbers are minutes except sleep efficiency measured in percentage(mean ± SD)

Table 3.2 Sleep statistics of participants in the study.

alongside an Actiwatch Spectrum from Philips Respironics to record actigraphy data. This device captures measurements of movements defined as "activity counts"[3] and aggregates them into 30-second epochs. The Actiwatch was securely fastened to the participant's non-dominant wrist. These actigraphy signals and their associated features can be derived from most research-grade wearable devices. The sensors for the Compumedics PSG comprised: cortical EEG, bilateral EOG, chin EMG, abdominal and thoracic respiratory inductance plethysmography, airflow, ECG, leg movement sensor and finger pulse oximetry. These sensors collected three types of signals: bioelectrical potentials (EEG, EOG, EMG, ECG), waveforms received from transducers (thermistors on the airflow devices, inductance respiratory bands, piezo leg sensors and position sensors from the leg device) and auxiliary devices (oximetry measures of oxyhemoglobin saturation and nasal pressure records). Full details of the setup, protocol and sampling rates are available [4,5]. All participants included in the study had at least one full night of PSG recording with concurrent actigraphy and ECG. An illustration of the experimental set-up is provided in Figure 3.1. All nocturnal recordings were transmitted to a centralized reading centre at the Brigham and Women's Hospital (Boston, MA, USA), and data were scored by trained technicians using AASM guidelines. For the training labels, the expert scoring and epoch staging annotations on PSG data are provided by Bild et al. [221]. Note that the MESA Sleep dataset is the **only large open-access dataset** combining gold-standard measures of sleep through PSG with wearable sensor data from actigraphy as well as ECG (HR/HRV) and thus the only existing dataset appropriate for the purposes.

Table 3.1 summarises the main demographic characteristics of the participants by training and test splits.

[3]https://www.salusa.se/Filer/Produktinfo/Aktivitet/TheActiwatchUserManualV7.2.pdf

[4]https://sleepdata.org/datasets/mesa/pages/equipment/montage-and-sampling-rate-information.md

[5]https://sleepdata.org/datasets/mesa/files/documentation

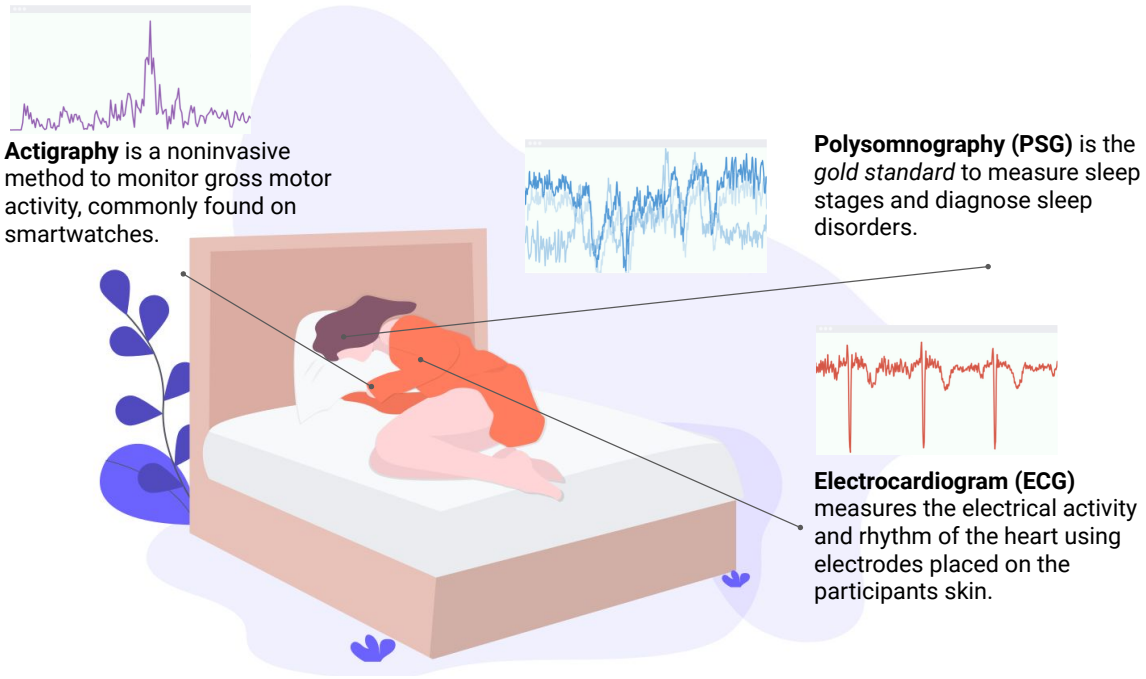


Figure 3.1 Experimental setup and tasks: the models are trained using a combined-sensing, multimodal approach which incorporates two time-series signals: actigraphy and ECG-derived HR and uses Gold-Standard PSG labels for training

3.3.2. Data Pre-processing and Feature Extraction

In this chapter, PSG, ECG and actigraphy records are synchronised into 30-second sleep epochs for 1,743 of the 2,237 participants included in the study. A total of 494 participants were excluded on the basis of: (1) lack of concurrent PSG, ECG and actigraphy data; (2) lack of enough quality standard data ($<$ 1.5 h of usable data from the concurrent three sensing methods); or (3) lack of data integrity or misalignment of data, the actigraphy outlier epochs have been removed based on human expert annotations. These outliers are either non-wearing periods or equipment failure periods. For actigraphy epochs labelled as outliers, their corresponding HR/HRV epochs were also removed [222].

The experiments in this chapter include participants with sleep disorders to thoroughly evaluate the performance of different methods; full details are presented in Supplementary Table S1. Similarly, this chapter did not exclude a total of 30 subjects (about 2% of the total cohort) who do not have any REM epochs at all, although these sleep patterns are physiologically very unlikely. The sleep stages for subjects in this dataset were scored by individual sleep technicians, blind to the disease status of the participants, into five classes (wake, N1, N2, N3, REM) according to AASM guidelines [221].

For the ECG signal, the derived features are only based on RR intervals instead of using the raw ECG signal. The rationale behind this was to make this work as transferable as possible to data collected from research-grade devices such as miniaturised ECGs or wrist wearables that incorporate PPG sensors (i.e., the Empatica E4 wristband). Participants whose ECG records did not include a full night of sleep or whose data was corrupted were excluded from further analysis.

QRS complexes (R-points) were detected using Compumedics Somte (Abbotsford, VIC, Australia) software Version 2.10 (Builds 99 to 101). The R-points were classified as normal sinus, supraventricular premature complex or ventricular premature complex. Data cleaning, filtering and noise removal took place during this step of the process using the Python package HRV-analysis[6]. First, RR interval outlier data was filtered using a threshold method with a range between 300 to 2000 ms following the method previously described by Tanaka et al. [223], then the ectopic beats were removed by the methods described in Malik et al. [224]. Second, for the removed R-points, the data has been linearly interpolated. In this chapter, the RR intervals have been grouped into 30 seconds to match the time interval of actigraphy data. Recalling the description in Section 3.2, the HRV describes the physiological variation of the beat-to-beat interval that can be extracted from the time distance between adjacent R wave peaks. Thus, 30 cardiac features were calculated from each 30-second window that matches the epoch of the actigraphy data. Following the approach used by Radha et al. [215], the features haven been extracted in four domains (time, geometrical, frequency and non-linear domains). Table 3.4 details the full set of cardiac features used in this chapter.

This chapter adopted two strategies for extracting actigraphy-related features. For DL approaches, which can automatically extract high-level features, the activity counts have been used as input, which can be directly extracted from the device (at a sampling rate of 1/30 Hz). For the other ML models, a total of 370 handcrafted time-series features were extracted, as described in Table 3.3. These features have been commonly used in the literature (i.e., [225, 226, 205]).

For each sleep epoch T, the statistics (i.e., mean, variance, median, kurtosis) have been calculated for actigraphy data that consider both centred and non-centred sliding windows of N sleep epochs (with $N = \{1, 2, ..., 19\}$), where each sleep epoch contains a scalar value. The other commonly used metrics have been calculated, such as the raw and natural logarithm values of the activity counts for each epoch T. These features are listed in Table 3.3.

The full feature set (i.e., both activity and cardiac features) were normalised using the z-score method. A summary of the pipeline used in this chapter is shown in Figure 3.2.

3.3.3. Sleep Stage Classification Tasks

The objectives in this chapter structured five different tasks and tested several hypotheses based on multimodal fusion and new model development. The first task, **Task 1**, aims to establish benchmarks for sleep-wake (binary) classification using single modality (either actigraphy or HR/HRV) and multimodality approaches (combining both modalities). In doing so, this chapter compares conventional statistical learning methods and simple neural network methods across modalities. This task is the most explored one among the research community in this area [202, 200, 201, 226, 227] and this chapter also aimed to augment the benchmarks previously reported by Palotti et al. [205].

Task 2 consisted of the same systematic evaluation, but this time, the simplest sleep staging paradigm was introduced (Wake, NREM, REM). Here, the AASM scores provided in the MESA

[6]https://pypi.org/project/hrv-analysis/

Feature name	Description
Activity Count ✶	*Raw activity count from the actigraphy device*
Log Activity Count ✶	*Natural Logarithm of the activity count*
Mean Activity ✶	*Mean value for the window of activity of size N.* $1 \leq N < 20$
Median Activity ✶	*Median value for the window of activity of size N.* $1 \leq N < 20$
Std Activity ✶	*Standard deviation value for the window of activity of size N.* $1 \leq N < 20$
Variance Activity ✶	*Variance value for the window of activity of size N.* $1 \leq N < 20$
Minimum Activity ✶	*Minimum value for the window of activity of size N.* $1 \leq N < 20$
Maximum Activity ✶	*Maximum value for the window of activity of size N.* $1 \leq N < 20$
NAT Activity ✶	*Number of epochs, in a window of size N, which the value for the activity count is larger than 50 and lower than 100. Devised from [200].* $1 \leq N < 20$
Any Activity ✶	*Number of epochs that contain any activity in the window of size N.* $1 \leq N < 20$
Skewness of Activity ✶	*Skewness for the window of activity of size N.* $4 \leq N < 20$
Kurtosis of Activity ✶	*Kurtosis for the window of activity of size N.* $4 \leq N < 20$

Table 3.3 Full set of features extracted from the actigraphy signal.

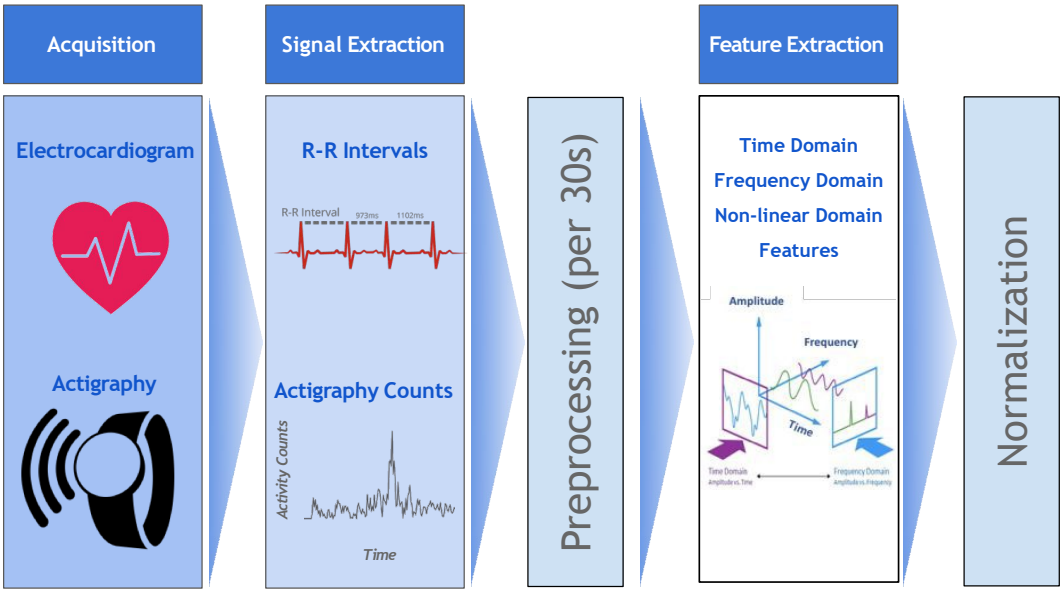


Figure 3.2 Multimodal data processing pipeline: after removing low-quality data, the signals from the actigraphy device and ECG are synchronised and features are extracted and normalised.

dataset are simplified and collapsed into a simpler representation of sleep staging. Wake and REM remain the same, but N1, N2 and N3 are grouped together to become NREM sleep as an entity. The feasibility of this task has also been tested by other studies [215, 61].

Taking a step further in the level of granularity, **Task 3** classifies the data into Wake, REM, Light Sleep and Deep Sleep. Here, light sleep captured both N1 and N2, which is often considered a transition state between light and deep sleep and usually takes up the largest percentage of

Time Domain Features	
Mean HR ♥	*Mean heart rate for that window*
Maximum HR ♥	*Maximum heart rate for that window*
Minimum HR ♥	*Minimum heart rate for that window*
Std HR ♥	*Standard deviation for the heart rate for that window*
SDNN ♥	*Standard deviation of Normal-to-Normal interval (NNi)*
SDSD ♥	*Standard deviation of NNi differences*
NN50 ♥	*Number of NNi differences greater 50ms*
pNN50 ♥	*Ratio between NN50 and total number of NNi*
NN20 ♥	*Number of NNi differences greater 20ms*
pNN20 ♥	*Ratio between NN20 and total number of NNi*
RMSSD ♥	*Root mean of squared NNi differences*
Median NNi ♥	*Median of NNis*
Range NNi ♥	*Range between smallest NN intervals to largest NN intervals*
CVSD ♥	The coefficient of variation of successive differences , the RMSSD divided by mean NNi
Coeff. of Variation of NNI ♥	*The Coefficient of Variation of NNi, i.e. the ratio of sdNN divided by mean NNi*
Geometrical Domain Features	
Triangular Index ♥	*The HRV triangular index measurement is the integral of the density distribution (that is, the number of all NN intervals) divided by the maximum of the density distribution (class width of 8ms)*
Frequency Domain Features	
Low Frequency ♥	*Low Frequency is the variance (i.e., power) in HRV in the Low Frequency (.04 to .15 Hz). Reflects a mixture of sympathetic and parasympathetic activity*
High Frequency ♥	*High Frequency is the variance (i.e., power) in HRV in the High Frequency (.15 to .40 Hz). Reflects fast changes in beat-to-beat variability due to parasympathetic (vagal) activity*
Variance in Low Freq. ♥	*VLF is the variance (i.e., power) in HRV in the Very Low Frequency (.003 to .04 Hz). Reflect an intrinsic rhythm produced by the heart which is modulated by primarily by sympathetic activity*
Low/High Freq. Ratio ♥	*The LF/HF ratio is sometimes used by some investigators as a quantitative mirror of the sympathy/vagal balance*
Norm. Low Freq. Ratio ♥	*Normalized low frequency ratio calculated from the raw values of low frequency band (LF or HF) divided by the total spectral power*
Norm. High Freq. Ratio ♥	*Normalized high frequency ratio calculated from the raw values of high frequency band (LF or HF) divided by the total spectral power*
Mean NNi ♥	*Mean over the NN intervals*
Total Power ♥	*Total power of the density spectral*
Non-Linear Domain Features	
Cardiac Sympathetic Index ♥	*Cardiac Sympathetic Index [124]*
Mod. Cardiac Symp. Index ♥	*A modified cardiac sympathetic index calculated by* $\frac{SD2^2}{SD1}$
Cardiac Vagal Index ♥	*Cardiac Vagal Index [124]*
SD1 ♥	*Poincaré plot standard deviation perpendicular the line of identity*
SD2 ♥	*Poincaré plot standard deviation along the line of identity*
SD1/SD2 Ratio ♥	*Ratio of SD1 to SD2*

Table 3.4 Full set of cardiovascular related features grouped by domain.

time during a full sleep cycle [228]. Given the heterogeneity and prevalence of N2, the difficulty of the task has risen significantly. The models are expected to perform worse than they did on previous tasks.

Task 4 explored the classification of sleep stages based on AASM rules (Wake, REM, N1, N2, N3). This task has the highest level of granularity, and it is, in fact, a task in which even the current state-of-the-art DL approaches on gold-standard PSG recordings often do not achieve satisfactory performance [219]. This task faces two challenges. The first is the class imbalance, as N1 and N3 sleep epochs account for only 11% and 7% of the data, respectively. The second

challenge is the nature of modalities that do not capture direct cortical signals, compromising the performance in more granular classification tasks.

3.3.4. Models and Settings

Conventional heuristic approaches have been readily used in the past 30 years for Task 1 (binary sleep-wake classification). It has recently been shown that feature-based ML and DL approaches greatly outperform all these methods [205].

ML and DL techniques are increasingly used in medical sciences [219, 61]. In this chapter, supervised learning techniques on time-series data have been adopted. This entails generating models that learn mappings between input and output spaces. For instance, Random Forest (RF) approaches have shown strong performance on activity recognition tasks [229]. Similarly, Radu et al. [192] showed promising results using DL approaches on multimodal sensor data for activity and context recognition tasks. Indeed, wearable sensors exploiting multimodal approaches have shown the advantages of these methods over single-modality approaches for human activity recognition tasks [230]. Going beyond traditional activity recognition tasks, ML and DL models have been shown to outperform conventional heuristic approaches for actigraphy-based sleep-wake classification [205, 225]. DL models have also shown great promise in the automatic classification of sleep stages using EEG or multimodal sensor data [153]. This chapter expands that work to multimodal wearable and minimally obtrusive sensors by systematically evaluating how the most well-established ML and DL models perform when using combined sensing.

For all included tasks and modalities, this chapter explores the common shallow ML and DL architectures, which include linear support vector machines, logistic regression, random forest, perceptrons, convolutional neural networks (CNN) and long-short term memory networks (LSTM). CNN and LSTM are commonly used neural network models in sleep-wake [231] and sleep stage classification research based on EEG signals [153, 154, 156]. Finally, this chapter introduced an *ensemble method* which aims to combine the unique *perspectives* and *capabilities* of DL classifiers with different window sizes containing discriminant power from different temporal dependencies that could be characteristic of different sleep stages.

This chapter hypothesised that given the large amounts of data, DL models would be better suited. Details on the ML and DL classifier settings can be found in Table B.1. Deeper architectures (more layers) are explored in the Appendix A.3, Figure A.1 and A.2, but for comparison purposes, this chapter only employed single layer architectures in the main results section.

3.3.5. Experimental Design

Once the feature sets were built for the two input modalities, the dataset has been randomly split into training and test sets following an 80/20 split where 80% (1,395 subjects) went to the training set and 20% (348 subjects) went to the test set. More details, including demographic information, can be found in Table 3.1, and a summary of sleep statistics is introduced in Table 3.2 .

Algorithm Type	Modality	Input Dimension	Features Used (Full list on Tables 3.3 and 3.4)
ML	*[Actigraphy]*	$x \in \mathbb{R}^{370}$	370 features were derived from Activity Counts
	[HR/HRV]	$x \in \mathbb{R}^{30}$	30 features were derived from RR intervals
	[HR/HRV, Actigraphy]	$x \in \mathbb{R}^{400}$	Concatenation of the two modalities above
DL	*[Actigraphy]*	$x \in \mathbb{R}^{l}$	Activity Counts
	[HR/HRV]	$X \in \mathbb{R}^{l \times 8}$	8 features were derived from RR intervals: Mean NNi, Standard Derivation of RR interval (SDNN), RR interval differences (SDSD), Very Low Frequency, Low Frequency, High Frequency Bands, Low Frequency to High Frequency Ratio and Total Power.
	[HR/HRV, Actigraphy]	$X \in \mathbb{R}^{l \times 9}$	Concatenation of the two modalities above

Table 3.5 Experiment settings based on input modalities, where l is the window length of the input ($l = \{21, 51, 101\}$), the inputs are for each sleep epoch

The inputs to the single modality and multimodal experiments can be found in Table 3.5. When using multimodal approaches, a *channel-wise* stacking approach was adopted prior to inputting the resulting matrix into each model. The channel-wised stacking method is widely used for time series studies [232, 151, 152]. All benchmark tasks adopt these methods. Following the method used in [205], the hyperparameter search is described below:

- **ML hyperparameter search:** This chapter employed 5-fold cross-validation on the training set.
- **DL hyperparameter search:** This chapter employed a hold-out method to randomly split the training dataset into a validation set of 279 subjects (20%) and a training set of 1,116 (80%).

The full detailed list for the hyperparameter tuning can be found in Table A.3. Furthermore, the hyper-parameter tuning results of CNNs and LSTMs and the best hyper-parameter settings used in this chapter can be found in Figure A.1 and Figure A.2 in Appendix A.3. The Scikit-learn[7], Keras[8] and Tensorflow[9] were used to implement models. For the feature set, this chapter adopted previously used approaches [215, 205] for movement and cardiac sensor feature extraction in traditional ML and DL setups, which were mentioned in Section 3.2. In the ML experiments, each input vector contains 400 features that combined 370 statistical features extracted from actigraphy and 30 HR/HRV features, as described in the feature engineering section. As such, the single modality approaches for actigraphy input 370 features, whereas for HR/HRV, 30 features for each sleep epoch were included for each input vector. These were used as inputs for the feature-based ML benchmarks.

In DL experiments, ECG signals are expensive and may not be available in most wearable sensors, in order to make this chapter as transferable and device-agnostic as possible, the input

[7]https://scikit-learn.org
[8]https://keras.io
[9]https://tensorflow.org

of the study did not consider the use of ECG raw signals. Thus, instead, this chapter used an 8-dimensional HR/HRV feature set (see Table 3.5) that can be derived from many wearable cardiac sensors, such as Epatica[10], or ActiHeart[11]. For movement data, this chapter simply uses the activity counts that can be acquired directly from the wrist-worn actigraphy device. In the deep learning experiments, following the convention of previous studies, we tested window lengths of 21, 51, and 101 sleep epochs, corresponding to 10.5, 25.5, and 50.5 minutes, respectively. For PSG annotation in clinical settings, sleep technicians and physicians often look at *adjacent* information as well as contextual temporal information to inform their decisions in scoring sleep epochs (sleep stages and/or sleep events). They may look at information and trends within a 30-minute or 1-hour period as well as contextual information regarding the distribution of previous sleep stages to reach a decision [233].

Motivated by this, in this chapter, the ensemble model combines DL classifiers (a combination of CNNs and LSTMs) with various sliding window lengths (21, 51 and 101 sleep epochs, equating to $10.5, 25.5$, and 50.5 minutes, respectively). Following the previously described ensemble model pipelines, this chapter explored two score-level fusion methods *model-averaging* and *maximum posterior selection*.

The sleep stage ensemble classification model is based on standard single-layer CNN and LSTM networks. Figure 3.3 illustrates the structure for individual classifiers and their score-level fusion mechanism. At the training stage, each classifier is trained independently, given the hypothesis that data from sliding windows of different lengths carry different discriminative information for each sleep-stage class.

This chapter used highly overlapping sliding windows with sleep classified at each sleep epoch (i.e., sample-wise classification). Assuming at timestamp (i.e., sleep epoch) t, the m^{th} classifier's output is a K-dimensional probability vector $\mathbf{p}_t^m \in \mathbb{R}^K$, where K is the total sleep class number (for a certain task). Probability vectors from all $M = 6$ models can then be combined using the two different fusing strategies. For *model-averaging*, the fused score can be calculated via:

$$\mathbf{p}_t^{fusion} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}_t^m$$

and label $\hat{k}_t$ can be assigned to the class with the highest probability, i.e.,

$$\hat{k}_t = \underset{k}{\operatorname{argmax}}\, \mathbf{p}_t^{fusion}.$$

The second strategy *maximum posterior selection* simply assigns labels $\hat{k}_t$ to the class with the largest probability among all the M classifiers:

$$\hat{k}_t = \underset{k}{\operatorname{argmax}}\, \mathbf{P}_t, \quad where\, \mathbf{P}_t = [\mathbf{p}_t^1, \mathbf{p}_t^2, ...\mathbf{p}_t^M].$$

[10]https://www.empatica.com
[11]www.camntech.com

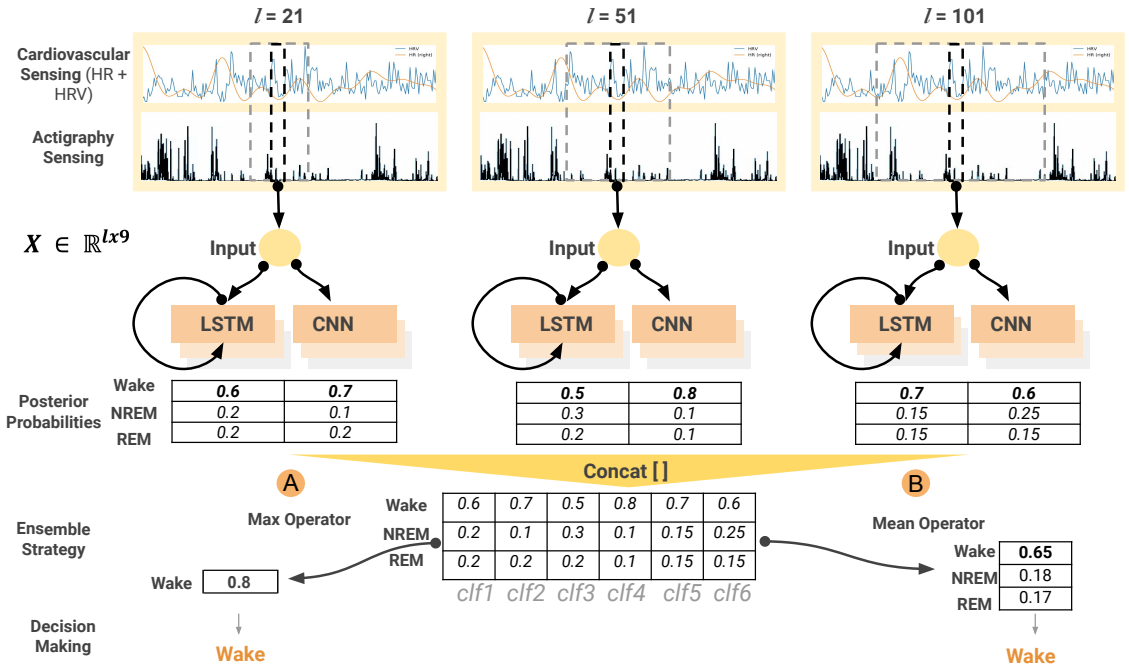


Figure 3.3 Ensemble model illustration. The model starts by taking inputs from different window lengths (l) from the multimodal sensors. A total of six different classifiers are used, combining a mixture of CNNs and LSTMs and exploiting their individual strengths. This produces a probability matrix which is formed by the concatenation operation, which becomes part of the ensemble architecture. Finally, the decision-making layer takes place by either (A) using a maximum calculation or (B) a mean calculation across all classifiers

An example of these two fusing strategies can be found in Figure 3.3.

3.3.6. Evaluation Metrics

This chapter adopted commonly used metrics in machine learning and medical sciences to evaluate the performance of the different classification algorithms based on task and modality combinations. The performance metrics were derived at both a *subject level* (first derived on an individual-by-individual basis and then averaged across the population) and a *group level*.

To assess class imbalance and evaluate performance, several popular metrics were adopted based on True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) classifications. These evaluation metrics can be seen in chapter 2, which include accuracy, recall, specificity, precision, mean F_1 score and Cohen's Kappa (κ):

With the exception of Task 1 (binary sleep-wake classification), all other tasks are multi-class classifications. The performance metrics are calculated in a class-wise manner and reported as the mean values. This chapter generates confusion matrices (corresponding to the best classifiers) for each task to further understand error types, and computed Cohen's Kappa to evaluate the agreement across the whole population. Two-tailed t-tests were used to calculate statistical significance. This chapter also proposed a measure, namely **Time Deviation**, to intuitively understand how long (in minutes), a classifier is either under or over-estimating a certain sleep stage across the whole population.

Task 1			Task 2			Task 3			Task 4		
Sleep Stages	# Epochs	%	Sleep Stages	# Epochs	%	Sleep Stages	# Epochs	%	Sleep Stages	# Epochs	%
Wake	652,509(314,784)	34%(20%)	Wake	652,509(314,784)	34%(20%)	Wake	652,509(314,784)	34%(20%)	Wake	652,509(314,784)	34%(20%)
Sleep	1,251,391	66%(80%)	NREM	1,022,346	54%(65%)	Light	893,472	47%(57%)	N1	171,027	9%(11%)
			REM	229,045	12%(15%)	Deep	128,874	7%(8%)	N2	722,445	38%(46%)
						REM	229,045	12%(15%)	N3	128,874	7%(8%)
									REM	229,045	12%(15%)
Total	1,903,900(1,566,175)	100%	Total	1,903,900(1,566,175)	100%	Total	1,903,900(1,566,175)	100%	Total	1,903,900(1,566,175)	100%

Table 3.6 Number of 30-second sleep epochs for each of the four tasks studied in this chapter. (*The numbers in parentheses were obtained within sleep period time which measured from the first to the last non-wake detected sleep epoch.)

Given N participants, for sleep class k the time deviation TD_k can be expressed as:

$$TD_k = \frac{1}{N}\sum_{i=1}^{N}(Pred_k^i - GT_k^i),$$

where $Pred_k$ is the classifier's prediction and GT_k refers to the ground truth for sleep class k. Both $Pred_k$ and GT_k were measured in minutes. This metric can help to better understand the classifier's performance/bias for a certain sleep class at the population level.

Many sleep classification studies used accuracy or F_1 to measure their model's performance, yet these are high-level metrics which do not consider class-wise performance. Confusion matrices, on the other hand, provide class-wise predictions and corresponding error types. However, for clinicians and other health practitioners, these matrices are not the most obvious way to represent the time deviation of sleep stages, as they include too many low-level details. This chapter proposed **Time Deviation**, is a mid-level metric, which summarises the class-wise performance in an intuitive manner. As such, it can be used as a complementary metric to what is offered by traditional metrics (confusion matrix, accuracy, and F_1), allowing healthcare practitioners to gain an intuitive understanding towards a classifier's reliability.

3.4. Results

The experiments were conducted on a total of 1,743 nights of sleep, representing 1,903,900 sleep epochs of 30 seconds. The prevalence of sleep stages (AASM convention used) within these epochs is reported in Table 3.6. For consistency, all of the architectures and models were evaluated during the sleep recording period across tasks, with performances reported in Table 3.7 for binary classification and Table 3.8 for multistage classification. For the performance within the sleep period, the results can be found in Appendix A.2, Table A.2. Within each table, results were sorted by mean accuracy in descending order. The performance of the benchmark study during the sleep period for all tasks can be found in Table A.2 of Appendix A.2. For each task, a full breakdown of all classifiers is presented in the supplementary materials. This section only shows the top three DL classifiers alongside the best classifier from ML. Table 3.10 provides a summary of sleep measured by PSG and the time spent in different sleep stages for the best classifier in each task.

Sleep-Wake Classification Benchmarks*									
Method Specifics			Performance Metrics						Time Deviation**
Modality	Sensors	Top 3 Classifiers	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Sleep (mins)
Multimodality	♥ 🚶 *[HR/HRV, Actigraphy]*	CNN (101)	**84.4 ± 1.0**	**67.9 ± 2.0**	**84.8 ± 1.3**	92.4 ± 1.2	87.6 ± 1.1	**62.0 ± 2.0**	36.2 ± 7.3
		LSTM (101)	**84.4 ± 1.0**	67.4 ± 1.9	84.7 ± 1.2	92.5 ± 1.1	**87.8 ± 1.0**	61.6 ± 2.1	**36.0 ± 6.7**
		CNN (51)	84.3 ± 1.0	67.3 ± 2.0	84.5 ± 1.3	**92.7 ± 1.2**	87.6 ± 1.1	61.7 ± 2.1	39.0 ± 7.2
		Random Forest (300)	82.3 ± 1.0	65.7 ± 2.1	83.7 ± 1.3	90.6 ± 1.1	57.6 ± 2.1	57.1 ± 2.1	32.9 ± 7.3
Single Modality	♥ *[HR/HRV]*	LSTM (101)	**79.5 ± 1.2**	**62.2 ± 2.1**	**81.8 ± 1.4**	88.9 ± 1.3	**84.1 ± 1.1**	**51.5 ± 2.2**	35.3 ± 4.4
		CNN (101)	79.1 ± 1.2	57.0 ± 2.1	79.9 ± 1.5	**91.0 ± 1.4**	83.9 ± 1.3	49.8 ± 2.1	54.4 ± 4.7
		LSTM (51)	78.6 ± 1.2	61.1 ± 2.0	81.2 ± 1.4	88.2 ± 1.3	83.4 ± 1.2	49.5 ± 2.1	**34.5 ± 4.6**
		Random Forest (300)	70.3 ± 1.2	39.2 ± 2.3	73.6 ± 1.4	86.7 ± 1.9	77.6 ± 1.5	27.1 ± 1.7	70.4 ± 12.5
	🚶 *[Actigraphy]*	CNN (101)	**84.9 ± 1.0**	67.1 ± 2.0	84.7 ± 1.3	**93.8 ± 1.0**	**88.3 ± 1.0**	**63.0 ± 2.0**	43.0 ± 6.9
		CNN (51)	84.4 ± 1.0	67.6 ± 2.0	84.6 ± 1.3	92.9 ± 1.1	87.8 ± 1.1	62.2 ± 2.1	39.0 ± 7.1
		LSTM (101)	84.3 ± 1.0	**69.7 ± 1.8**	**85.5 ± 1.2**	91.2 ± 1.1	87.6 ± 1.0	62.0 ± 2.0	**26.5 ± 6.6**
		Random Forest (300)	81.2 ± 1.0	63.4 ± 2.0	82.9 ± 1.3	89.7 ± 1.1	85.4 ± 1.0	54.1 ± 2.1	32.6 ± 7.2

Table 3.7 Sleep wake classification results (mean ± standard error at 95% confidence interval) and predicted minutes by multimodal and single modality approaches (full recording period); (*Full Table available on supplementary, **Average time deviation from ground truth across all subjects ± standard error)

3.4.1. Task 1: Sleep-Wake Classification

The best-performing algorithms for Task 1 are presented in Table 3.7, and a full breakdown of all classifiers are presented in the supplementary tables for this task. The baseline approaches were used here, namely, *Always Sleep* and *Always Wake*, which showed that 66.5% of the epochs is sleep. Given the fact that for the purpose of this chapter, the classification during the night period was explored. The minimum accuracy threshold which is 66.5% was established as the baseline performance, as the dataset is imbalanced. Furthermore, although not reported in this chapter, several of the well-established heuristic algorithms were tested such as Cole-Kripke [202] and Sadeh [200] on single-modality actigraphy data. The results agree with what was reported by Palotti et al. [205]. All of these approaches were outperformed by both the feature-based ML and DL models explored in this chapter.

All traditional ML modalities showed similar performance. Corroborating what had been shown in the related work that, when these algorithms are applied to actigraphy data, they result in high sensitivity but poor specificity [205]. Interestingly, for this task, adding HR and HRV to actigraphy for a combined sensing modality on the top classifier of CNN (101) did not significantly improve F_1 ($p = 0.347$), accuracy ($p = 0.499$) or Cohen's κ ($p = 0.506$). As expected, HR/HRV alone did not yield comparable performance to actigraphy alone or the combined sensing approach.

3.4.2. Task 2: Wake, Non-REM sleep, REM Sleep Classification

Task 2 evaluated sleep stages from a low granularity perspective by aggregating the different partitions of NREM. As observed in the supplementary table for this task, although some ML models had reasonable performance, the DL approaches were superior. It is important to note that at this level of granularity, NREM is overestimated while REM is underestimated for almost all models except for CNN (51) and CNN (101). In contrast to what was observed in Task 1, all models explored have a higher specificity than sensitivity and accuracy higher than F_1 score due to the imbalanced dataset.

Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	NREM
Multimod.	♥ 🚶	LSTM (51)	**76.2 ± 1.0**	**85.6 ± 0.5**	**72.2 ± 1.3**	68.8 ± 1.2	67.9 ± 1.3	58.4 ± 1.8	-13.2 ± 6.8	-10.7 ± 3.8	**23.9 ± 7.1**
		LSTM (101)	76.1 ± 0.9	85.1 ± 0.5	71.9 ± 1.4	66.8 ± 1.2	66.4 ± 1.3	57.4 ± 1.9	**-3.2 ± 6.8**	-23.3 ± 3.4	26.5 ± 7.0
		CNN (101)	76.0 ± 1.0	**85.6 ± 0.6**	**72.2 ± 1.2**	**69.7 ± 1.3**	**68.1 ± 1.3**	**58.6 ± 1.9**	-32.7 ± 7.2	**2.5 ± 4.5**	30.2 ± 7.7
		Random Forest (300)	70.5 ± 0.9	79.9 ± 0.5	59.2 ± 1.5	53.0 ± 0.7	50.3 ± 0.7	47.6 ± 1.7	-20.0 ± 7.6	-63.6 ± 2.9	83.5± 7.8
Single Modality	♥	LSTM (101)	**73.8 ± 1.2**	**84.3 ± 0.6**	**69.8 ± 1.5**	**66.1 ± 1.3**	**64.9 ± 1.5**	**50.0 ± 2.2**	-27.8 ± 8.5	-8.6 ± 4.2	36.4 ± 8.1
		LSTM (51)	72.9 ± 1.1	83.8 ± 0.6	67.9 ± 1.4	64.1 ± 1.3	62.9 ± 1.4	45.5 ± 2.1	-17.9 ± 8.5	-16.2 ± 4.3	34.1 ± 8.3
		CNN (101)	71.0 ± 1.2	83.6 ± 0.6	66.3 ± 1.4	65.4 ± 1.4	62.7 ± 1.4	46.1 ± 2.0	**-12.9 ± 9.4**	**2.3 ± 5.0**	**10.6 ± 9.1**
		Random Forest (300)	59.6 ± 1.0	73.8 ± 0.4	48.4 ± 1.4	43.4 ± 0.6	39.2 ± 0.8	19.7 ± 1.4	-26.8 ± 13.5	-65.3 ± 3.1	92.0 ± 13.2
	🚶	LSTM (101)	**71.4 ± 0.9**	**80.1 ± 0.6**	**51.7 ± 1.1**	**52.9 ± 0.7**	**49.8 ± 0.8**	**49.7 ± 1.7**	**-16.8 ± 7.3**	**-67.0 ± 3.0**	**83.8 ± 7.7**
		CNN (101)	71.0 ± 1.0	79.5 ± 0.6	50.1 ± 0.8	52.1 ± 0.8	49.1 ± 0.8	48.0 ± 1.8	-34.9 ± 7.5	-67.6 ± 3.0	102.5 ± 7.9
		LSTM (51)	70.9 ± 0.9	79.7 ± 0.6	49.0 ± 0.8	52.4 ± 0.7	49.2 ± 0.8	48.3 ± 1.7	-20.3 ± 7.4	-67.6 ± 3.0	87.9 ± 7.8
		Random Forest (300)	68.6 ± 0.9	78.8 ± 0.5	53.4 ± 1.1	51.0 ± 0.7	48.5 ± 0.8	44.3 ± 1.7	-20.5 ± 7.2	-60.7 ± 2.9	81.2 ± 7.4

Task 3: Wake, Light Sleep, Deep Sleep, REM

Method Specifics			**Performance Metrics**						**Time Deviation***			
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	Deep Sleep	Light Sleep
Multimod.	♥ 🚶	LSTM (51)	**70.3 ± 1.0**	**87.4 ± 0.4**	57.9 ± 1.3	**54.0 ± 1.0**	**51.9 ± 1.0**	**53.8 ± 1.9**	**-1.0 ± 6.9**	-5.6 ± 4.0	-36.2 ± 3.5	**42.8 ± 7.4**
		LSTM (101)	70.2 ± 1.0	86.9 ± 0.4	**59.9 ± 1.5**	52.4 ± 1.0	51.3 ± 1.1	51.7 ± 1.8	-18.9 ± 6.6	-24.7 ± 3.7	**-32.4 ± 3.5**	76.0 ± 7.3
		CNN (101)	69.0 ± 1.0	87.0 ± 0.4	58.0 ± 1.4	53.7 ± 1.0	51.2 ± 1.1	51.6 ± 1.8	-15.9 ± 7.5	**4.4 ± 4.8**	-34.5 ± 3.5	46.1 ± 8.1
		Random Forest (300)	63.6 ± 1.0	83.3 ± 0.4	44.7 ± 1.3	40.1 ± 0.6	36.7 ± 0.6	34.4 ± 1.3	-15.2 ± 7.6	-61.3 ± 2.9	-38.9 ± 3.6	115.3 ± 8.3
Single Modality	♥	LSTM (101)	**67.4 ± 1.2**	**86.2 ± 0.4**	**56.2 ± 1.6**	**51.3 ± 1.1**	**49.5 ± 1.2**	**44.6 ± 2.2**	**-13.1 ± 8.4**	-13.5 ± 3.8	**-33.7 ± 3.5**	60.4 ± 8.1
		LSTM (51)	66.2 ± 1.1	85.6 ± 0.4	54.4 ± 1.5	49.5 ± 1.1	47.4 ± 1.1	41.2 ± 2.1	-15.0 ± 8.1	-14.6 ± 4.1	-36.4 ± 3.5	65.9 ± 7.9
		CNN (101)	64.3 ± 1.1	85.3 ± 0.4	54.4 ± 1.6	50.2 ± 1.1	47.1 ± 1.1	40.9 ± 2.1	-23.0 ± 9.5	**8.2 ± 5.1**	-34.8 ± 3.5	**49.6 ± 8.9**
		Random Forest (300)	53.3 ± 1.0	79.2 ± 0.4	35.5 ± 1.1	33.2 ± 0.5	28.6 ± 0.6	12.6 ± 1.1	-4.3 ± 14.0	-64.7 ±3.0	-39.0 ± 3.6	-39.0 ± 3.6
	🚶	LSTM (101)	**64.1 ± 1.0**	82.9 ± 0.5	35.6 ± 0.7	**39.6 ± 0.7**	**35.8 ± 0.7**	**33.5 ± 1.4**	-32.5 ± 7.4	-67.6 ± 3.0	**-39.3 ± 3.6**	139.4 ± 8.5
		CNN (101)	63.9 ± 1.0	**83.0 ± 0.4**	**36.3 ± 0.9**	**39.6 ± 0.7**	35.7 ± 0.7	**33.5 ± 1.4**	**-26.4 ± 7.6**	-67.5 ± 3.0	**-39.3 ± 3.6**	**133.2 ± 8.7**
		LSTM (51)	63.6 ± 1.0	82.7 ± 0.4	35.6 ± 0.8	39.3 ± 0.7	35.5 ± 0.7	33.0 ± 1.4	-36.3 ± 7.1	**-67.3 ± 3.0**	**-39.3 ± 3.6**	143.0 ± 8.2
		Random Forest (300)	61.4 ± 1.0	82.6 ± 0.4	39.6 ± 0.9	38.1 ± 0.5	35.0 ± 0.6	31.2 ± 1.3	-15.6 ± 7.3	-59.3 ± 2.9	-37.1 ± 3.6	112.1 ± 8.1

Task 4: Wake, REM, N1,N2,N3

Method Specifics			**Performance Metrics**						**Time Deviation ****				
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	N3 Sleep	N2 Sleep	N1 Sleep
Multimod.	♥ 🚶	LSTM (51)	**63.7 ± 1.0**	88.7 ± 0.3	47.1 ± 1.4	43.0 ± 0.8	39.9 ± 0.8	56.3 ± 1.8	22.2 ± 7.1	-12.9 ± 3.9	-35.2 ± 3.5	**71.9 ± 7.5**	-46.0 ± 3.0
		LSTM (101)	63.6 ± 1.0	88.7 ± 0.3	47.8 ± 1.3	43.3 ± 0.8	40.5 ± 0.9	**57.0 ± 1.8**	**-3.3 ± 6.8**	-15.9 ± 3.9	**-32.3 ± 3.5**	97.7 ± 7.5	-46.2 ± 3.0
		CNN (101)	63.1 ± 1.1	**88.8 ± 0.3**	**51.5 ± 1.4**	**44.7 ± 0.9**	**41.9 ± 0.9**	56.2 ± 1.8	-26.2 ± 7.1	**8.2 ± 5.0**	-34.3 ± 3.5	92.4 ± 8.0	**-40.2 ± 3.1**
		Random Forest (300)	56.9 ± 1.0	86.2 ± 0.3	36.4 ± 1.2	33.1 ± 0.5	28.8 ± 0.5	46.3 ± 1.6	18.6 ± 8.1	-54.9 ± 3.1	-38.7 ± 3.6	123.6 ± 8.4	-48.6 ± 3.2
Single Modality	♥	CNN (21)	**55.6 ± 1.1**	86.4 ± 0.3	40.4 ± 1.2	37.3 ± 0.8	33.6 ± 0.9	36.2 ± 1.8	-15.6 ± 10.1	**-3.9 ± 5.8**	-39.1 ± 3.6	103.0 ± 9.8	-44.4 ± 3.0
		CNN (101)	**55.6 ± 1.1**	**86.7 ± 0.3**	**44.9 ± 1.4**	**38.9 ± 0.9**	**35.9 ± 1.0**	**37.1 ± 1.8**	**1.1 ± 10.7**	-12.0 ± 4.8	**-29.5 ± 3.6**	**81.2 ± 9.4**	**-40.8 ± 3.1**
		CNN (51)	54.2 ± 1.1	86.0 ± 0.3	41.2 ± 1.3	35.6 ± 0.8	32.1 ± 1.0	32.3 ± 1.9	35.7 ± 12.1	-28.2 ± 5.1	-36.4 ± 3.5	69.5 ± 11.0	-40.6 ± 3.1
		Random Forest (300)	46.6 ± 1.0	83.1 ± 0.3	29.9 ± 1.0	27.1 ± 0.4	22.3 ± 0.5	17.6 ± 1.4	48.5 ± 14.3	-61.3 ±3.1	-38.9 ± 3.6	97.8 ± 13.7	-46.2 ± 3.2
	🚶	LSTM (51)	**56.9 ± 1.0**	**85.7 ± 0.4**	26.1 ± 0.8	32.2 ± 0.6	27.1 ± 0.7	46.9 ± 1.7	-12.9 ± 7.5	-67.6 ± 3.0	**-39.3 ± 3.6**	169.2 ± 8.5	**-49.4 ± 3.2**
		LSTM (101)	**56.9 ± 1.0**	**85.7 ± 0.4**	25.3 ± 0.7	**32.3 ± 0.6**	27.1 ± 0.7	**47.1 ± 1.7**	**-3.3 ± 7.5**	-67.6 ± 3.0	**-39.3 ± 3.6**	159.7 ± 8.7	**-49.4 ± 3.2**
		CNN (101)	56.8 ± 1.1	85.8 ± 0.3	**27.7 ± 0.9**	32.2 ± 0.5	**27.2 ± 0.6**	46.9 ± 1.7	9.6 ± 8.3	**-65.6 ± 3.0**	**-39.3 ± 3.6**	**144.7 ± 9.1**	**-49.4 ± 3.2**
		Random Forest (300)	54.4 ± 1.0	85.6 ± 0.3	31.7 ± 0.8	31.1 ± 0.4	27.2 ± 1.6	42.7 ± 1.6	16.2 ± 7.6	-56.0 ±2.8	-36.7 ± 3.5	120.8 ± 8.0	-44.3 ± 3.3

Table 3.8 Sleep stage classification results (mean ± standard error at 95% confidence interval and predicted minutes by multimodal and single modality approaches (full recording period); (*Full Table available on supplementary, **Average time deviation from ground truth across all subjects ± standard error)

As reflected in the top part of Table 3.8, the best classifiers for this task were all DL models for all sensor modalities. These models were significantly better than the best traditional ML model (Random Forest), with $p < 0.001$ for all metrics evaluated. The best DL algorithm with respect to accuracy was LSTM (51) which was also statistically better than CNN (21) ($p < 0.001$) achieving an accuracy of 76.2%. However, it was not significantly better than CNN (101) in terms of F_1, sensitivity and specificity ($p = 0.364, p = 0.138, p = 0.063$ and $p = 0.399$). Nevertheless, CNN (101) achieved the lowest mean time deviation with a 2.5-minute overestimation of REM sleep. Interestingly, for this task, most algorithms' specificity significantly improved (e.g. LSTM (51) $p < 0.001$, reaching a specificity of 86%) when compared to Task 1 with the exception of the perceptron model.

In this task, it becomes apparent that multimodality is required for better performance at multistage classification, with the single modality approaches being significantly ($p < 0.001$) outperformed in all performance metrics and yielding much larger time deviations.

3.4.3. Task 3: Wake, Light Sleep, Deep Sleep and REM-sleep Slassification

Task 3 explored sleep staging at a higher level of granularity than Task 2, with class imbalances being perhaps more apparent, as shown in Table 3.6. Here, DL approaches continued to

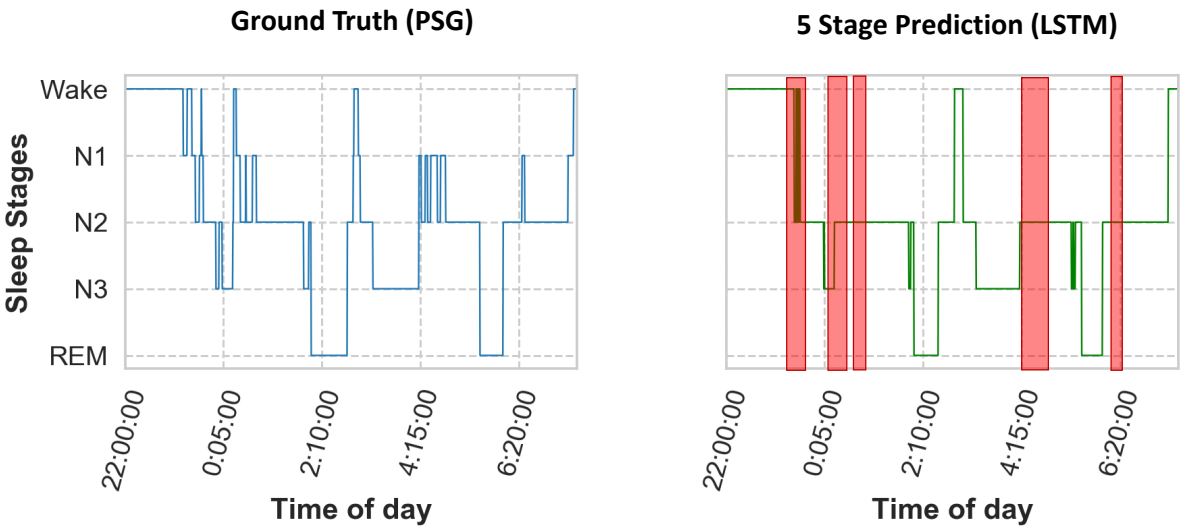


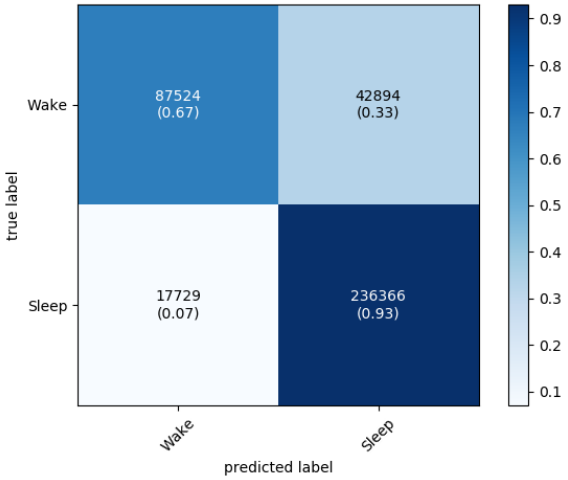
Figure 3.4 Classification performance for multimodal, 5-stage classification using LSTM. The top figure is the ground truth PSG, and the figure at the bottom is the predicted stages by the model. Highlighted in red are areas where the model does poorly.

outperform all feature-based ML models except for the Random Forest, which was not significantly worse than the CNN (21). The full results are available in supplementary tables. The best-performing model was LSTM (51), although it was closely followed by LSTM (101) and CNN (101). Multimodal approaches were significantly better ($p < 0.001$) across all metrics upon the comparison of the best classifiers for each category explored, depicting the value of these combined sensing approaches for multistage classification. Across all sensing modalities and all algorithms, deep and REM sleep were underestimated, with the exception of CNN (101) in the multimodal setup. In contrast, light sleep was overestimated, with Wake being slightly underestimated across all setups, due to the class imbalance, except for LSTM (51).

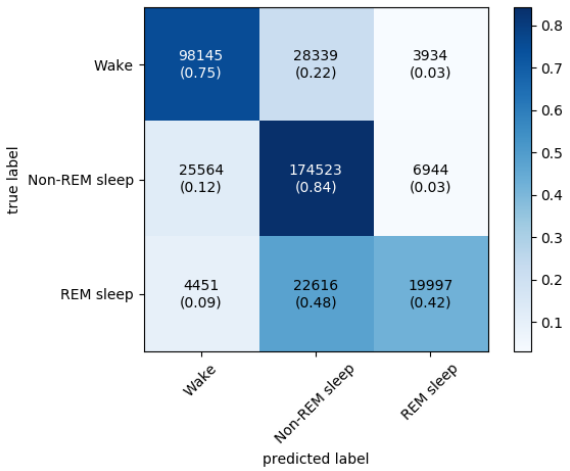
3.4.4. Task 4: Wake, N1, N2, N3, REM Sleep Classification

Finally, Task 4 aimed to classify sleep stages following AASM rules (N1, N2, N3, REM and Wake). This task is the most complex due to its level of granularity and high-class imbalance, and as expected, the models performed worse here than in the previous tasks. An example of the best-performing model LSTM (101) and the *mistakes* it makes is highlighted in Figure 3.4.

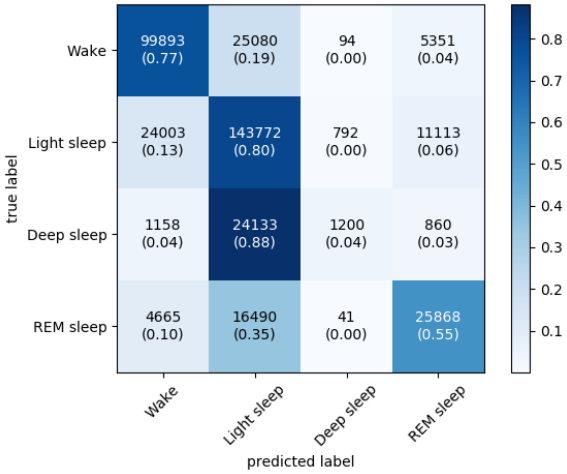
Like in previous tasks, the performance of DL algorithms was significantly better than feature-based ML algorithms as depicted in Table 3.8. The three best-performing DL algorithms were not significantly different from each other with respect to accuracy($p > 0.05$) and F1 scores ($p > 0.05$). The best-performing algorithm was LSTM (51), with an accuracy of 63.7% and an F_1 score of 39.9%. Even in the best-performing multimodal approach, N2 tended to be severely overestimated (71 minutes more on average across the population). Nevertheless,



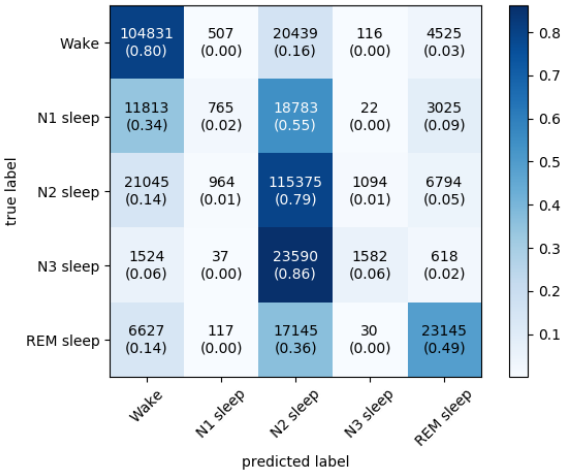
(a) Task 1: Wake, Sleep (CNN (101))



(b) Task 2: Wake, NREM, REM (LSTM (51))



(c) Task 3: Wake, Light Sleep, Deep Sleep, REM (LSTM (51))



(d) Task 4: Wake, N1, N2, N3, REM (LSTM (51))

Figure 3.5 Confusion matrix for the best classifier per Task

the multimodal and HR/HRV approaches were good at classifying Wake and REM, with only moderate deviations in time for those classes.

It is important to note that although the performance in terms of accuracy for the single modality approaches was comparable, each method struggled or had strengths at very different things. For instance, HR/HRV was significantly better at classifying REM sleep in this modality than actigraphy. Similarly, upon evaluation of the algorithms only during the sleep period only (Table A.2, multimodal approaches were significantly better at detecting awakenings, yielding a more accurate wake after sleep onset (WASO) metric. In this thesis, we consider the wake-up minutes as the value of WASO which is between sleep onset and sleep offset.

Figure 3.5 shows the confusion matrix for the best classifiers per task, allowing us to observe how models have an *easier* time classifying REM and Wake and struggle to classify N1 and N3 (NREM). The observed time deviation in minutes substantiates this finding.

Finally, this chapter evaluated the performance of different ensemble methods for each task. To validate the performance of the proposed ensemble model, a t-test was conducted on both the

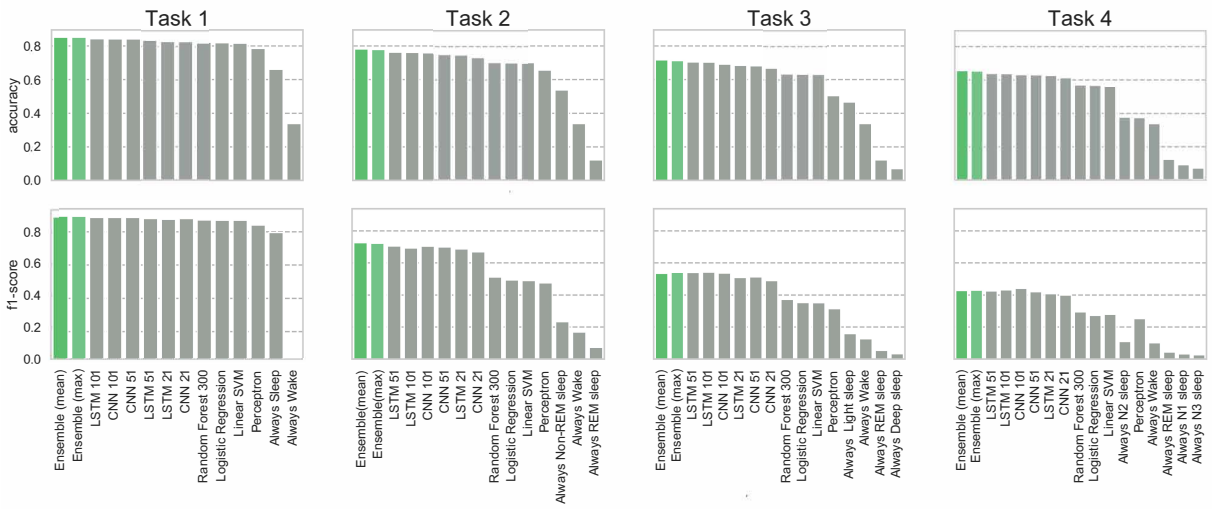


Figure 3.6 Performance (accuracy, F_1) per Task and model. Task 5 (ensemble architectures) are depicted against all benchmarks per each task on green

subject level as well as the group of subjects level. The difference between the two experiments lies in that the second approach, randomly divides all test subjects into 29 groups, each group containing 12 individuals. The purpose is to test whether the benefits of using ensemble methods are due to random chance.

The results of the two ensemble architecture models explored (based on different score-level fusion approaches) are shown in Table 3.9. The results show that there is no significant difference between the two ensemble models for all evaluated performance metrics. However, they achieve better accuracy than single-classifier approaches for all tasks and are significantly better on several performance metrics.

In Task 2, the ensemble approaches significantly outperformed LSTM (51) in terms of accuracy ($p < 0.05$), F_1 score ($p < 0.05$) and Cohen's κ ($p < 0.05$) and CNN (101) in terms of accuracy ($p < 0.05$) and Cohen's κ ($p < 0.05$) based on both subject and group level t test. These models were the best two performers for that task prior to the introduction of the ensemble approach. Interestingly, on Task 4 (highest level of class granularity) the ensemble models only outperformed the best classifier (LSTM (51)) in terms of Cohen's κ and accuracy (for both subject and group level t test).

A summary of results per class and model is presented in Figure 3.6.

3.4.5. Feature Importance Analysis

To understand how different modalities contribute to the prediction of each sleep stage, models that rank feature importance can be implemented. Many traditional ML approaches can provide feature importance ranking, such as logistic regression, linear Support Vector Machine or Random Forests. Of those, Random Forest is one of the most powerful traditional ML models, and it can rank feature importance by calculating the mean Gini impurity or mean information gain over all its decision trees. However, these approaches only yield features that are important to the holistic classification task and do not provide information on how these features contribute to

Ensemble method	Accuracy	Cohen's κ	F_1	Precision	Recall	Specificity	Time Deviation (mins)				
Task 1 (2 Stages)							**Sleep**				
Maximum selection	85.3 ± 1.0	**64.4 ± 2.0**	88.4 ± 1.1	**85.7 ± 1.2**	92.8 ± 1.1	**70.1 ± 1.9**	**33.4 ± 6.7**				
Mean over classifiers	**85.4 ± 1.0**	64.3 ± 2.1	**88.5 ± 1.0**	85.4 ± 1.3	**93.4 ± 1.1**	69.1 ± 2.0	37.5 ± 6.8				
Task 2 (3 stages)							**Wake**	**REM sleep**	**NREM sleep**		
Maximum selection	77.9 ± 1.0	61.4 ± 1.8	69.6 ± 1.3	74.5 ± 1.3	70.6 ± 1.2	**86.5 ± 0.5**	**-16.1 ± 6.8**	**-11.1 ± 4.2**	**27.3 ± 7.5**		
Mean over classifiers	**78.2 ± 0.9**	**61.9 ± 1.8**	**69.8 ± 1.3**	**75.2 ± 1.3**	**70.7 ± 1.3**	**86.5 ± 0.5**	-21.7 ± 6.8	-13 ± 4.2	34.7 ± 7.5		
Task 3 (4 stages)							**Wake**	**REM sleep**	**Deep sleep**	**Light sleep**	
Maximum selection	71.1 ± 1.0	55.7 ± 1.8	**52.4 ± 1.0**	**58.3 ± 1.3**	**54.8 ± 1.0**	**87.7 ± 0.4**	**-11.1 ± 6.7**	**-3.5 ± 4.6**	**-37.4 ± 3.5**	**52.0 ± 7.8**	
Mean over classifiers	**71.6 ± 1.0**	**56.1 ± 1.8**	52.1 ± 1.0	57.1 ± 1.2	54.3 ± 1.0	87.6 ± 0.4	-17.8 ± 6.7	-8.9 ± 4.4	-38.5 ± 3.5	65.2 ± 7.8	
Task 4 (5 stages)							**Wake**	**REM sleep**	**N3 sleep**	**N2 sleep**	**N1 sleep**
Maximum selection	65.2 ± 1.0	59.6 ± 1.8	**41.4 ± 0.8**	**49.7 ± 1.4**	**45.2 ± 0.8**	**89.3 ± 0.3**	**3.0 ± 6.9**	4.7 ± 5.1	**-37.1 ± 3.5**	**76.4 ± 7.8**	-47 ± 3.1
Mean over classifiers	**65.4 ± 1.0**	**60.1 ± 1.8**	41.2 ± 0.8	48.6 ± 1.4	44.9 ± 0.8	89.2 ± 0.3	-5.1 ± 6.9	**-2.1 ± 4.8**	-38.4 ± 3.5	91.9 ± 7.8	**-46.3 ± 3.1**

Table 3.9 Results (mean ± standard error at 95% confidence interval) of different ensemble methods for each task.(Mean over classifiers and Maximum selection are ensemble models)

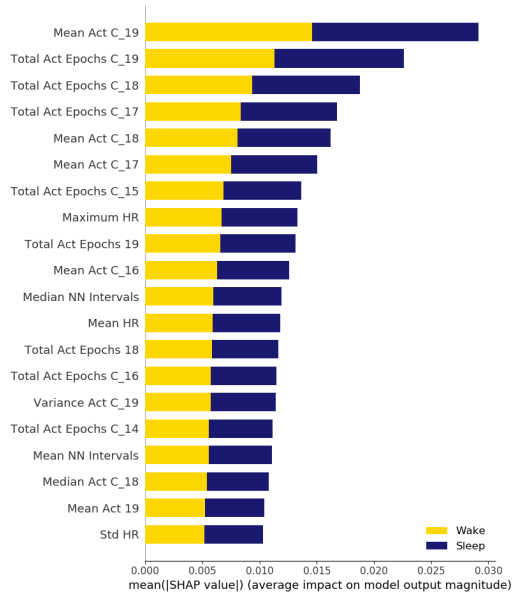
Minutes of Sleep Stages						
Task	**Methods**	**Wake**	**Sleep**			
1	Ground truth	187.8±81.6 (179.2-196.4)	365.7±81.8 (357.1-374.3)			
	Mean over classifiers	150.2±73.2 (142.5-157.9)	403.3±92.0 (393.6-413.0)			
		Wake	**REM**	**NREM**		
2	Ground truth	187.8±81.6 (179.2-196.4)	68.6±27.2 (65.7-71.5)	299.7±66.5 (292.7-306.7)		
	Mean over classifiers	166.1±77.2 (158.0-174.2)	57.4±39.0 (53.3-61.5)	332.7±87.8 (323.5-341.9)		
		Wake	**REM**	**Deep Sleep**	**Light Sleep**	
3	Ground truth	187.8±81.6 (179.2-196.4)	68.6±27.2 (65.7-71.5)	41.8±33.7 (38.3-45.3)	258.8±65.7 (251.9-265.7)	
	Maximum selection	176.7±78.3 (168.5-184.9)	65.4±43.1 (60.9-69.9)	5.2±6.2 (4.5-5.9)	310.6±85.4 (301.6-319.6)	
		Wake	**REM**	**N1**	**N2**	**N3**
4	Ground truth	187.8±81.6 (179.2-196.4)	68.6±27.2 (65.7-71.5)	50.1±30.5 (46.9-53.3)	209.1±58.7 (202.9-215.3)	41.8±33.7 (38.3-45.3)
	CNN (101)	161.6±79.1 (153.3-169.9)	76.7±46.8 (71.8-81.6)	9.6±14.6 (8.1-11.1)	301.5±87.0 (292.4-310.6)	7.7±8.7 (6.8-8.6)

Sleep Parameters						
Task	**Methods**	**Total Sleep Time**	**Wake After Sleep Onset**	**Sleep Period Duration**	**Sleep Efficiency (Recording Period)**	**Sleep Efficiency (Sleep Period)**
	Ground truth	365.7±81.8 (357.1-374.3)	89.4±62.5 (82.8-96.0)	455.1±90.0 (445.6-464.6)	66.5±12.9 (65.1-67.9)	80.9±11.8 (79.7-82.1)
1	Mean over classifiers	360.7±84.0 (351.9-369.5)	70.9±57.4 (64.9-76.9)	474.2±92.9 (464.4-484.0)	65.5±13.2 (64.1-66.9)	76.9±14.0 (75.4-78.4)
2	Mean over classifiers	359.5±84.3 (350.6-368.4)	81.7±60.1 (75.4-88.0)	469.1±93.0 (459.3-478.9)	65.3±13.2 (63.9-66.7)	77.4±13.9 (75.9-78.9)
3	Maximum selection	358.2±84.1 (349.4-367.0)	86.5±61.4 (80.0-93.0)	463.3±92.3 (453.6-473.0)	65.0±13.2 (63.6-66.4)	78.1±13.7 (76.7-79.5)
4	CNN (101)	361.1±83.3 (352.3-369.9)	94.6±68.6 (87.4-101.8)	486.5±90.7 (477.0-496.0)	65.6±13.1 (64.2-67.0)	75.0±14.2 (73.5-76.5)

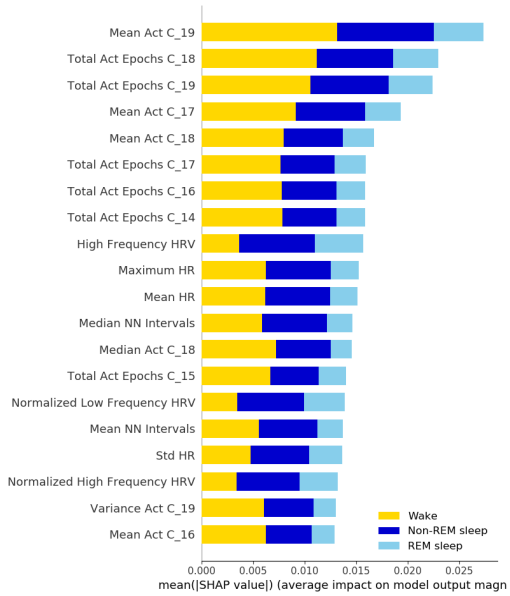
Table 3.10 Sleep parameters and predicted minutes of each sleep stage in the *test* dataset. Numbers are minutes except for the sleep efficiencies which are reported as percentages.
(*Results are in mean +- SD/ and numbers in parentheses indicate the range in 95% CI (Mean over classifiers and Maximum selection are ensemble models))

recognising certain classes (e.g., a sleep stage like REM sleep). This chapter used SHAP [197] with Random Forest, which can generate class-wise feature importance. More technical details of SHAP can be found in [197].

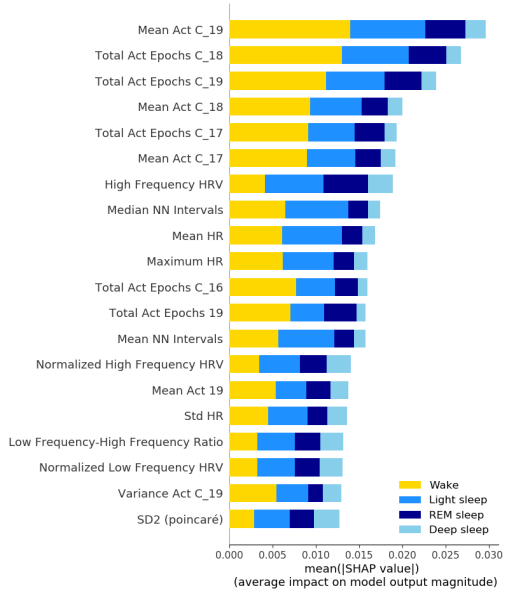
By using this SHAP implementation with Random Forest, the feature importance score can be calculated per class, as shown in Figure 3.7. This chapter reports the top 20 features on Tasks 1-4, respectively. It is interesting to see how the top-ranked features differ from task to task, pointing towards what contributes to more granular levels of classification. For instance, in Task 1 (i.e., binary sleep-wake classification), the most informative features are from movement sensors (15 features out of 20), in contrast to those obtained from cardiac sensing (5 out of 20). However, cardiac features become more and more important as multi-stage classification tasks get more granular. In Tasks 2-4, with the increased class granularity, the most important features are cardiac features (8, 10, and 13 cardiac features, respectively) indicating the key role of cardiac sensing in distinguishing detailed sleep patterns.



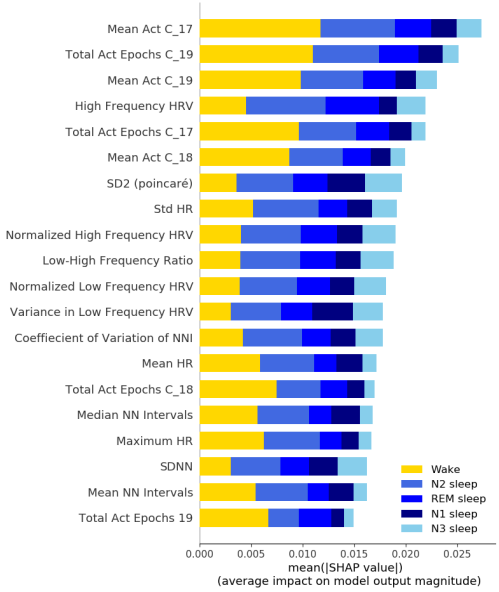
(a) SHAP for Task 1 : Wake, Sleep



(b) SHAP for Task 2 : Wake, NREM, REM



(c) SHAP for Task 3 : Wake, Light Sleep, Deep Sleep, REM



(d) SHAP for Task 4: Wake, N1, N2, N3, REM

Figure 3.7 SHAP values (Random Forest) for class-wise feature importance ranking in Task 1-4

In multi-stage sleep classification (i.e., Tasks 2-4), it is also interesting to see feature importance associated with the different classes. Specifically, this chapter observed high frequency HRV is the most discriminant feature in recognising REM sleep, a finding that is consistent across all 3 multi-stage classification tasks. However, high frequency HRV is not as valuable in recognising wake status. In Tasks 3 and 4, the non-linear HRV features such as SD2 which is then normalised Poincare plot parameter, SDNN, and coefficient of variation of NNI become more important than time domain features of HRV. Among these features, SD2 is ranked higher than many of the other HRV features except high frequency HRV in Task 4.

3.5. Discussion

3.5.1. Summary

This chapter presents the first systematic analysis of sleep-wake and sleep-stage classification using multimodal sensor data in a large, diverse population of both healthy and sleep-disordered participants. The main aim of this chapter was to understand how different models performed based on details of the task (sleep-wake or multistage sleep classification) and sensor combination. To achieve this, a series of traditional ML and DL approaches were tested for each individual modality (i.e., actigraphy, ECG) and sensor combination (multimodal sensor fusion). Furthermore, four different tasks were performed to gain a deeper understanding of the strengths and limitations of the different approaches.

These tasks include sleep-wake (Task 1); wake, NREM and REM (Task 2); wake, light sleep, deep sleep and REM (Task 3); and wake, N1, N2, N3 and REM (Task 4). The framework and analysis provided were based on sensor modalities and signals that can be obtained from research-grade wearable devices. Hence, RR-based metrics were used instead of raw ECG signals. Unlike raw ECG, these metrics may be derived from commercial research-grade wearable devices and in the near future also from non-clinical smartwatches that use both actigraphy/accelerometers and photoplethysmogram (PPG) [234, 235]. This chapter aims to provide a set of benchmarks for commercial and research studies and to inspire others to create open-access large population repositories to study the role of sleep and other physical behaviours in health and disease.

This chapter systematically evaluated how sensor modality affects classification outcomes and how model choice leads to differences in performance. Yuda et al. also explored a multimodal approach to sleep classification. Although their work is strong methodologically, the cohort is much smaller than that presented here, almost 70% of their cohort is male and the majority of their subjects had sleep disorders, limiting the generalisability of their findings [61]. Furthermore, they only explored the classification of three sleep stages. This chapter shows that although multimodal sensor approaches do not lead to great improvements in classification performance for sleep-wake classification tasks, they are essential to classify sleep stages. For instance, actigraphy by itself struggles to classify REM sleep in Tasks 2-4 (Table 3.8), but its performance improves when combined with HR/HRV. To date, conventional sleep-wake classification algorithms have mostly exploited count-based movement data [200, 203] and models that combined HR information have mostly been confined to commercial devices based on device-specific algorithms.

Furthermore, this chapter highlights the strengths and limitations of the different models used, for instance, while CNN models do well at classifying high frequency transitions, LSTMs excel at classifying smooth patterns. LSTMs outperformed all other classifiers at multistage classification due to their deep temporal modelling characteristics which align well with the multi-class, time-series classification problem that sleep stages introduce. Meanwhile, CNNs were the best performers for binary sleep-wake classification tasks due to their ability to track exponentially longer sequences, such as those used in this type of task where the objective is

less granular and has lower transition frequencies than the multi-class scenario. As such, the ensemble approach aimed to exploit individual model *strengths* to achieve better performance.

In sum, this chapter presents the first systematic analysis of single modality (actigraphy, HR/HRV) and multimodal sensing approaches for sleep-wake and sleep-stage classification using the most common feature-based ML and DL frameworks. Furthermore, a new ensemble architecture is introduced, outperforming all other models.

3.5.2. Transparency in Algorithm Development in Machine Learning for Sleep Health

All analyses were performed in MESA [210, 211], a publicly available dataset for which access can be requested through https://sleepdata.org/datasets/mesa. The experimental code is also available on GitHub: https://github.com/bzhai/multimodal_sleep_stage_benchmark.git. The open-source approach aims to create transparency and promote reproducibility in human sleep science, and encourage others to use the resource to develop novel, more accurate models that leverage multimodal data. Here this chapter provides an example of how the performance of well-established methods can be surpassed by an *ensemble architecture* of DL models of different window sizes. Similarly, the experimental results found that the model performance was influenced by optimal hyperparameter search, as reflected in the Appendix A.3. In particular, the experimental results observed that although there was no significant improvement in terms of accuracy and F_1 on the search space, certain patterns did emerge. For most CNNs and LSTMs, increasing the length of the sliding window improved performance (except for Task 1, sleep-wake only).

This chapter advocates for and demonstrates the value of including performance metrics beyond the conventional accuracy, specificity, precision, recall and F_1 scores. For instance, introducing time deviation metrics allowed us to understand what precisely each model over- or underestimated. This is of particular value for the translational applications of this chapter which may be implemented by HCI researchers, clinicians or epidemiologists and have an impact on the field of digital health. These metrics are more interpretable for non-machine learning experts who may seek to understand how certain inferences should be interpreted. Clear, interpretable measures that allow non-specialists to understand the limitations of models is critical both to the development of better study cohorts and to understanding the inferences made by these models.

3.5.3. Sleep Classification Performance by Task

Binary sleep-wake classification (Task 1) using actigraphy had been previously explored by Palotti et al. in the same cohort [205]. The experimental results corroborate this study, with the CNN architecture narrowly outperforming the LSTM architecture. Interestingly, the multimodal approach did not add much to this binary classification task when exploring conventional metrics but did yield a lower time deviation of overall sleep time than actigraphy alone. Models based only on cardiac signals had a slightly worse performance than both actigraphy alone and the multimodal approach, with accuracy estimates in the high 70s (79% for the LSTM (101)).

Task 2 consists of Wake, NREM and REM classification. This represents a valuable yet holistic overview of sleep stages and is what most free-living commercial devices aim to measure. Actigraphy and HR/HRV yielded an accuracy of around 74% and F_1 scores between 49% and 65%. The time deviation metrics demonstrated that the actigraphy cannot accurately determine REM sleep, whilst HR/HRV perform better. Both sensor modalities tend to overestimate time spent in NREM sleep. This finding also pertained to the multimodal approach, where NREM overestimation was the most error-prone estimate of the three states classified, at around 24 minutes mean deviation from gold-standard measures per participant on average when using the LSTM (101) and 24 minutes deviation when using the LSTM (51). NREM could be overestimated because it is the most common state among all participants on average, meaning that errors could be magnified. Accuracy estimates for the multimodal approach were in the high 70s for the majority of the classifiers, with LSTM (51) reaching 76% accuracy. These results are in line with the best performance previously reported in the literature. However, none of these previous studies had the scale and diversity that the MESA dataset offers [221].

Task 3 explored classification into Wake, REM, light sleep and deep sleep. In this classification, N1 and N2 were considered part of light sleep and N3 was classified as deep sleep. Actigraphy and HR/HRV reached accuracies of around 67% through LSTM (101). However, F_1 scores for the single modality approaches were between 35-50%. Similarly, both approaches overestimated time spent in light sleep and also struggled to pick up REM sleep. The multimodal approach outperformed the single modality approach with a higher accuracy of around 70% and an F_1 score of 52% for LSTM (51) (the highest performing model). Interestingly, LSTM (51) has a very strong performance at classifying wake and REM but struggles to discern light sleep and deep sleep, overestimating light sleep. There are two speculative reasons for this phenomenon. First, this may be caused by model bias, as LSTM predictions appear smoother in terms of sleep continuity. Secondly, it may also be due to the high prevalence of light sleep nested in the night causes the class imbalance.

Task 4 aimed to evaluate classifiers that followed AASM scoring rules of Wake, REM, N1, N2 and N3. This task is the most complex of the four, given the high level of granularity required and the imbalance severity between sleep stages increased. Actigraphy and HR/HRV performances at this task were poor, with F_1 scores ranging from 27-36%. Both heavily overestimated the most prevalent state, N2. The multimodal approach struggled to discern among the different NREM stages and, again, overestimated time spent in N2. Its performance on Wake and REM was much better but, intriguingly, worse than what had been observed in Task 3. Accuracy did not exceed 64% and F_1 scores were between 40-42%. A visual illustration of this task is presented in Figure 3.4, where overestimation of N2 can be observed, alongside how the model struggles to discern transitions between N1, N2 and N3.

Across all multistage classification tasks, LSTMs outperformed every other modelling approach. This is most likely due to its ability to learn temporal dependencies from longer window sizes, contrasting with CNN models which focus on local dependencies. This makes LSTMs a particularly attractive candidate for multistage sleep classification given the intrinsic

transitional nature of the task. The ensemble model approach serves as an example of how multistage classification benchmarks ought to be improved by new model architectures. The example approach is a rather simple one and thus only improves the performance marginally. Nevertheless, the results are promising. By incorporating different temporal domains and classifier types, these new models are able to pick up *nuances* that may have been tougher to identify by using a single convolutional or recurrent neural network.

Understanding sleep stage dynamics at a population level could be of value for digital health, epidemiology and clinical studies. The research on behavioural change (i.e., [209]) can take advantage of ubiquitous systems for sleep stage classification to recommend changes for better sleep hygiene and healthier sleep architectures.

3.5.4. Physiological Underpinnings of Classifiers and Sensor Modality Contributions

The classification tasks aimed to explore how the different modalities performed with regard to the level of granularity and detail generated. Physiologically, sleep stages are quite different and one objective of this chapter was to understand the individual contributions of each sensor, as well as model biases and preferences. Following the AASM staging convention:

1. N1 (the first stage of NREM sleep) is the stage in which the change between wakefulness and sleep occurs. During this stage, heart rate, breathing and eye movement slow, with occasional muscle twitches. Similarly, slow-wave activity starts to appear on the PSG's EEG signal.

2. N2 (the second stage of NREM sleep) is the transition period between light and deep sleep. Heartbeat and breathing slow, muscles relax even further and body temperature drops. This stage is the most repeated across all sleep cycles. Together with N1, it is often referred to as *light sleep*.

3. N3 (the third stage of NREM sleep) is often referred to as *deep sleep*. Heart rate and breathing are at their lowest, muscles are very relaxed and it is rare for the person to awaken during this stage. These changes are also observed on the PSG's EEG signal, where the lowest frequency and highest amplitude waves can be found. Together with N1 and N2, this stage constitutes NREM sleep.

4. Finally, as explored in the introduction, REM sleep occurs in a cyclical fashion, approximately every 90 minutes. Breathing is faster and irregular, heart rate increases and, in healthy people, the body is in a state of temporary paralysis that prevents sharp movements related to dreams.

Given the physiological differences between sleep stages, depending on the sensors used, the performance at classifying certain sleep stages may differ. This is of high importance when considering the deployment of these technologies in clinical settings or for the exploration of the association between sleep characteristics and disease end-points in population-based research. Understanding time spent at different stages over a long-term period is of great importance

for the greater sleep scientific community. For instance, during non-REM sleep, slow-wave activity has been shown to support memory consolidation [236] and reduce next-day anxiety [12]. In [236], for example, these slow-wave oscillations have been shown to affect the way the brain cerebrospinal fluid dynamics work, leading to oscillations in blood volume that draw this fluid across the blood-brain barrier.

This chapter used SHAP to further understand how different sensor features contribute to the individual classification of sleep stages. Through this method, it became apparent that whilst actigraphy features were the most informative for sleep-wake classification, when moving to multistage classification tasks, HR/HRV features were also important. This is reflected in Figure 3.7 where the top features contribute more to the model than the bottom ones, indicating their higher predictive power. For example, frequency domain features were very informative for recognizing non-REM sleep. Similarly, the application of this method to the different tasks allows for the direct comparison of feature importance across different levels of sleep architecture granularity. When exploring SHAP results, for Task 1, the results show that the most informative features came from actigraphy, although maximum HR and NN intervals were also notable contributors. Activity coming from the wrist actigraphy was particularly important for Wake classification. This finding carried through all 4 tasks and makes sense given the considerably higher amount of movement present during wake than in any sleep stage unless a sleep disorder is present. For Tasks 2–4, SHAP results helped us understand why the multimodal approach performs significantly better ($p < 0.001$) than individual sensors at sleep-stage classification. The results show that although HRV features are not particularly useful for the Wake classification, they added a lot of value to NREM and REM predictions. Interestingly, the empirical experimental results demonstrated that frequency domain HRV features were amongst the most informative for light and REM sleep classification, confirming the initial hypothesis derived from previous clinical reports [129, 212]. These findings emphasise the importance of including HR/HRV measurements combined with the movement for multistage classification tasks using wearable devices.

3.5.5. Summary

The strengths of this chapter derive from its novelty, population size and generalisability. It is the first systematic assessment of multistage sleep classification using non-obtrusive sensors. This makes an important contribution to the literature with potential applications for clinicians, researchers and the wellness industry. The population used is uniquely diverse, including a breadth of racial backgrounds, balanced sex and a representative sample of sleep-disordered participants. This enhances the generalisation capabilities of the findings in contrast to previous studies [215, 61].

In conclusion, this chapter introduces a systematic benchmark approach to sleep-wake and sleep stage classification using ML and DL approaches in single-modal and multimodal settings. This approach advocates for model transparency, alongside reproducibility by exploring these methods in the only open-access dataset, which includes participants with sleep disorders. The

findings indicate that multimodal approaches combining movement with HR and HRV data were a valuable tool for the monitoring of sleep stages when those stages were aggregated to the level of NREM, REM, and Wake. This chapter further provides information regarding the performance of specific algorithms and guidance regarding algorithm selection depending on the classification tasks. Moreover, this chapter introduces a deep ensemble model architecture which shows promising improvements in performance across the different multistage tasks explored. Overall, the findings highlight the promise of using wearable sensors as a low-burden, cheap and scalable approach for large, population-based studies.

The main purpose of this chapter is to investigate the feasibility of using a wearable device with a similar pattern to detect sleep and the performance of the baseline experiments. Based on this scope, this chapter only explored benchmark models, but more complex networks can be designed, for example, by training deeper architectures, adding residual-connections [158] or using attention mechanisms [162]. Given the temporal dependencies of these tasks, attention mechanisms may be well suited to improve model performance [237]. Another architecture that may yield interesting results is the addition of a dense layer to merge representations learned by RNNs and CNNs, also exploiting the unique contributions of each classifier (temporal representations by the RNNs and spatial representations by CNNs). Furthermore, this chapter has only explored a way to combine activity features and cardiac sensing features by concatenating them before sending them to a deep-learning neural network. **Chapter 5** will systematically address these outstanding research questions using advanced multimodal fusion approaches with deep learning architectures. The experiments in this chapter were performed on datasets collected using gold-standard equipment in sleep labs. It does not guarantee that wearables such as smart bands or smartwatches will achieve similar performance. Therefore, in **Chapter 5**, the author will also carry out a further study of sleep stage classification based on wearable devices.

Chapter 4. DisSleepNet: Disentanglement Learning for Personal Attribute-free Three-stage Sleep Classification Using Wearable Sensing Data

4.1. Introduction

In Chapter 3, the findings demonstrated that sleep stage monitoring can be achieved using activity counts and HRV features. Through a large-scale benchmark study, we tested various task settings and modality combinations. Empirical results suggested that three-stage sleep classification is a feasible task based on these two modalities. However, the MESA dataset was collected from medical studies that involved large numbers of ageing and disease participants. The cardiac sensing data could strongly be affected by personal health conditions, such as ageing, body mass index (BMI), and sleep breathing disorders (e.g., sleep apnoea), which are considered as personal attribute (PA).

If a deep learning model for sleep stage classification is trained on data from healthy people, it will inevitably suffer from performance degradation when tested on subjects with ageing, disease, or underlying health conditions, and the opposite process may hold true as well. The model works well when the difference in data distribution between training and testing datasets is small. In real-world scenarios, the target cohort data that the model attempts to predict may be completely inaccessible during training, let alone some annotated sleep data. Therefore, the latent representations learnt by these methods through supervised learning will inevitably be affected by these factors.

Several approaches could mitigate the influence of covariates during the training process, and most of those works have been investigated in computer vision applications, such as person re-identification [238], face recognition [239], gait recognition [240] and visual navigation [241]. An efficient approach is to use transfer learning methods to reduce the model performance degradation on different datasets. This way, the model is trained on the data collected from several similar datasets and fine-tuned on the data collected from the target distribution [242, 243]. A notable limitation of these methods is their need to fine-tune the model each time to deal with the covariates. Introducing auxiliary information and adversarial learning loss terms into the training pipeline is another efficient approach to extracting invariant representations of the individual attributes [244, 173, 177]. These discriminative architectures can learn effective representations that are invariant to dataset-specific PAs, thus increasing the robustness of the model [245, 246, 72]. For automated, large-scale sleep stage monitoring using wearable devices, these methods require PAs from the target population during the testing phase, which may be hard or costly to acquire in real-world scenarios, such as the apnoea hypopnoea index. In contrast,

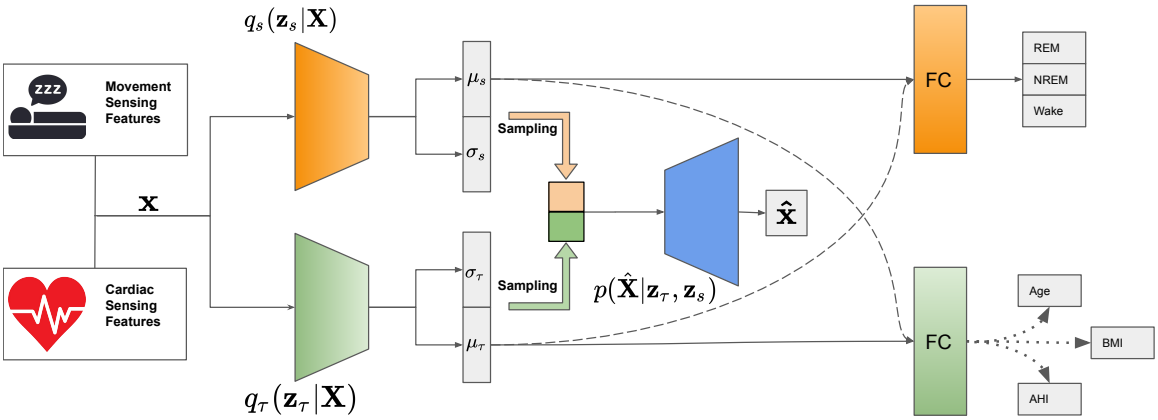


Figure 4.1 The proposed disentanglement model used in this study. The multimodal handcraft features $\mathbf{x}$ in the input data for the two probabilistic encoders that comprise a PA-specific encoder $q_\tau(\mathbf{z}_\tau|\mathbf{x})$ and a PA-free encoder $q_s(\mathbf{z}_s|\mathbf{x})$. The PA-specific encoder learns (μ_τ, σ_τ) that are more dependent on the PAs, and the PA-free encoder learns (μ_s, σ_s) related to sleep stage classification. The dotted line implies that each experiment will only disentangle one personal attribute at a time. Two disentanglers were introduced to further encourage the decorrelation of PA-specific and PA-free representations as shown by the dotted line. In the inference stage, only the PA-free encoder $q_s(\mathbf{z}_s|\mathbf{x})$, and the μ_s are used for sleep stage classification.

the approach proposed in this chapter is more flexible and does not require this information during the testing stage.

Another less-exploited but the potentially more promising approach might be to use disentanglement representation learning. This method can be used to learn less inter-correlated features where a single latent unit is sensitive to the changes in a single generative factor while being relatively invariant to the changes of the other factors [167]. Several recent works demonstrated that using disentangled representation learning with adversarial learning together can remove unwanted information to a certain extent during the training process [172–174]. It has been suggested that learning disentangled generative factors can be useful for a large variety of tasks. A disentangled representation could boost the performance of state-of-the-art (SOTA) machine learning approaches where although the models still struggle, humans excel [175]. These scenarios may require knowledge transfer, where the reasoning of the new data can be facilitated by recombining the learnt factors. Numerous applications in computer vision tasks (for example, human pose representation learning [176], gait recognition [240] and face recognition [177], etc.) have demonstrated that disentangled representation can achieve higher performance. However, many of these methods have been designed for computer vision tasks. For sleep stage classification, limited studies have explored the solutions to the covariate problem using EEG signals [154, 155].

For three-stage sleep monitoring using cardiac and actigraphy data, personal health conditions, for instance, atrial fibrillation and restless leg syndrome, can be considered as the factors that influence model generalisation. The modulation of the autonomic nervous system (ANS) regulates cardiovascular functions, which can be measured by ECGs during sleep onset and various sleep stages. HRV is a non-invasive indicator of ANS activity. Particularly, the HRV and brain imaging analysis demonstrated close connectivity between the autonomic cardiac modulations and activities of certain brain areas during REM sleep [54]. Nevertheless, the sleep

stage is one of many factors correlated to the modulation of the ANS. Other factors such as ageing, obesity that were measured with BMI, and sleep apnoea such as obstructive sleep apnoea (OSA) can influence the modulation of ANS [247, 248]. Additionally, for the ageing population, OSA and severe obesity are normally accompanied by autonomic dysfunction [249, 250]. These risk factors may affect the use of ECG data for sleep stage monitoring and cause conditions like abnormal ECG waves compared to healthy adults [247].

Furthermore, these risk factors or PAs may cause the models to be biased towards specific populations. Previous works have shown that the classification performance can be improved by combining the embedded representation of PAs and medical imaging representations [251]. However, some PAs may be unavailable during the testing phase and must be obtained from the medical diagnosis process. For instance, in some countries, sleep apnoea diagnosis often requires a PSG study [51]. Thus, a feasible solution is to learn the disentangled representation that does not require PAs in the testing phase, as the features related to PAs have been reduced during the training process.

The empirical evidence in this work demonstrated that traditional CNNs suffer from performance degradation when the distributions of training and test datasets are diverse. Thus, this chapter proposes DisSleepNet to learn the disentangled representations that contain fewer PA-related features, which can lead the model to SOTA performance, compared to the previous work [5].

Specifically, this work employed the beta-variational auto-encoder (VAE) structure to learn the variations across different personal attributes [168, 170]. The constraints proposed in beta-VAE can push the model to learn more efficient latent representations that are disentangled if the data contains some underlying factors that may be less correlated with each other. However, the PA-free latent representations may still contain PA-specific information as long as these features do not significantly influence the decision boundaries. To further separate the two representations, inspired by [252], two probabilistic encoders are proposed to produce less correlated representations in the feature space to encourage PA-free latent representations that separate it from PA-specific representations.

4.2. Related Work

4.2.1. Impacts of Personal Attributes on Sleep Stage Classification

Long-term sleep monitoring with ambulatory PSG devices is time-consuming and a very challenging task if monitored for more than two consecutive nights. Actigraphy and wearable ECG devices provide a viable solution to monitor the three sleep stages, as shown in several previous studies [5]. Autonomic nervous activity varies greatly from wakefulness to sleep, for example, sympathetic tone decreases gradually as sleep changes from wakefulness to NREM sleep and increases in the REM sleep stage [253, 254]. This evidence shows that the ANS system modulation can reflect the changes in sleep stages.

Modulation of the ANS is affected not only by different sleep stages but also heart disease/disorders such as the severity of obesity and respiratory disorders [255]. Comparison of the subjects with sleep apnoea and healthy subjects can show several varying HRV characteristics. OSA is one of the most prevalent sleep-breathing disorders and is characterised by upper-airway obstruction in adults during sleep [256, 51]. It affects HRV during sleep, which reflects in the periodic changes of heart rate [257].

The cyclical changes in heart rate are affected not only by the sleep stage but also by sleep apnoea. In [255], their study demonstrated that subjects with moderate and severe sleep apnoea had reduced high frequency (HF) values during sleep and wake after sleep onset (WASO), compared to healthy controls. Moreover, a study investigated by Wikulnd et al. [258] discovered a significant decrease in HF activity in patients with sleep apnoea when awake. This may indicate the presence of autonomic dysfunction in patients with sleep apnoea.

OSA can be seen in obese patients. BMI is a common unit to measure the severity of obesity (e.g., BMI $>30 kg/m^2$ is considered as obesity). Studies have found that obesity relates to the imbalance of ANS activity. This is characterised by an increase in parasympathetic tones and inappropriate activation of the sympathetic nervous system [259, 53]. This indicates that the adaptation of HR may be faulty in response to the changing requirements in obese subjects. In [53], the study results suggest that the HRV complexity exhibited significant reductions during NREM sleep in the obese patient group, compared to the control group.

4.2.2. Learning Disentangled Representation

Transfer learning in computer vision has been well-studied to alleviate the problem of model degradation caused by dataset differences. One of the most common solutions is to fine-tune the model with a proportion of data from the test dataset. Most of these works focus on methods for learning dataset-invariant feature representations [260–262].

Adversarial learning has been investigated for representation learning in various applications of computer vision. It has been mainly adopted in generative models such as generative adversarial network (GAN) [263] and VAEs [168]. Its objective is to minimise the divergence between the distribution of real and counterfeit images. To address distribution discrepancy, the key idea is to use a discriminator that classifies whether a data point is drawn from the source or target distributions [264]. This method encourages the learning of invariant features to the datasets through an adversarial objective to minimise the distance between the source and target distributions [265].

In human activity recognition (HAR), users usually perform the same class of activities differently due to their varied personal characteristics, such as habits, age and physical strength, and this makes the corresponding sensory data highly disparate. Bai et al. [266] designed a Discriminative Adversarial MUlti-view Network (DAMUN) for HAR to minimise the discrepancies between the wearable sensing data representations of different subjects. The model explicitly decreases the subject divergence, thereby ensuring that all subjects are presented in a consistent

representation space. It improves the generalisation ability of the feature extractor that can produce subject-invariant features.

Aside from HAR based on the accelerometer data, other bio-signals also carry the information of PAs to some extent. In [267], they proposed an adversarial inference approach in deep learning models to learn session-invariant person-discriminative representations that can improve the robustness of the model in terms of longitudinal usability. Zhao et al. [72] adopted an adversarial learning regime that could remove extraneous information that is specific to individuals or measurement conditions while retaining all information relevant to the sleep stage classification task. The adversary discriminator ensures conditional independence between the learned representation and the training dataset-specific characteristic(a.k.a, the measurement conditions and personal health conditions).

4.3. Method

This section will first present the technical description of the disentanglement methods used for the study. This includes the baseline methods, network structures, evaluation metrics and implementation details. this chapter then describes the experimental design on the largest sleep dataset with actigraphy and cardiac sensing data so far. This includes the data pre-processing, feature extraction methods and experimental settings to investigate the impacts of different personal attributes on three-stage sleep classification.

4.3.1. Problem Statement

Personal attributes, for example, age, obesity, and sleep apnoea may restrain the neural network to learn a more generalised representation. To alleviate these challenges, the work in this chapter proposes to use disentanglement learning to reduce the impact of PA on the model.

Following previous work [5], given an ith frame-wise multimodal time-series data that corresponds to a sleep epoch, a joint distribution can be denoted as $p(\mathbf{X}, y)$ on $\mathscr{X} \times \mathscr{Y}$, where $\mathscr{X}, \mathscr{Y}$ denote the sleep data instance space and sleep label instance space, and $\tau \in \{age, obesity, sleep\ apnoea\}$ denotes the PA, respectively. This chapter aims to train a deep learning model $f(\cdot)$, such that the trained model is less influenced by these three personal factors during the inference stage. In real-world scenarios, this setting is more challenging than the conventional transfer learning setting. Firstly, it's an arduous task to collect enough data for training that covers all different PA values, for example, the ageing population over 80 years. Secondly, for the unseen population, the model does not access any information other than the sensing data. This may be due to the limited resources of the sleep laboratory. Compared to standard machine learning settings, the proposed model does not need PAs during the inference process.

4.3.2. Personal Attribute-Free Feature and Personal Attribute-Specific Feature

Age, obesity, and sleep apnoea affect our cardio-respiratory systems and inevitably affect the use of body movement and heart-sensing data. To reduce the influence of these factors during model training, the model is designed to generate two kinds of representation features, namely PA-free features and PA-specific features. PA-free features included commonalities related to three-stage sleep. The PA-specific features could be attributable to age, obesity, and sleep apnoea severity.

For each training sample $\mathbf{X}^{(i)} \in \mathbb{R}^{C\times L}$, where C is the number of intermediate features (e.g., activity counts, and heart rate) and L is the temporal length. The network consists of two probabilistic encoders, namely $q_\tau(\mathbf{z}_\tau|\mathbf{X}^{(i)})$ and $q_s(\mathbf{z}_s|\mathbf{X}^{(i)})$ to extract latent features $\mathbf{z}_\tau$ and $\mathbf{z}_s$ that denote the PA-specific (subscript τ represents PAs) and PA-free latent variables (subscript s represents sleep stages). Compared to the deterministic auto-encoder structure, the VAE model these latent variables as joint distributions between the feature space $\mathbf{z} \in \mathscr{Z}$ and observation space $\mathbf{X} \in \mathscr{X}$ instead of a single value. This way, changes in data can be captured. Suppose we have a prior distribution $p(\mathbf{z})$ placed over the latent variables, a simple assumption for the prior is a multivariate Gaussian distribution $\mathscr{N}(0,\mathbf{I})$, which technically model the PA uncertainty.

In this chapter, as shown in Figure 4.1, the proposed network tries to make the latent representation $\mathbf{z}_s$ to be sensitive to sleep stage classification and more invariant to the PAs. Meanwhile, the design makes the PA-specific latent units $\mathbf{z}_\tau$, to be sensitive to changes in a single generative factor (e.g., changes in one dimension of the latent feature vector) and less sensitive to sleep stage classification. After computing the latent representation, the latent features are then concatenated and fed into the decoder $p(\hat{\mathbf{X}}|\mathbf{z}_\tau,\mathbf{z}_s)$ that can reconstruct the input $\mathbf{X}$.

As the generated data $\hat{\mathbf{X}}$ is influenced by variations of latent factors. If the true distribution of $p(\mathbf{X})$ can empirically be approximated on the training data set, then the training objective is to maximise the marginal likelihood as:

$$\mathbb{E}[\log p(\mathbf{X})] = \mathbb{E}_{p(\mathbf{X})}[\log \mathbb{E}_{p(Z)}[p(\mathbf{X}|\mathbf{z}_s,\mathbf{z}_\tau)]] \tag{4.1}$$

However, direct estimation of the likelihood typically becomes intractable. As it is often computationally challenging especially if the model has more than a couple of interconnected layers, whether in the directed or undirected graphical model frameworks [167]. A widely adopted method is to estimate the posterior via an amortised inference distribution $q(\mathbf{z}|\mathbf{X})$ and jointly optimise a lower bound of the log-likelihood as:

$$\begin{aligned}\mathscr{L}_{ELBO}(x) = &-\alpha D_{KL}(q_s(\mathbf{z}_s|\mathbf{X})||p(\mathbf{z}_s)) \\ &-\beta D_{KL}(q_\tau(\mathbf{z}_\tau|\mathbf{X})||p(\mathbf{z}_\tau)) + \mathbb{E}_{q_s(\mathbf{z}_s|\mathbf{X}),q_\tau(\mathbf{z}_\tau|\mathbf{X})}[\log(p(\mathbf{X}|\mathbf{z}_s,\mathbf{z}_\tau))]\end{aligned} \tag{4.2}$$

Where the first two terms calculate the Kullback–Leibler (KL) divergence between the estimated distribution of latent variable and the commonly assumed prior, which regulates the variables in the latent space as close as possible to Gaussian prior. The last item is the reconstruction error.

The above optimisation process is deemed to be unsupervised. This means that the learnt representation is not optimised for three-stage sleep classification. Additionally, the latent variables, $\mathbf{z}_\tau$ and $\mathbf{z}_s$, are designed to be "independent" of each other. However, in an unsupervised manner, using two encoders may be insufficient to guarantee that the disentangled latent features do not overlap. Therefore, incorporating the PAs $\tau^{(i)}$ and sleep stages $y^{(i)}$ to guide the learning of features during the training process can be a viable approach. To realise it, a disentangling regressor, D_τ, is introduced to regress the PAs, and a disentangling classifier D_y is introduced to predict sleep stages. This work considers the estimation of personal attributes as a regression task. The loss function is denoted as:

$$\mathscr{L}_{D_y,D_\tau}(\mathbf{z}_\tau,\mathbf{z}_s,y^{(i)},\tau^{(i)}) = \frac{1}{N}\sum_i[\ell_{CE}(y^{(i)},D_y(\mathbf{z}_s)) + \sum_\tau \ell_{MSE}(\tau^{(i)},D_\tau(\mathbf{z}_\tau))] \tag{4.3}$$

Where the τ represents the value of corresponding PA (e.g., age), N is the total number of training samples, and ℓ represents the loss function. In 4.3, the mean absolute error (MSE) loss function is adopted for the PA regression task and the cross-entropy (CE) loss is adopted for the sleep stage classification task.

4.3.3. The Independent Excitation Mechanism

D_y and D_τ enable the learning of the PA-specific latent features and PA-free latent features. The disentangled learning process still can not entirely reduce the feature-overlapping phenomenon. It is still possible that PA-free features contain PA-specific information if these features do not influence the decision boundary significantly, and the same may happen in PA-specific latent space as well. To further maximise the independence of these two features, inspired by [268, 252], their proposed method can encourage the $\mathbf{z}_s$ to be more sensitive to sleep stage classification. Meanwhile making $\mathbf{z}_\tau$ less irrelevant to the sleep stages. The independence excitation mechanism increases the error of D_τ when $\mathbf{z}_s$ is fed into D_τ. In another situation, the PA-free features were encouraged to be irrelevant to PAs. The independence excitation objective function for multivariate PA is denoted as:

$$\mathscr{L}_{IE_m}(\mathbf{z}_\tau,\mathbf{z}_s,y^{(i)},\tau^{(i)}) = -\frac{1}{N}\sum_i[\ell_{CE}(y^{(i)},D_y(\mathbf{z}_\tau)) + \sum_\tau \ell_{MSE}(\tau^{(i)},D_\tau(\mathbf{z}_s))] \tag{4.4}$$

The independence excitation objective function for univariate PA is denoted as:

$$\mathscr{L}_{IE_u}(\mathbf{z}_\tau,\mathbf{z}_s,y^{(i)},\tau^{(i)}) = -\frac{1}{N}\sum_i[\ell_{CE}(y^{(i)},D_y(\mathbf{z}_\tau)) + \ell_{MSE}(\tau^{(i)},D_\tau(\mathbf{z}_s))] \tag{4.5}$$

where the loss function of the first part of the equation is the cross-entropy loss, and the MSE loss is used for the second loss term.

4.3.4. Full Objective

All these objectives, i.e., representation disentanglement (L_{ELBO}), PA-specific and PA-free training and independent excitation can be used together, and in doing so the full objective function can be denoted as:

$$\mathscr{L} = \mathscr{L}_{ELBO} + \eta \mathscr{L}_{D_s,D_\tau} + \gamma \mathscr{L}_{IE} \tag{4.6}$$

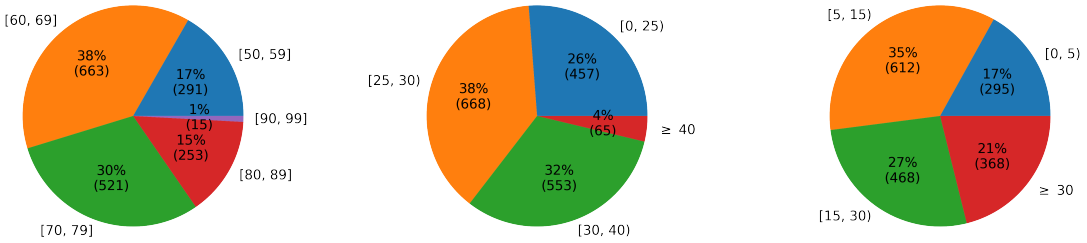
Where η and γ are positive parameters that control the balance between the independent excitation loss and the sleep stage classification and the PA regression tasks. This study fixed the weights during the entire training process in all experiments. The PA-specific latent features are not beneficial to the prediction task as they are not designed for three-stage sleep classification. Therefore, during the testing/inference stage, the test data samples $\mathbf{X}^{(j)}$ are fed into the probabilistic encoder $q(\mathbf{z}_s|\mathbf{X}^j)$ to obtain PA-free latent features $\mathbf{z}_s$. The $\mathbf{z}_s$ comprises two vectors $\mu_{\mathbf{z}_s}$ and $\sigma_{\mathbf{z}_s}$, where only the $\mu_{\mathbf{z}_s}$ is used for the classifier DC_y to predict the sleep stages.

4.3.5. Dataset Description

The Multi-Ethnic Study of Atherosclerosis (MESA) is a multi-site prospective study that includes 6,814 men and women. The study had 2,237 participants enrolled in the sleep exam, which included seven days of wrist-worn actigraphy; they underwent concurrent PSG for one night (wrist-worn actigraphy collected concurrently) [269] The actigraphy recorded the activity counts at 1/30Hz and ECG at 100Hz. The data processing pipeline is consistent with the previous chapter [5]. After the data pre-processing, 1,743 of 2,237 participants met the data quality criteria. The full details of the study setup, protocol and sampling rates are available at [5, 269]. According to the feature set used in previous research [5], this study used the same features, including activity counts and eight HRV features derived from the NN interval data in each sleep epoch [5]. The feature set comprises the mean NNi, standard derivation of NN interval (SDNN), NN interval differences (SDSD), very low frequency (VLF), low frequency (LF), high frequency (HF) bands, low frequency to high frequency ratio (LF/HF) and total power.

For each sleep epoch, the intermediate feature vector is constructed based on eight HRV features and the activity counts (a scalar value per sleep epoch). In addition to these handcraft features, the personal attributes are extracted, which include age, BMI, and apnoea hypopnoea index (AHI).

For the **age** attribute, the minimum and maximum ages were 54 and 94. The age bin was set to every 10 years, which is a commonly used bin size in epidemiological sleep research [270]. As in Figure 4.2 (a), the majority of participants were between the ages of 54 and 79, which represents 85% of the total number of subjects. The smallest group was between 90 and 99 years. Since there were too few participants in this group, this group has been combined with the group whose ages were between 80 and 99 years. To investigate the effect of age on model performance, two groups with a wide span were selected: subjects for training were between 50 and 69 years and for testing were between 80 and 95 years old.



(a) Age **(b)** BMI **(c)** AHI

Figure 4.2 The number of subjects in each subgroup is organised by personal attributes. (a) the proportion of subjects group by the age attributes, (b) the proportion of subjects group by the BMI attribute, (c) the proportion of subjects group by the OSA severity attribute which is measured in AHI.

In terms of **obesity**, the subjects were segmented into four groups according to the settings of the National Heart, Lung, and Blood Institute, on the basis of BMI as follows: no obesity where BMI was less than 25; grade 1 obesity where BMI was between 25 and 30; grade 2 obesity where BMI was between 30 and 40; and grade 3 obesity where BMI was 40 or greater [271]. These standard categories have been increasingly used in public health studies [272]. To understand the negative impact of BMI on the model, the subjects were further divided into non-obesity and obese groups and used their BMI score as the corresponding PA value.

For **sleep apnoea**, AHI is commonly used to measure the severity of sleep breathing disorders. According to the MESA study protocol, the dataset was segmented into four subgroups corresponding to the current definitions used by Medicare for reimbursement and includes all apnoeas plus hypopnoeas with a $\geq$4% desaturation [271]. The clinical cutoffs rather than distributional ones were used, a set of commonly used AHI clinical cut-off points (number of events per hour) were adopted to divide the subjects into four groups that were defined as follows: no sleep-disordered breathing where AHI was <5; a mild severity where AHI was between 5 and 15; a moderate severity where AHI was between 15 and 30; and a severe where AHI was 30 or greater [273]. The AHI scores were used in regression tasks.

For the segmentation of each PA, the subjects have been divided into training and testing datasets based on the PA combinations. The training dataset is further randomly divided so that 20% of the samples are used as the validation set and the remaining 80% of the samples are used for training. The validation set is used to select the best model for testing.

4.3.6. Baselines and Implementation Details

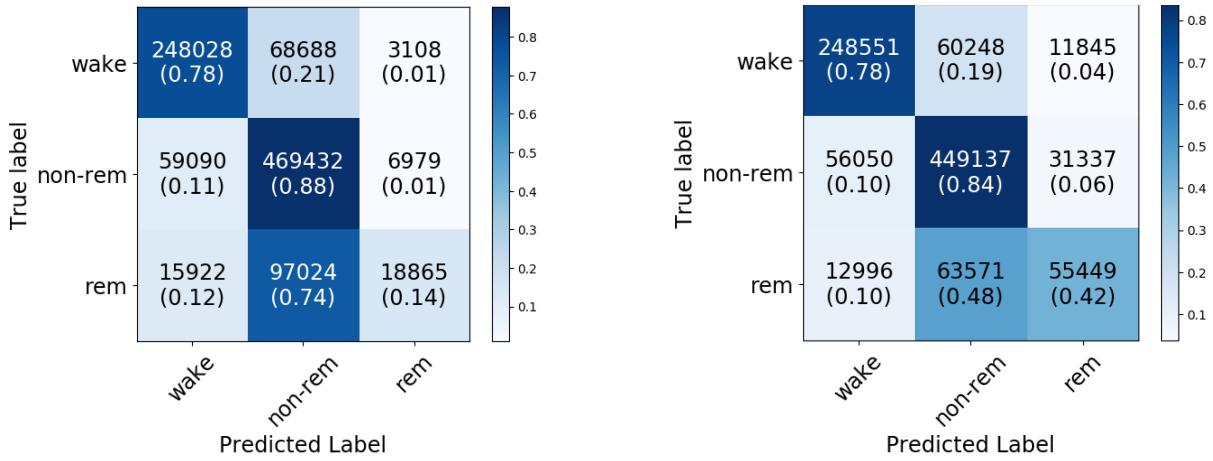
This work adopted a four-layer CNN used in [5] based on the window length of 101 as a closely related baseline. The same network structure was also used for the encoders q_s and q_z. Furthermore, the input feature set includes activity counts and 8 HRV features. Due to the class imbalance problem in the MESA dataset, the mean F_1 (average over classes) is used as the measurement of the robustness of the model. The details of the backbone network structure are shown in Figure 5.2. For the hyper-parameters of α, β, η, and γ, the hyperparameter values are tested based on the following settings: $\eta \in 1e4, 1000, 100, 10$, $\gamma \in 1e4, 1000, 100, 10$,

$\beta \in 0.002, 0.02, 0.2, 2, 10, 100$, and $\alpha \in 0.1, 1, 10, 100$. The hyperparameter set to the values of $\eta = 1000, \gamma = 1000, \beta = 10, \alpha = 10$ for the DisSleepNet model. The batch size was set to 1024 to speed up the training process, changing the batch size did not make any significant difference. The Adam optimiser with a learning rate of 1e-4 is utilised, the validation set is used to select the best model for testing and the maximum training epoch is set to 20. The VAE network with two auxiliary tasks is also considered as an additional baseline mode.

4.4. Results

This section first explains how each factor may influence the model performance, experiments were conducted on each PA, and the results of each model are shown in Table 4.1 4.2, and 4.3 Secondly, to understand the effect of the three personal factors combined namely age, obesity and sleep apnoea, and the results are shown in Table 4.4. The random seed for all experiments was fixed, and each experimental setting was run ten times to calculate the mean F_1 score. The t-test was conducted to compare the performance of DisSleepNet and the baseline methods.

4.4.1. Experimental Results for Sleep Apnoea



(a) CNN **(b)** DisSleepNet

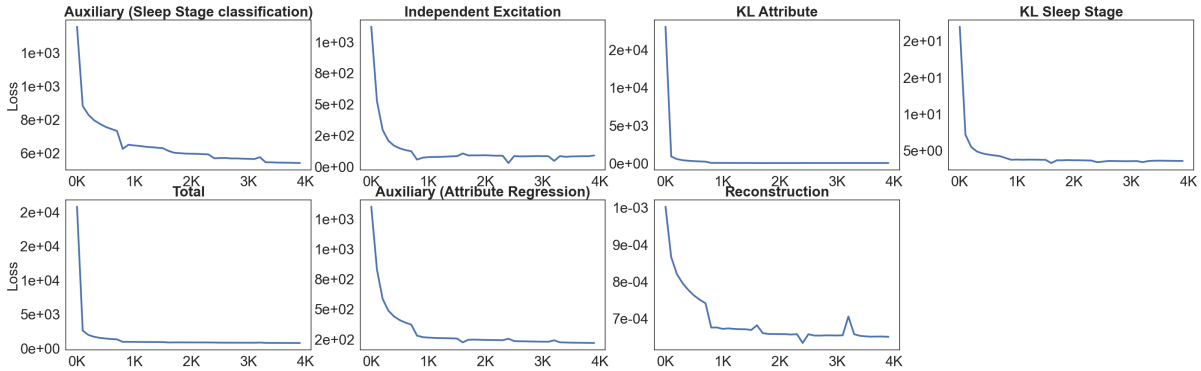
Figure 4.3 The confusion matrices derived from the prediction results of moderate/severe OSA subjects' data, where the models were trained on the healthy and mild OSA subjects' data

OSA is an important factor negatively affecting cardiorespiratory function in clinical sleep medicine. As shown in Table 4.1, the results demonstrated that the use of the DisSleepNet can achieve a higher improvement on the mean F_1 by 7.68 ($p < 0.05$) if it was trained on the normal and mild groups and tested on the moderate and severe group subjects. In contrast, training the DisSleepNet on the moderate and severe groups and testing it on the normal and mild groups improved F_1 by 3.96 ($p < 0.05$).

Figure 4.4 shows the change in the loss in the setting of training on patient data with AHI $\geq$ 15 and testing on patients with AHI < 15. This setup achieved the biggest improvement.

Training AHI Range (# of subjects)	Testing AHI Range (# of subjects)	Mean F_1 ($\mu \pm \sigma$)		
		CNN	VAE + Ancillary Tasks	DisSleepNet
AHI<15 OSA (907)	AHI≥15 OSA (836)	60.99 ± 0.17	59.41 ± 0.29	**64.95 ± 0.31**
AHI≥15 OSA (836)	AHI<15 OSA (907)	60.98 ± 0.64	65.46 ± 0.31	**68.66 ± 0.34**

Table 4.1 Three-stage sleep classification prediction results based on the obesity groups (mean ± std)



(a) CNN

Figure 4.4 Reduction of different types of loss during training to disentangle AHI attributes (AHI ≥15). (Note: The x-axis represents the number of per thousand batches trained)

4.4.2. Experimental Results for Age

To better understand the influence of age on the model generalisation, two settings were tested. The first setting is to train the network on data from subjects aged between 50 and 69 and tested the model on a population aged between 80 and 99 years. Several previous studies have suggested that sleep patterns in older adults can vary widely within a 10-year gap, compared to younger adults [274]. Additionally, the inverse setting was tested using subjects aged between 80 and 99 years for the training data set and subjects between 50 and 69 years for the testing data set. As shown in the second row of 4.2, the DisSleepNet achieved the highest mean F_1 score.

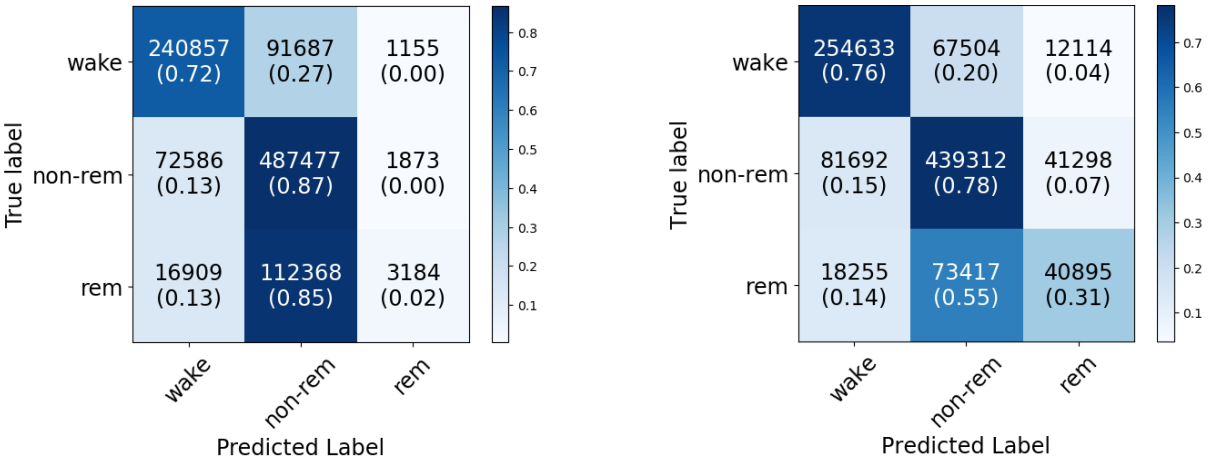
In a real-world scenario, the data from the group aged between 50 and 79 years are more relatively convenient to collect than the group aged between 80 and 99 years. This way, the entire dataset was used for the study. Thus, the confusion matrices demonstrated results trained on the group aged between 50 and 79 years and tested on the group aged between 80 and 99 years, as shown in Figure 4.5.

The confusion matrices derived from the last row experiment in Figure 4.5 show that the DisSleepNet improved the classification performance of REM sleep and wake stages.

Training Age Range (# of subjects)	Testing Age Range (# of subjects)	Mean F_1 ($\mu \pm \sigma$)		
		CNN	VAE + Ancillary Tasks	DisSleepNet
50s-69s (954)	80s-99s (268)	61.26 ± 0.31	60.36 ± 0.1	**62.1 ± 0.39**
80s-99s (268)	50s-69s (954)	59.73 ± 0.21	58.05 ± 0.74	**62.16 ± 0.19**
60s-99s (1452)	50s-59s (291)	69.25 ± 0.43	68.9 ± 0.11	**70.9 ± 0.22**
50s-79s (1475)	80s-99s (268)	63.01 ± 0.36	62.1 ± 0.11	**64.44 ± 0.09**

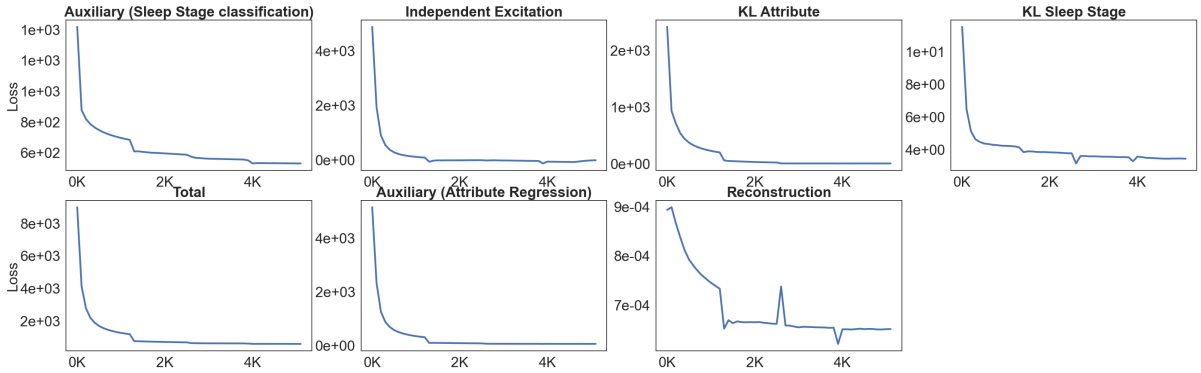
Table 4.2 Three-stage sleep classification prediction results based on various age groups

The third setting is to train the model on data from the group aged between 60 and 99 years and test it on the group aged between 50 and 59 years. The rationale behind this setup is to



(a) CNN **(b)** DisSleepNet

Figure 4.5 The confusion matrices derived from the prediction results of the group aged between 80 and 90 years, where the models are trained on the group aged between 50 and 79 years



(a) CNN

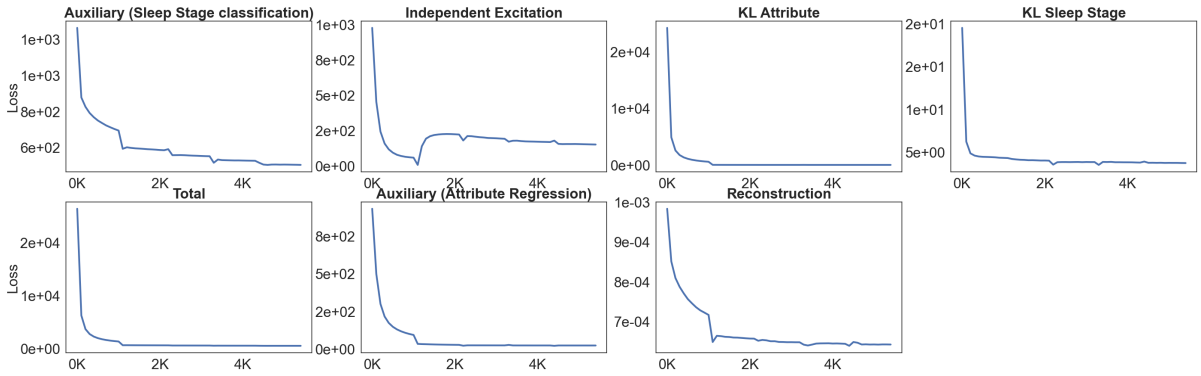
Figure 4.6 Reduction of different types of loss during training to disentangle age attributes (50s-79s). The x-axis represents the number of per thousand batches trained.

understand the effect of age factors on the model training of large datasets. As can be seen in the third row of the table 4.2, models trained on the group aged between 60 and 99 years achieved a higher F_1 score, compared to the training on the group aged between 50 and 79 years and testing on the group aged between 80 and 90 years. Finally, the t-test results between the DisSleepNet and the CNN revealed that all the improvements are statistically significant ($p < 0.05$). Figure 4.6 shows the change in the loss in the setting of training on patient data with age $\geq$ 60s and testing on patients with age between 50s and 59s. This setup achieved the highest average F_1 score and modest improvement.

4.4.3. Experimental Results for Obesity

Training Obesity Range (# of subjects)	Test Obesity Range (# of subjects)	Mean F_1 ($\mu \pm \sigma$)		
		CNN	VAE + Ancillary Tasks	DisSleepNet
BMI < 25 (457)	BMI $\geq$ 25 (1286)	60.97 $\pm$ 0.38	61.74 $\pm$ 0.34	**64.88 $\pm$ 0.44**
BMI $\geq$ 25 (1286)	BMI < 25 (457)	66.8 $\pm$ 0.24	66.38 $\pm$ 0.3	**68.95 $\pm$ 0.17**

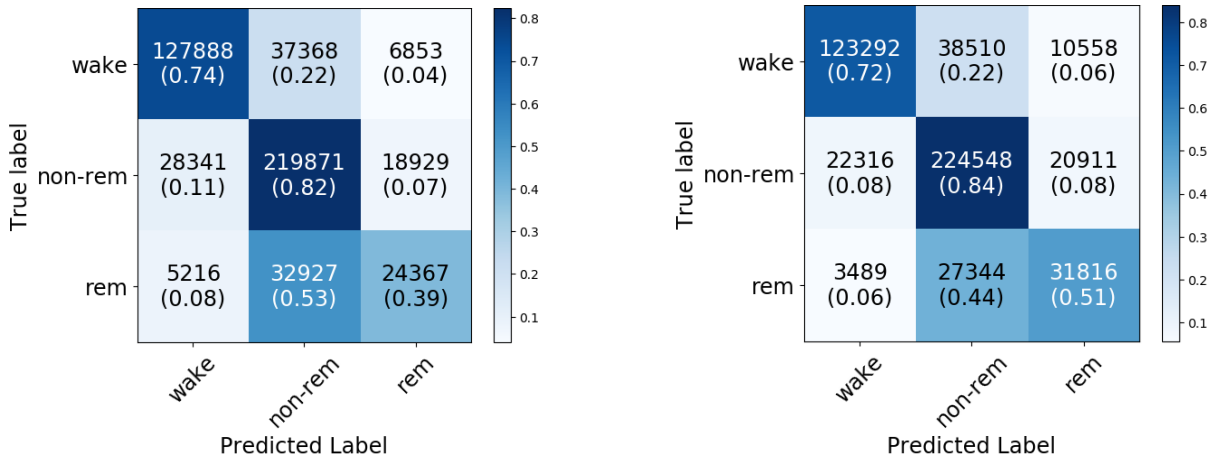
Table 4.3 Three-stage sleep classification prediction results based on the obese subjects (mean $\pm$ std)



(a) CNN

Figure 4.7 Reduction of different types of loss during training to disentangle BMI attributes (BMI $\geq$ 25). The x-axis represents the number of per thousand batches trained.

To examine whether BMI affects the performance of the three-stage sleep classification, the MESA dataset was divided into two groups, namely a normal group and an obese group. Table 4.3 shows that when training on the healthy group and testing on the obese group, the improvement of mean F_1 is higher than when training on the obese group and testing on the healthy group. The first setting achieved a 3.91 ($p < 0.05$) improvement. The mean F_1 of the second setting achieved a 2.15 ($p < 0.05$) improvement.



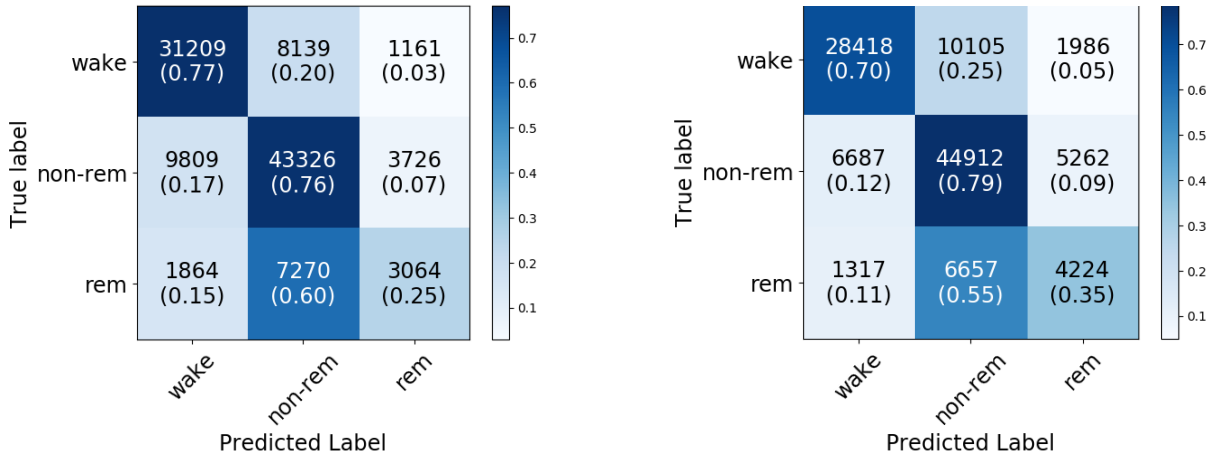
(a) CNN **(b)** DisSleepNet

Figure 4.8 The confusion matrices derived from the prediction results of obese subjects' data, where the models are trained on the subjects with a normal range of BMI.

Figure 4.7 shows the change in the loss in the setting of training on patient data with BMI $\leq$ 25 and testing on patients with BMI > 25. This setup achieved the largest mean F_1 score. Figure 4.8 shows the confusion matrices for understanding the performance improvements by removing the BMI factor. These figures were derived from the models that were trained on the subjects with a normal range of BMI and tested on the obese subjects. The performance improvement is mainly due to the DisSleepNet alleviating the misclassification of REM sleep.

Training PAs Range (# of subjects)	Testing PAs Range (# of subjects)	Mean F1 ($\mu \pm \sigma$)		
		CNN	VAE + Ancillary Tasks	DisSleepNet
Age<=79, BMI<30, AHI<15 (556)	Age>79, BMI≥30, AHI≥15 (40)	59.59 ± 0.39	57.96 ± 0.44	**59.67 ± 0.05**
Age<=79, BMI≥30, AHI≥15 (326)	Age>79, BMI<30, AHI<15 (99)	60.38 ± 0.43	51.66 ± 0.74	**61.6 ± 0.26**

Table 4.4 Three-stage sleep classification prediction results based on joint disentanglement of gender, age, BMI and AHI



(a) CNN **(b)** DisSleepNet

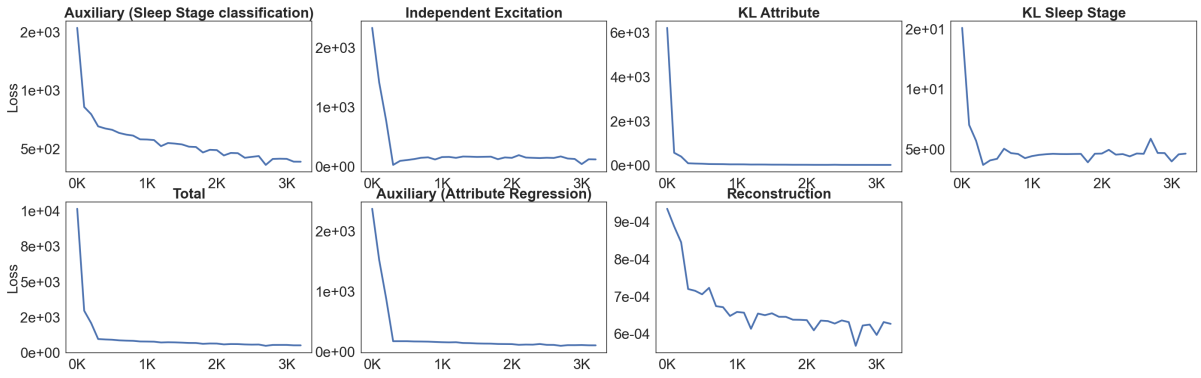
Figure 4.9 The confusion matrices were derived from the second joint disentanglement setting

4.4.4. Experimental Results for Joint Disentanglement of PAs

The joint disentanglement experiments aim to understand how multiple PAs affect model performance. In this work, two scenarios were tested. The first scenario assumes that the subjects in the training dataset are aged between 50 and 79 years, with a BMI less than 30 (normal and grade 1 obesity), and an AHI between 0 and 15 (normal and mild sleep apnoea). Under this setting, the test dataset comprised an elderly population with high BMI and AHI scores. This chapter did not test the settings where the training dataset comprised only healthy people, as the number of subjects was too small. Moreover, an alternative setting was tested where the training dataset comprised moderate-to-severe obesity and moderate-to-severe OSA populations, and the test dataset comprised of healthy and mild subjects. As shown in Table 4.4, the DisSleepNet could improve the mean F_1 score by up to 1.22 ($p < 0.05$) compared to the baseline method. The results show that the effect of PAs can be jointly reduced using the DisSleepNet, although the improvement is marginal. Figure 4.10 shows the change in the loss in the second set. This setup provides the biggest improvement.

4.5. Discussion

As can be seen in the results, the DisSleepNet has achieved higher or comparable mean F1 scores on all settings of PAs. This greatly illustrates that the use of DisSleepNet could reduce the influence of PAs on the model. For the age factor, the models were tested on the data collected from very distinct groups. Particularly, when training models on higher age groups and testing them on lower age groups, we see that the DisSleepNet achieves larger mean F_1 margins than the



(a) CNN

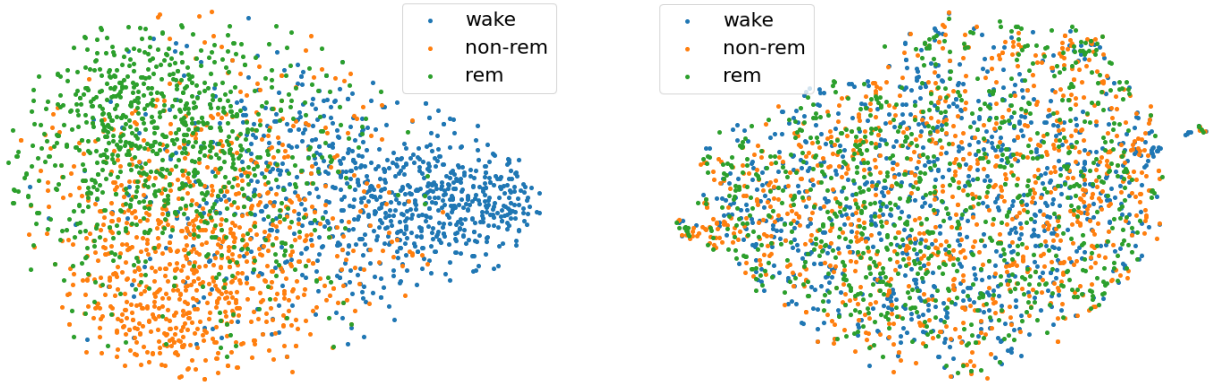
Figure 4.10 Reduction of different types of loss during training to jointly disentangle attributes. The x-axis represents the number of per thousand batches trained.

models trained on the reverse setting. In Figure 4.5, the confusion matrices demonstrated that the improvement actually came from correcting the misclassified NREM sleep as REM sleep. This may indicate that different age values may affect the model performance differently.

Obesity is usually associated with several chronic diseases, such as hypertension, cardiovascular diseases and diabetes. Further, it is associated with shorter sleep duration and increased sleep disturbance. In Table 4.3, DisSleepNet can reduce the effect of obesity on model performance as tested by these various settings. Training DisSleepNet on the healthy group improves model performance more than training on the obese group. One possible reason is that the training sample size of the first setting is smaller than that of the second setting, and future work should investigate different datasets and possibly use sufficient training samples from healthy subjects.

OSA is an important factor that affects model performance when the test dataset does not contain patients with the same severity of OSA. As several previous studies have shown, model test performance may degrade if the model is trained on a dataset with limited OSA severity patients [275]. The baseline model and DisSleepNet trained on the normal and mild OSA groups achieved higher performance on the test dataset, compared to training on the moderate/severe OSA group. The improvement of the mean F_1 score in this study corroborates previous studies [6], which suggested that sleep breath disorders could negatively influence the model performance. By comparing different PAs, the empirical evidence suggests that not all PAs have the same effect on the model performance. This work showed that sleep apnoea had the largest negative impact on the model, followed by BMI, and age had the least impact on model performance.

The proposed DisSleepNet can successfully learn the representation that is less influenced by PAs. The t-distributed stochastic neighbour embedding (t-SNE) embedding of the PA-free representation, $\mathbf{z}_s$, and the PA-specific representation, $\mathbf{z}_\tau$, are shown in Figure 4.11. Figure 4.11 (b) depicts that no obvious pattern divides the representation into clusters, compared to the PA-free representation shown in Figure 4.11 (a), where the clusters of embedding are more distinct and samples with the same sleep stage tend to be grouped into the same cluster. The number of clusters is the same as the number of sleep stages.



(a) t-SNE plot of PA-free features

(b) t-SNE plot of PA-specific features

Figure 4.11 Visualisation of the t-SNE embedding of learnt PA-free features $\mathbf{z}_s$ (a) and PA-specific features $\mathbf{z}_\tau$. Each sleep stage is represented by a distinct colour.

4.6. Summary

In this chapter, the author proposed a novel disentangling network to reduce the influence of PAs on three-stage sleep classification. The proposed method effectively makes the model less susceptible to PA by disentangling the representation into PA-specific and PA-free features. Empirical results show that obesity and sleep apnoea is common challenging factors that affect model performance. The proposed method outperformed the baseline network under various different settings on the MESA dataset. In addition, this study set out to explore the effects of three PAs on model performance. Insights from this study may help health researchers in the ubiquitous computing community to develop deep learning models that are less affected by individual attributes or covariates.

Chapter 5. Ubi-SleepNet: Advanced Multimodal Fusion Techniques for Three-stage Sleep Classification using Ubiquitous Sensing

5.1. Introduction

In chapter 3, experimental results suggested the feasibility of using these two modalities for three-stage (wake/REM/NREM) sleep classification, and the findings corroborate sleep physiology studies [52, 126] NREM sleep was not deemed to be easily separated into N1/N2/N3 without employing EEG signals. The results of the three-stage sleep classification look promising. However, the network architecture and the way of combining the two modalities are only tested on baseline methods and may not achieve optimal performance. Moreover, the HRV features derived from the MESA dataset are based on RR intervals that are calculated from ECG signal which is still expensive and impractical for long-term sleep monitoring. Nowadays, many consumer wearable sensing devices are available on the market for entertainment and health self-tracking purposes. These devices often sense human activities using accelerometers [50], photoplethysmography (PPG) [6], pressure sensors [58] and radio signals [57, 59], etc. Among them, cardiac and movement (upper limb) sensing data are considered promising modalities in terms of reliability and availability. They can be easily collected from lightweight research/consumer-grade devices (e.g., Apple Watch [6]).

The easy-to-collect nature of cardiac and movement sensing data provided a scalable method for large-scale and long-term sleep monitoring. Longitudinal sleep monitoring with accurate details in (three) sleep stages is meaningful to health and medical research. For instance, deep NREM sleep (or slow wave sleep - SWS) is known to be the most "restorative" sleep stage, which controls hormonal changes that affect glucose regulation [42]. Long-term reduction in NREM sleep may adversely affect glucose homeostasis and increase the risk of type 2 diabetes [43]. REM sleep dysregulation has played a central role in depression and Parkinson's studies [44, 45]. The phenomenon includes reduced REM sleep latency, along with increased REM sleep duration and REM sleep density, which have been considered to be an objective indicator of depressive disorder and inversely correlated to its severity [46, 47, 34]. The increased health research density in digital phenotype by using inexpensive, mass-produced consumer wearables demand reliable algorithms that can classify sleep stages in longitudinal settings [48].

Benchmarks in chapter 3 and a study by [6] used cardiac and movement sensing data for three-stage sleep classification based on publicly accessible sleep datasets. Both works used very basic multimodal fusion techniques (i.e., feature concatenation [6, 5]) in neural networks, which tested the feasibility of classifying three-stage sleep. However, the models used in these benchmark studies did not achieve good performance due to overestimating NREM sleep time

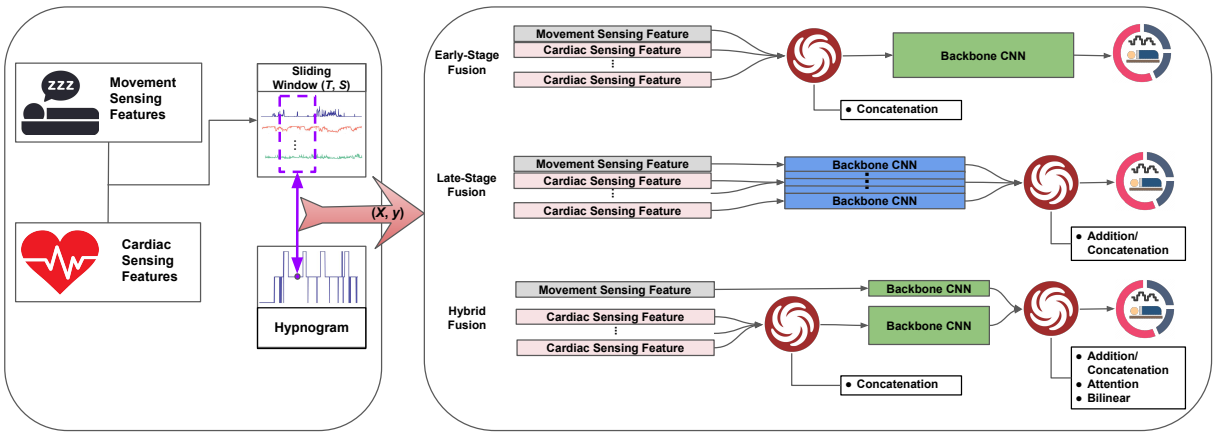


Figure 5.1 An overview of the three-stage sleep classification system. Features were extracted for each sleep epoch (30s). The sliding window method divides the sleep data into multiple segments with window length T and stride S, where $T = 101$, and $S = 1$.

and underestimating wake time. The simple fusion technique may not fully utilise the advantages of multimodal data, especially in heterogeneous multimodal data scenarios. Given that, it is desirable to explore advanced fusion techniques to boost performance further.

Firstly, this chapter systematically studied three fusion strategies for three-stage sleep classification, including early-stage fusion, late-stage fusion and hybrid fusion, to answer the question, *"At what stage should the cardiac and movement sensing representation be merged?"*

Secondly, this chapter employed three fusion methods (simple operations, attention mechanism, tensor methods) to answer the question, *"How to better combine cardiac and movement sensing representations?"* The simple baseline operations (concatenation and addition) as well as the advanced fusion methods (the attention mechanism-based method [180] and bi-linear pooling-based method [276]) were studied. The pipeline of this study is demonstrated in Figure 5.1.

These fusion techniques were comprehensively evaluated on two public datasets, which are the Apple Watch dataset [6] and the Multi-Ethnic Study of Atherosclerosis (MESA) dataset [269, 5, 271]. The Apple Watch dataset includes cardiac and movement signals collected from consumer-grade devices from a cohort of 31 young and healthy adults. For the MESA dataset, only the cardiac and movement sensing signals were used, which can be acquired from research-grade devices. The dataset consists of 1743 subjects from the ageing population.

For these two representative datasets, the results suggested that three-stage sleep classification can be reliably achieved by employing advanced fusion techniques on the cardiac and movement sensing data, which can be easily acquired from consumer/research-grade devices. Several models developed in this chapter achieved state-of-the-art performance for three-stage sleep classification. This chapter also evaluated the module parameter size and its corresponding inference time, which may play a vital role in ubiquitous computing applications.

Moreover, this chapter also investigated a visualisation method to explore the decision-making process of the multimodal fusion model for three-stage sleep classification. The exploratory user research demonstrated that the gradient class activation map (Grad-CAM) [3] based sleep data

visualisation can be understood and used by humans, which facilitates the transparency of using DL in sleep health research.

The work in this chapter contributes to the long-term non-intrusive three-stage sleep monitoring solution that may be deployed with mass-produced and inexpensive consumer-grade wearables, which may potentially be used for large-scale population-based sleep health studies and long-term sleep self-tracking.

5.2. Ubiquitous Sensing Techniques for Sleep Monitoring

5.2.1. Cardiac Activities and Sleep Physiology

The autonomic nervous system (ANS) and sleep are closely related in anatomical, physiological, and neurochemical bases [277]. Cardiovascular autonomous control plays an essential role in sleep, and it will be different when transitioning to different sleep stages. The modulation of the ANS regulates cardiovascular functions during sleep onset and sleep stages [125, 126].HRV analysis is a classical tool for ANS analysis. Research on HRV in sleep stages noted that REM sleep was characterised by a likely sympathetic predominance, while NREM sleep followed an opposite trend [130–132]. The transition between Wake, NREM and REM sleep is accompanied by changes of several HRV characters, such as the HR, Low-Frequency (LF) power, High-Frequency (HF) power and LF/HF ratio [130, 52, 131].

Not all sleep stages are associated with brain activity. A study conducted by Desseilles et al. [52] through HRV and brain imaging analysis found close connectivity between autonomic cardiac modulations and the activity of certain brain areas during REM sleep. There is no conclusive connectivity between the brain and cardiac activity during NREM sleep. Therefore, it may not be easy to discern each NREM sleep stage accurately without EEG signals.

5.2.2. Consumer and Research-grade Wearables for Sleep Monitoring

Traditionally, gold-standard human sleep assessment was conducted in laboratory settings using PSG, which commonly involved EEG, EMG and EOG, as shown in Chapter 2. It is impractical to measure sleep using this method for more than two consecutive nights. Recent advances in miniaturised sensing technologies have enabled the deployment of these simplified array EEG devices in clinical trials. Compared to consumer wearables such as smartwatches, these devices are often expensive and still uncomfortable to wear for long-term sleep monitoring.

Consumer-grade wearable devices with diverse modalities offer a potential solution to ambulatory sleep tracking. Such sensors provide valuable, inexpensive, unobtrusive measurement tools to collect biological signals. Many of these wearables can communicate with smartphones, facilitating data collection and storage during large-scale population studies. Therefore, exploring the use of consumer-grade wearables in sleep and health studies becomes prevalent as the HR/HRV data and movement sensing data are generally available on these wearables [6].

Many leading consumer products such as Fitbit^{TM} and Xiaomi^{TM} band provide sleep stage tracking services. However, these consumer products commonly lack minimal validation, with

poor algorithm transparency on data processing/sleep stage classification, resulting in these devices being precluded in clinical, research, or occupational settings [278]. Nevertheless, another consumer product, Apple Watch, provides access to the accelerometer data and heart rate data, making it feasible to develop an algorithm for sleep health studies and self-trackers.

The sleep stage classification based on ECG/PPG signals has also been investigated by [279, 280]. The results demonstrated promising performance. However, PPG data is generally unavailable on many consumer wearables, and ECG requires the skin electrodes to be placed near the heart. Collecting these raw signals may require research-grade wearables (e.g., Empatica™ E4), which demand additional financial costs for daily sleep monitoring.

Several previously published studies demonstrated that using HR/HRV features and movement sensing together could discern three sleep stages and achieved promising results [194, 6, 5]. Heterogeneous modalities may carry supplementary information for sleep stage classification. There is still much to be understood regarding how to construct this fusion architecture and which fusion method will be the most effective for sleep-stage classification. This work adds to this knowledge. Exploring multimodal fusion strategies and methods to better integrate different physiological signals is of great significance for health research and self-monitoring of sleep using ubiquitous computing technology.

5.3. Advanced Fusion Techniques for Three-stage Sleep Classification

This section will first discuss the current progress of multimodal fusion strategies and methods and their applications in sleep monitoring. Secondly, this section will then presents the study structure, followed by a technical description of three fusion strategies (early-stage, late-stage and hybrid fusion) and three methods (simple operation, attention mechanism and tensor-based method)

5.3.1. Overview of Multimodal Fusion

Multimodal fusion in machine learning has been extensively studied in pattern recognition applications, such as in image and video captioning[179], visual question answering [180], audio-visual speech recognition [149] and emotion recognition [181]. In the field of ubiquitous computing, multimodal fusion has also been adopted for human activity recognition [182], sleep stage classification [5], fatigue assessment [183] and person identification [184]. The simple concatenation method was commonly adopted in these studies to combine the raw inputs or combine the representations obtained from the pre-trained model of each modality [192]. Other researchers explored more advanced fusion methods, such as the attention-based fusion scheme for human activity recognition [182]

For monitoring sleep, several previous works achieved promising results for sleep stage classification by concatenating multimodal intermediate features and feeding them into DL models [5, 194]. However, these studies focus on the choice of modalities rather than the fusion

techniques. Different modalities may contain complementary information. It is difficult to explicitly identify the best suitable cross-modal fusion architectures.

In terms of movement sensing and cardiac sensing, they are different in signal-to-noise ratio, data generation process and measurement frequency. Moreover, the activity count is better in sleep/wakefulness classification, but it is difficult to discern different sleep stages [50]. For healthy adults, the difference in heart rate variability between REM sleep and wake is less than the difference in NREM and REM sleep [52].

The choice of fusion strategy and fusion method may thus influence the model classification performance. In recent years, the DL-based computational models have outperformed shallow machine learning models for sleep stage classifications, not only on unimodal data but also on the multimodal data [5, 281, 153]. Therefore, this work will only focus on multimodal fusion techniques based on DL networks.

5.3.2. Problem Statement

Based on the movement sensing and cardiac sensing data, the goal of this work is to comprehensively study how to use advanced fusion techniques to reliably classify three-stage sleep. As demonstrated in Fig. 5.1, a sliding window method was adopted with window size T and stride S to segment sleep recordings into frames. In each frame, one common approach is to extract the handcraft features (e.g. heart rate features that were deemed to be intermediate / mid-level features) from each sleep epoch that can provide physiologically meaningful features to the model [5, 280], or to use neural networks to extract the deep features. The time steps t represents one sleep epoch (i.e. every 30 seconds). Given that, this work aims to map the data in a sliding window to a sleep stage that corresponds to the centre point of the window (e.g., the purple point in the hypnogram in Fig. 5.1).

Suppose the ith frame-wise time-series input data for cardiac sensing can be denoted as $\mathbf{X}_{car}^{(i)} \in \mathbb{R}^{C_{car} \times T}$, where C_{car} denotes the number of features/input channels and T denotes the sliding window length. For movement sensing, the input data can be denoted as $\mathbf{X}_{mov}^{(i)} \in \mathbb{R}^{C_{mov} \times T}$. The details of feature extraction will be introduced in Section 3.3. The goal of deep multimodal fusion is to determine a multilayer neural network $f(\cdot)$ whose output $\hat{y}^{(i)}$ is expected to be the same as the target $y^{(i)}$ as much as possible for each sample $(\mathbf{X}_{mov}^{(i)}, \mathbf{X}_{car}^{(i)})$. This can be implemented by minimising the empirical loss $\mathscr{L}$ for classification denoted as:

$$\min_{f} \frac{1}{N} \sum_{i=1}^{N} \mathscr{L}\left(\hat{y}^{(i)} = f(\mathbf{X}_{mov}^{(i)}, \mathbf{X}_{car}^{(i)}), y^{(i)}\right) \tag{5.1}$$

5.3.3. Fusion Strategy

Traditional fusion strategies include feature level fusion (e.g., [185]), score-level fusion (e.g., [186]) or decision-level fusion (e.g., [187]). In the end-to-end DL era, the boundary between multimodal representation and fusion has been blurred. Representation learning is interlaced

with classification (or regression) objectives. Nevertheless, the fusion strategy for DL models may still be carried out in three stages, such as early fusion, late fusion and hybrid fusion [188].

Fusion at different stages may influence the results of representation learning. For example, the early and late fusion may inhibit intra-modal or inter-modal interaction [188]. Neverova et al. noted that highly correlated modalities should be fused together [189]. Hazirbas et al. demonstrated that the performance of fusion is highly affected by the choice of which layer to fuse [190]. For sleep stage classification, the way to fuse heterogeneous intermediate features is worthy of exploration. It is meaningful to gain a comprehensive understanding at what stage the model should fuse these inputs to achieve the most performance improvements on the three-stage sleep classification task. Three commonly used fusion strategies were evaluated, including early-stage fusion, late-stage fusion and hybrid fusion, as shown in Figure 5.1.

Early-stage Fusion

In the early-stage fusion, data from different modalities (e.g., intermediate features) are concatenated (stacked) in the input stage. It is popular because of its simplicity, yet it is sub-optimal [156]. Early-stage fusion firstly concatenates the cardiac (denoted as subscript *car*) and movement (denoted as subscript *act*) sensing data then feeds them into neural networks h to make a corresponding prediction.

$$\hat{y}^{(i)} = h(\text{concatenate}(\mathbf{X}^{(i)}_{car}, \mathbf{X}^{(i)}_{mov})). \tag{5.2}$$

where the concatenate$(\cdot)$ is the matrix concatenation function.

Late-stage Fusion

Late-stage fusion is another prevalent way to fuse (high-level) representation from multiple sources. This fusion strategy allows high-level representations to have better intra-modal coherence. Late-stage fusion processes each modality's cth channel input data with a network q and then combines all their high-level representations via an aggregation operation followed by the classification layers. It is denoted as:

$$\hat{y}^{(i)} = \varphi(\text{Agg}(q(\mathbf{x}^{(i)}_{mov,1}), \cdots, q(\mathbf{x}^{(i)}_{mov,C_{mov}}), q(\mathbf{x}^{(i)}_{car,1}), \cdots, q(\mathbf{x}^{(i)}_{car,C_{car}}))) \tag{5.3}$$

where Agg$(\cdot)$ is the aggregation function and φ denotes the classifier (e.g. fully connected layers), and $\mathbf{x}^{(i)} \in \mathbb{R}^{1\times T}$ and T is the window length. The cardiac intermediate features are denoted as $\mathbf{X}^{(i)}_{car} = [\mathbf{x}^{(i)}_{car,1}, \mathbf{x}^{(i)}_{car,2}, \cdots, \mathbf{x}^{(i)}_{car,C_{car}}]$. In this study, the aggregation function represents various fusion methods that will be introduced in the next section. q denotes neural networks that learn the latent representation (e.g., for CNNs, it is the feature maps) of the cth intermediate feature, where C_{car} and C_{mov} are the numbers of the intermediate features for cardiac sensing and movement sensing respectively.

Hybrid Fusion

With hybrid fusion, the fusion may occur at multiple stages/layers of the DL models [192]. It is commonly understood that the DL model hierarchically encodes features at different levels, starting from low-level to higher-level features as the layers go deeper [157]. This study did not cover all possible combinations of fusion architecture. Therefore, following previous work [192], the work in this chapter considers a simple scenario, which is firstly to fuse different input channel data belonging to the same modality (sharing a representation learning network) and then to fuse the high-level features from both modalities at the later stage. Formally, the hybrid-fusion strategy can be written as:

$$\hat{y}^{(i)} = \varphi(\mathrm{Agg}(g_{mov}(\mathbf{X}^{(i)}_{mov}), g_{car}(\mathbf{X}^{(i)}_{car}))) \tag{5.4}$$

where φ is the classifier (e.g., fully connected neural networks) and g denotes the modality-specific networks (e.g., CNNs) that can learn representation from a specific modality such that the g_{mov} does not share network parameters with g_{car}. $\mathrm{Agg}(\cdot)$ is the aggregation function that can be implemented as concatenation, attention mechanism [180] and tensor-based method [282].

5.3.4. Fusion Method

Based on their complexity, fusion methods can be divided into three types: simple operations, attention-based methods and tensor-based methods. For feature vectors from different modalities, concatenation and addition are two commonly used simple operations [188]. The attention mechanism is widely used for multimodal fusion. This usually refers to dynamically calculating a weight vector for each time step (or spatial position) and weighting a set of feature vectors [162, 283]. For tensor-based methods, bilinear pooling is a method of fusing two unimodal representations into a joint presentation by calculating their outer product. This method can capture the multiplicative interaction between all elements in two vectors [193].

For the early-stage fusion, this chapter only adopted concatenation as the only fusion method in this study. For the late-stage fusion, two commonly used simple methods were selected, which are concatenation and element-wise addition. Hybrid fusion provides aggregated representations for each modality, which facilitates flexible fusion methods. Apart from the simple operation methods, this chapter also evaluated the attention mechanism and the tensor-based method. The choice of the fusion method may be influenced by the application context.

Concatenation

For **early-stage fusion**, the concatenation method concatenates inputs of all modalities into one matrix, which can be denoted as:

$$\mathbf{K}^{(i)}_{early} = \mathrm{concatenate}(\mathbf{X}^{(i)}_{car}, \mathbf{X}^{(i)}_{mov}) \tag{5.5}$$

where $\mathbf{K}_{early}^{(i)} \in \mathbb{R}^{(C_{mov}+C_{car})\times T}$. $\mathbf{X}_{car}^{(i)} \in \mathbb{R}^{C_{car}\times T}$ is the intermediate feature matrix of cardiac sensing, $\mathbf{X}_{mov}^{(i)} \in \mathbb{R}^{C_{mov}\times T}$ is the intermediate feature matrix of movement sensing. The C_{car} represents the number of intermediate features of cardiac sensing and T represents the number of temporal steps.

For **late-stage fusion**, suppose we have the cardiac latent representation denoted as $\mathbf{X}_{car,c}^{\prime(i)} \in \mathbb{R}^{U\times L}$, which is learned from a neural network g via $\mathbf{X}_{car,c}^{\prime(i)} = g(\mathbf{x}_{car,c}^{(i)})$. The movement representation matrix is computed in the same way, which can be formally denoted as $\mathbf{X}_{mov}^{\prime(i)} = g(\mathbf{x}_{mov}^{(i)})$ where the $\mathbf{X}_{mov}^{\prime(i)} \in \mathbb{R}^{U\times L}$ is the latent representation of movement sensing. L is the temporal length and U is the representation's dimension. For example, in a convolutional neural network, U is the number of feature maps. At the late stage, as the feature maps of each input intermediate feature were kept separately, the concatenation operation concatenates these representations together, as follows:

$$\mathbf{K}_{late}^{(i)} = \text{concatenate}(\mathbf{X}_{mov,1}^{\prime(i)}, \cdots, \mathbf{X}_{mov,C_{mov}}^{\prime(i)}, \mathbf{X}_{car,1}^{\prime(i)}, \cdots, \mathbf{X}_{car,C_{car}}^{\prime(i)}) \tag{5.6}$$

In this study, for the activity counts (handcraft feature) and cardiac features, the late-stage fusion's representation is denoted as $\mathbf{K}_{late}^{(i)} \in \mathbb{R}^{(C_{mov}+C_{car})\times U\times L}$

For the **hybrid fusion**, the high-level representation of each modality is obtained from their own sub-network. The movement sensing representation is denoted as $\mathbf{X}_{mov}^{\prime\prime(i)} = g_{mov}(\mathbf{X}_{mov}^{(i)})$ and the cardiac sensing is formally denoted as $\mathbf{X}_{car}^{\prime\prime(i)} = g_{car}(\mathbf{X}_{car}^{(i)})$. The concatenation method for the hybrid fusion can be written as:

$$\mathbf{K}_{hybrid}^{(i)} = \text{concatenate}(\mathbf{X}_{car}^{\prime\prime(i)}, \mathbf{X}_{mov}^{\prime\prime(i)}) \tag{5.7}$$

where $\mathbf{K}_{hybrid}^{(i)} \in \mathbb{R}^{2U\times L}$

Addition

The second simple operation is the element-wise addition denoted as $\oplus$. For the **late-stage fusion**, the addition operation is to integrate the high-level representation of each channel from each modality. The method is formally denoted:

$$\mathbf{Q}_{late}^{(i)} = \mathbf{X}_{mov,1}^{\prime(i)}\oplus, \cdots, \mathbf{X}_{mov,C_{mov}}^{\prime(i)} \oplus \mathbf{X}_{car,1}^{\prime(i)}, \cdots, \oplus\mathbf{X}_{car,C_{car}}^{\prime(i)} \tag{5.8}$$

where $\mathbf{Q}_{late}^{(i)} \in \mathbb{R}^{U\times L}$.

For the **hybrid fusion**, the addition method will aggregate the high-level representation of each modality. Formally, it can be denoted as:

$$\mathbf{Q}_{hybrid}^{(i)} = \mathbf{X}_{car}^{\prime\prime(i)} \oplus \mathbf{X}_{mov}^{\prime\prime(i)} \tag{5.9}$$

where $\mathbf{Q}_{hybrid}^{(i)} \in \mathbb{R}^{U\times T}$.

Attention Mechanism

Attention methods have been broadly adopted in multimodal fusion tasks. For example, in VQA tasks the method used is to fuse the visual representations with the language representation [180]. In this study, the attention vector in the attention model will weigh one modality according to the context of the other modality. The meaning behind this is to filter the most significant information from a unimodal, which is jointly relevant for three-stage sleep classification. Therefore, two attention fusion methods were adopted. The first one is Attention-on-Movement (Attention-on-Mov) and the second one is Attention-on-Cardiac (Attention-on-Car). Given the cardiac representation matrix $\mathbf{X}'^{(i)}_{car}$ and the movement representation matrix $\mathbf{X}'^{(i)}_{mov}$, First, they are fed through a single-layer neural network and then a softmax function is applied to generate the attention distribution over the temporal dimension, which is denoted as:

$$\mathbf{H}^{(i)}_{att} = \tanh(\mathbf{W}_{car}\mathbf{X}'^{(i)}_{car} \oplus \mathbf{W}_{mov}\mathbf{X}'^{(i)}_{mov} + b_h) \tag{5.10}$$

$$\mathbf{P}^{(i)}_{att} = \text{softmax}(\mathbf{W}_{att}\mathbf{H}^{(i)}_{att} + b_{att}) \tag{5.11}$$

where $\mathbf{X}'^{(i)}_{mov} \in \mathbb{R}^{U\times L}$. Suppose we have linear transformation matrices that include $\mathbf{W}_{mov}, \mathbf{W}_{car} \in \mathbb{R}^{D\times U}$ and $\mathbf{W}_{att} \in \mathbb{R}^{L\times D}$, then $\mathbf{H}^{(i)}_{att} \in \mathbb{R}^{D\times L}$ and $\mathbf{P}^{(i)}_{att} \in \mathbb{R}^{L\times L}$, where D is the dimension of attention embedding space. The attention weight matrix is denoted as $\mathbf{P}^{(i)}_{att} = [\mathbf{p}^{(i)}_{att,1}, \cdots, \mathbf{p}^{(i)}_{att,L}]$ and each temporal step has an attention vector $\mathbf{p}^{(i)}_{att,l}$, where $\sum \mathbf{p}^{(i)}_{att,l} = 1$. The subscript *att* stands for attention and l is the temporal step index.

A reasonable approach is that applying attention weights on different modalities will have an impact on the results. Therefore, two scenarios were studied in this work. The first method is to weigh cardiac sensing representations based on the attention distribution and concatenate them to build the joint feature representation matrix. It can be written as:

$$\mathbf{V}^{(i)}_{car} = \mathbf{X}'^{(i)}_{car}\mathbf{P}^{(i)}_{att} \tag{5.12}$$

$$\mathbf{K}^{(i)}_{car} = \text{concatenate}(\mathbf{V}^{(i)}_{car}, \mathbf{X}'^{(i)}_{mov}) \tag{5.13}$$

This chapter refer to this method as Attention-on-Car and $\mathbf{K}^{(i)}_{car} \in \mathbb{R}^{2U\times L}$

The second method is to weight the latent feature of movement sensing using the attention distribution, then concatenate them to build the joint representation matrix, which can be denoted as:

$$\mathbf{V}^{(i)}_{mov} = \mathbf{X}'^{(i)}_{mov}\mathbf{P}^{(i)}_{att} \tag{5.14}$$

$$\mathbf{K}^{(i)}_{mov} = \text{concatenate}(\mathbf{V}^{(i)}_{mov}, \mathbf{X}'^{(i)}_{car}) \tag{5.15}$$

This chapter refers to this method as Attention-on-Mov and $\mathbf{K}^{(i)}_{mov} \in \mathbb{R}^{2U\times L}$ is the merged joint representation.

Bilinear Pooling Method

Bilinear pooling is a method to compute the outer product of matrices that can facilitate multiplication interaction between all elements in both matrices. It is a method often used to fuse visual feature vectors with textual feature vectors to create a joint representation space, even though their distribution may vary dramatically[276, 282]. During the NREM sleep period, the cardiac system is co-modulated by peripheral and sympathetic neural systems. The heart rate is generally below the average for the wake and REM sleep period and is accompanied by tiny tremors in limb movement. A hypothesis proposed in this chapter assumes the bilinear model may be able to capture such tiny differences between REM and NREM sleep. Given its superior representation learning capacity, it has achieved remarkable performance in fine-grained image classification tasks [284]. The bilinear model calculates the outer product of two matrices. In this work, suppose the two feature representation matrices $\mathbf{X}'^{(i)}_{car}$ and $\mathbf{X}'^{(i)}_{mov}$, and the bilinear representation can be written as:

$$\mathbf{k}^{(i)}_{bi} = \text{vec}(\mathbf{X}'^{(i)}_{car} \otimes \mathbf{X}'^{(i)}_{mov}) \tag{5.16}$$

The symbol of $\otimes$ denotes the Kronecker product of two matrices, and the vec denotes the matrix vectorisation. After the vectorisation, an element-wise signed square root is performed as denoted:

$$\mathbf{k}^{(i)}_{bi} \leftarrow \text{sign}(k^{(i)}_{bi})\sqrt{|k^{(i)}_{bi}|} \tag{5.17}$$

and then apply l_2 normalisation on the vector $\mathbf{k}^{(i)}_{bi}$. Afterwards, the normalised vector was passed to a linear function to reduce the feature dimension before feeding it into the classifier.

5.4. Experiment Design

This section describes the experimental design of advanced multimodal fusion strategies and methods for the three-stage sleep classification using wearable devices. Firstly, two open-access datasets used in the study are introduced, including the data collection, data pre-processing and feature extraction. Secondly, four backbone networks used with advanced multimodal fusion techniques are illustrated. Finally, the evaluation metrics used in the study are discussed.

5.4.1. Dataset Description

Apple Watch Sleep Dataset

The first dataset used in this chapter is the Apple Watch Sleep Study[1], which is an open-access dataset collected at the University of Michigan between 2017 and 2019 [6, 285]. The dataset consists of 31 healthy subjects with no known sleep disorders or cardiovascular diseases and neurological or psychiatric impairment disorders[6]. All subjects wore Apple Watch (Apple Inc. series 2 and 3) and performed continuous recording for 7 to 14 days, and then joined the PSG

[1] https://physionet.org/content/sleep-accel/1.0.0/

study in the sleep laboratory on the last day [6]. During the PSG study, all subjects wore Apple Watch, which recorded heart rate and triaxial acceleration [6]. The acceleration and heart rate were measured by Apple Watch and recorded by a custom-developed watch application using the built-in functions of the iOS Watch kit and HealthKit by creating a "Workout Session" in app [285]. The PSG recordings were annotated according to the AASM rules [6].

The heart rate is measured by the PPG sensor of the Apple Watch and recorded as beats per minute (BPM), and a sample is returned from the Apple API every few seconds. The heart rate data is timestamped and the interval is between 2s and 5s. After the data cleaning process, the feature engineering process was performed on triaxial acceleration data; following [76, 6], the activity counts were used as the movement feature. The final activity counts were added for each sleep epoch. Since the heart rate collected from Apple Watch is calculated in two to five seconds, this data may be deemed as "pseudo" instantaneous heart rate (IHR). In each sleep epoch, the summary statistics of the heart rate data were calculated (called HR statistics or HRS for short), which includes the mean, standard deviation, minimum, maximum, skewness and kurtosis of the heart rate. Together with the activity counts, a seven-dimensional vector was constructed for each sleep epoch from these intermediate summary features and called it the Apple ACT-HRS feature set.

MESA Dataset

Following the work in previous chapters, the MESA dataset is used for the experiments in this chapter. The data processing pipeline is consistent with the chapter 3. After the data pre-processing, 1,743 of 2,237 participants satisfied the data quality condition. Full details of the study setup, protocol and sampling rates are available in [5, 269, 271]. According to the feature set used in previous research [5, 6], this chapter used the same features, including activity counts and eight HRV features derived from the NN interval data in each sleep epoch [5]. The feature set consists of the Mean NNi, Standard Derivation of RR interval (SDNN), RR interval differences (SDSD), Very Low Frequency, Low Frequency, High Frequency Bands, Low Frequency to High Frequency Ratio and Total Power. These features have been investigated in several sleep physiology studies [214, 280, 286, 126, 52]. For each sleep epoch, the intermediate feature vector was constructed based on eight HRV features and the activity counts (a scalar value per sleep epoch), which is named the MESA ACT-HRV feature set.

In addition, the NN intervals were converted into IHR data, and the statistical features of IHR and combined activity counts were calculated as the second intermediate feature set. The purpose is to study the feature effects on the choices of fusion strategies and methods. In this chapter, this feature set is named the MESA ACT-HRS feature set.

5.4.2. Evaluation Metrics

For performance evaluation, accuracy, Cohen's κ, mean F_1 and time deviation([5]) were used. The time deviation that was used in chapter 3 is denoted as ($TD_k = \frac{1}{N}\sum_{i=1}^{N}(Pred_c^i - GT_c^i)$). For a sleep stage c, the $Pred_c$ refers to the predicted minutes and GT_c refers to the ground truth sleep

minutes. The superscript i represents the ith subject. The time deviation summarises the mean bias of the total minutes of each sleep stage predicted by the classifier in the population. To understand the impact of individual differences in performance evaluation, this study adopted the subject-level evaluation. The metrics of each subject were calculated individually and the mean value and 95% confidence interval of each metric were obtained for the population.

5.4.3. Experimental Procedure

Following the approaches in previous chapters, a highly overlapping sliding window method was adopted with $S = 1$ to segment the input time-series data. In chapter 3, the hyperparameter tuning results showed that the window length can impact the prediction performance. For convolutional neural networks, a longer window produced better results compared with a shorter window.

For each sleep epoch, the data of 50 adjacent (forward and backward) sleep epochs were selected to construct the inputs with a window length of 101. The details are shown in Figure 5.1. The hypnogram represents the stages of sleep over time in each sleep epoch. The prediction process was performed for each sleep epoch. The backbone network has the same structure for early-stage fusion and hybrid fusion. For the sliding window at the beginning and end of the recording, the empty sleep epoch inputs were filled with a value of -1. For the training, validation and testing, the experimental settings are as follows:

- **Apple Watch Sleep Dataset** Following the experiment setting of previous work [6], instead of using leave-one-subject-out-cross-validation, this chapter adopted leave-two-subjects-out cross validation. Each fold had two subjects for testing, except for the last fold, which only contained one subject (total 31 subjects and 16 folds). In each fold, the data in the training dataset were randomly split into a validation dataset (20%), and the rest 80% were used for training. The validation set was used to select the best model for the test dataset.

- **MESA Sleep Dataset** The dataset contains 1743 valid sleep records of subjects. This chapter employed the hold-out method to divide the entire dataset into a test set of 348 subjects (20%) and a training set of 1,395 subjects (80%) following the experimental protocol in chapter 3 The training set was further randomly split into a validation set (20%) and a training set (80%). Again, the validation set was used to select the best model for the test dataset.

All experiments conducted in this paper adopted the above setting for each dataset respectively.

In chapter 3, it was found that the performance improvement of three-stage sleep classification was more related to increasing the number of LSTM networks instead of increasing the number of CNN layers for the three-stage sleep classification task. Therefore, this study focused on the design of CNN architecture. All experiments in this work adopted the Adam gradient update rule [287] with learning rate $\alpha = 10^{-4}, \beta_1 = 0.9$, and $\beta_2 = 0.99$. No early-stopping or weight decay was adopted in training processing. The batch size was set to 1024 except for the experiments containing the bilinear method, which were set to 512. This setting is due to the GPU memory

limitation. For the attention method, the attention embedding dimension was set to 256. For the bilinear method, the size of the feature dimension was reduced to 1024 using a linear layer. The training epoch corresponding to both datasets was set to 20. The training was completed on a single GPU-equipped (Nvidia RTX 3090) machine, and each epoch took approximately three minutes.

5.4.4. Implementation Details

Hyperparameter Tuning and Backbone Networks

The fusion strategies and fusion methods may not benefit from a single-layer convolutional neural network. To find a feasible backbone deep CNN that was capable of serving the study, inspired by [148], a backbone network was designed and a hyperparameter search was conducted on 3-5 convolutional layer blocks (corresponding to 7-13 convolutional layers). From the hyperparameter tuning results, the network from the highest F_1 validation score group was selected. The number of hidden units was further gradually reduced in fully connected layers and the experimental results showed slight improvements in model performance. This network is named DeepCNN. To better understand the impact of modality fusion strategies and methods in different CNN architectures, inspired by [288, 289], this chapter further added a skip connection in each convolutional block and called it ResDeepCNN, as the skip connection became an indispensable component in a variety of neural architectures that could boost representation learning. Figure 5.2 lists the details of two network structures. The stride and padding values were set to 1 for all convolutional (Conv) layers, and the kernel size was set to 3. The kernel size and stride were set to 2 for all max pooling (Max-Pool) layers. A dropout layer was applied after each fully connected (FC) layer, and the dropout rate was set to 0.25. The DeepCNN network was selected from the hyperparameter search results. The skip connection inside each convolutional block is added and it is referred to as ResDeepCNN. As this chapter focuses on the fusion strategy and methods, the backbone network was merely designed to conduct feasible experiments. More details on the hyperparameter search of the backbone network can be seen in the appendix section B.1

Backbone Network Setting

For the early-stage fusion and hybrid fusion, DeepCNN and ResDeepCNN were the main networks for the experiments. This chapter slightly adapted the DeepCNN and the ResDeepCNN for the late-stage fusion experiments according to [152], which allowed each input channel to share the convolutional kernels but kept the feature representation separate. This means that, for a convolutional layer, the feature maps extracted from each input channel would not be fused with the feature maps of other channels. Instead, each input channel's feature maps would be fused before the classification module (fully-connected layers). For instance, for DeepCNN in the Apple Watch dataset, if the input was an intermediate feature matrix that contained cardiac and movement sensing and was denoted as $\mathbf{S}_0^{(i)} \in \mathbb{R}^{7\times 101}$ (one movement feature and six HRS

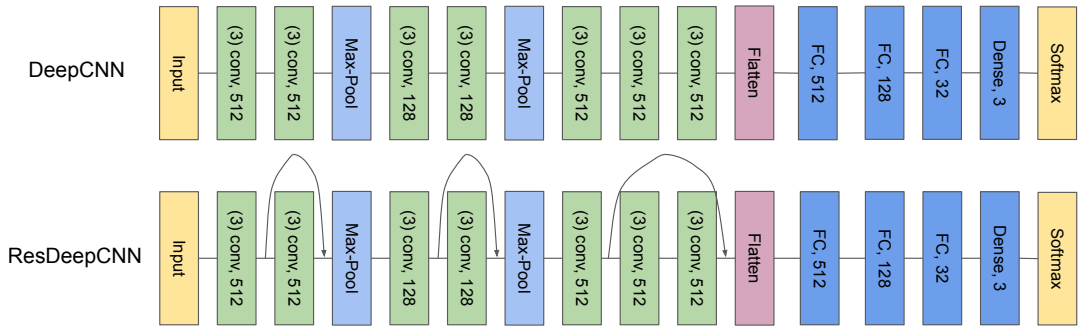


Figure 5.2 Backbone network used in this study.

features), the feature map function $\mathscr{F}_{l+1} : \mathbf{S}_l^{(i)} \mapsto \mathbf{S}_{l+1}^{(i)}$ was realised by a convolutional layer, where l denotes the lth convolutional layer. The output of the first convolutional layer was the feature map denoted as $\mathbf{S}_1^{(i)} = \mathscr{F}_1(\mathbf{S}_0^{(i)}) \in \mathbb{R}^{C_1 \times 7 \times 101}$, where C_1 was the number of feature maps of the first CNN layer. In this way, the intermediate feature of each input channel was kept separate.

5.5. Results

This section empirically compares each combination of multimodal fusion strategies and methods based on two scenarios. The first scenario is the MESA dataset which contains multimodal data that can be extracted from research grade-wearable devices. The second scenario is the Apple Watch dataset derived from consumer-grade smartwatches (Apple Watch Series 2 and 3) with the sleep stages annotated using the gold-standard PSG study.

The performance is reported in the order of three fusion strategies which included early-stage fusion, late-stage fusion, and hybrid fusion, and three fusion methods, including simple operation (concatenation and addition), attention mechanism, and bi-linear pooling method. The effects of different window lengths (51 and 21) were also investigated, and the corresponding results can be seen in Appendix B.3.1. The experiments using raw accelerometer data and HR statistical features can be seen in Appendix B.2.2.

For consistency, all fusion strategies and methods in each dataset were evaluated on the subject level during the sleep recording period. Accuracy, Cohen's κ, the mean F_1 score and time deviation (minutes) were calculated based on the predictions during the sleep recording period. In the end, this chapter compared the model parameter size and inference time for each strategy and method. These factors are important for model selection in the context of ubiquitous computing.

5.5.1. Apple Watch Dataset

Activity Counts and HRS Features

The first experiment was performed based on activity counts and the HRS feature set (ACT-HRS) derived from the consumer wearables.

Fusion Specifics			Performance Metrics			Time Deviation(min.)		
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy(%)**	**Cohen's κ**	**Mean F_1(%)**	**Non-REM sleep**	**REM sleep**	**Wake**
Early-Stage Fusion	DeepCNN	Concatenation	72.3 ± 2.5	40.0 ± 5.8	59.0 ± 3.6	-0.6 ± 22.3	11.8 ± 22.5	-11.2 ± 6.9
	ResDeepCNN	Concatenation	76.0 ± 2.4	45.7 ± 6.1	63.4 ± 3.5	12.3 ± 17.2	-4.0 ± 17.1	-8.2 ± 7.3
Late-Stage Fusion	DeepCNN	Concatenation	76.2 ± 2.7	47.0 ± 7.1	63.7 ± 4.5	8.4 ± 16.5	1.3 ± 19.4	-9.7 ± 7.1
		Addition	**78.2 ± 2.1**	49.9 ± 6.9	65.2 ± 3.8	11.4 ± 10.4	-0.6 ± 11.2	-10.8 ± 6.6
	ResDeepCNN	Concatenation	75.5 ± 2.9	47.0 ± 6.5	63.4 ± 4.2	10.4 ± 18.5	-2.1 ± 19.6	-8.3 ± 7.0
		Addition	78.2 ± 2.3	**52.0 ± 5.8**	**66.5 ± 3.8**	1.9 ± 11.7	4.7 ± 12.6	-6.5 ± 7.3
Hybrid Fusion	DeepCNN	Concatenation	72.9 ± 3.0	40.1 ± 6.6	60.6 ± 3.8	10.7 ± 20.1	0.6 ± 19.5	-11.4 ± 6.4
		Addition	72.1 ± 2.9	38.6 ± 6.5	58.9 ± 3.9	8.8 ± 20.4	1.3 ± 19.9	-10.2 ± 7.3
		Attention-on-Mov	73.5 ± 2.9	41.3 ± 6.2	60.4 ± 3.9	1.2 ± 18.7	5.0 ± 18.9	-6.2 ± 7.7
		Attention-on-Car	71.1 ± 3.4	37.4 ± 6.7	58.1 ± 4.1	7.5 ± 24.3	0.7 ± 23.1	-8.2 ± 7.5
		Bilinear	69.5 ± 3.5	29.5 ± 7.5	53.0 ± 4.2	1.2 ± 25.7	10.0 ± 25.0	-11.3 ± 8.5
	ResDeepCNN	Concatenation	74.4 ± 2.4	44.2 ± 5.7	62.0 ± 3.3	2.5 ± 16.7	5.3 ± 16.9	-7.8 ± 7.0
		Addition	74.9 ± 2.3	44.3 ± 5.4	62.3 ± 3.7	12.8 ± 14.2	-7.6 ± 15.6	-5.2 ± 8.2
		Attention-on-Mov	75.2 ± 2.6	45.0 ± 5.6	63.1 ± 3.4	10.4 ± 19.4	-2.0 ± 19.5	-8.4 ± 7.2
		Attention-on-Car	72.2 ± 3.0	41.5 ± 6.8	59.7 ± 3.8	-2.9 ± 25.3	12.0 ± 26.4	-9.1 ± 6.7
		Bilinear	70.8 ± 3.5	38.4 ± 7.5	58.1 ± 4.4	-2.8 ± 22.9	6.4 ± 24.8	-3.5 ± 8.0

Table 5.1 Three-stage sleep classification results (mean ± standard error at 95% confidence interval) for each combination of fusion strategies and methods with the Apple Watch dataset using ACT-HRS feature based on a window length of 101.

Table 5.1 lists the subject-level evaluation results of the Apple Watch dataset based on the window length of 101 during the sleep recording period. Since Apple Watch sampled the heart rate data with an unknown resolution and method, the experiments were only performed based on the ACT-HRS feature set.

Overall, the ResDeepCNN achieved the highest mean F_1 score of 66.5%, Cohen's κ of 52 and an accuracy of 78.2% using the addition method in late-stage fusion. The same methods used on DeepCNN were higher than those in early-stage fusion too.

In the hybrid fusion strategy, using the Attention-on-Mov method achieved the highest scores irrespective of backbone networks. The experimental results demonstrated a similar pattern in the experiments using window lengths 51 and 21. For window lengths 51 and 21, the highest-performed models in each category were lower than the highest-performed models using the window length of 101. This chapter listed these results for window lengths 51 and 21 in Appendix B.3.1.

5.5.2. MESA Sleep Dataset Results

The second experiment was conducted on the MESA dataset. It was by far the largest sleep dataset that contained activity counts and instantaneous heart rates, which might be extracted from research-grade wearable devices. Again, this chapter performed experiments on two different feature sets. The first feature set included activity counts and HRV features (ACT-HRV) [5], while the second feature set was ACT-HRS derived using the same feature extraction method in the Apple Watch dataset.

MESA Activity Counts and HRV Features

The reason for using the HRV features was that they had sleep physiological meaning. Table 5.2 shows the subject-level evaluation results based on the window length of 101. For the early-stage

and late-stage fusion, the results of the two backbone networks were comparable, which showed the skipping connections did not improve the classification performance.

Fusion Specifics			Performance Metrics			Time Deviation(min.)		
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy(%)**	**Cohen's κ**	**Mean F_1(%)**	**Non-REM sleep**	**REM sleep**	**Wake**
Early-Stage Fusion	DeepCNN	Concatenation	78.6 ± 0.9	62.8 ± 1.8	71.1 ± 1.3	27.1 ± 6.9	-6.5 ± 3.6	-20.7 ± 6.4
	ResDeepCNN	Concatenation	78.0 ± 1.1	60.2 ± 2.0	71.1 ± 1.4	54.1 ± 7.0	1.2 ± 3.8	-55.3 ± 6.6
Late-Stage Fusion	DeepCNN	Concatenation	79.6 ± 0.9	64.3 ± 1.8	72.5 ± 1.3	12.9 ± 6.6	0.1 ± 3.5	-13.0 ± 6.2
		Addition	78.5 ± 0.9	62.3 ± 1.8	71.1 ± 1.3	25.7 ± 6.4	-6.8 ± 3.3	-18.9 ± 6.4
	ResDeepCNN	Concatenation	79.3 ± 0.9	64.4 ± 1.7	72.6 ± 1.2	8.9 ± 6.4	4.2 ± 3.4	-13.1 ± 6.1
		Addition	78.6 ± 1.0	62.8 ± 1.9	71.4 ± 1.3	2.4 ± 6.9	-3.0 ± 3.5	0.6 ± 6.7
Hybrid Fusion	DeepCNN	Concatenation	77.6 ± 1.1	62.7 ± 1.8	71.4 ± 1.3	-12.7 ± 7.3	7.2 ± 3.9	5.5 ± 7.0
		Addition	79.0 ± 0.9	62.9 ± 1.7	70.7 ± 1.3	53.6 ± 6.9	-15.0 ± 3.3	-38.6 ± 6.3
		Attention-on-Mov	79.0 ± 1.0	63.9 ± 1.8	72.1 ± 1.3	2.3 ± 7.0	-4.9 ± 3.5	2.6 ± 6.9
		Attention-on-Car	78.1 ± 1.0	63.0 ± 1.7	71.6 ± 1.3	-2.1 ± 6.8	6.5 ± 4.0	-4.4 ± 6.3
		Bilinear	75.7 ± 0.9	58.6 ± 1.8	68.8 ± 1.2	3.7 ± 6.8	16.6 ± 4.0	-20.3 ± 6.3
	ResDeepCNN	Concatenation	79.7 ± 0.9	65.3 ± 1.7	72.7 ± 1.3	6.4 ± 6.7	-7.7 ± 3.4	1.3 ± 6.7
		Addition	**79.8 ± 0.9**	64.1 ± 1.7	72.7 ± 1.3	24.1 ± 6.9	-9.7 ± 3.2	-14.4 ± 6.5
		Attention-on-Mov	79.6 ± 1.0	**65.5 ± 1.8**	**73.3 ± 1.3**	-0.9 ± 6.4	6.1 ± 3.7	-5.1 ± 6.3
		Attention-on-Car	78.5 ± 1.0	62.7 ± 1.7	70.5 ± 1.3	37.6 ± 7.0	-9.5 ± 4.0	-28.1 ± 6.2
		Bilinear	75.7 ± 0.9	58.6 ± 1.8	68.8 ± 1.2	3.7 ± 6.8	16.6 ± 4.0	-20.3 ± 6.3

Table 5.2 Three-stage sleep classification results (mean ± standard error at 95% confidence interval) for each combination of fusion strategies and methods with the MESA test dataset using the ACT-HRV feature set based on a window length of 101.

For the hybrid fusion strategy, the ResDeepCNN achieved the highest accuracy, the Cohen's κ and the mean F_1 score of 79.6%, 65.5 and 73.3%, respectively, using the Attention-on-Mov method. The results were statistically significant ($p < 0.05$) and higher than the models in the early-stage fusion. Those metrics were higher than the Attention-on-Car models too. Similar to the Apple Watch dataset, the performance of models based on window lengths 51 and 21 tends to be worse than experiments performed with window length 101. These experimental results can be seen in Appendix B.3.1.

In terms of time deviation, DeepCNN achieved the optimal time deviation using the Attention-on-Mov method. The mean value of the NREM sleep time deviation was 0.9, and the mean value of the REM sleep time deviation was 6.1.

MESA Activity Counts and HRS Features

The heart rate statistical features were derived from the instantaneous heart rate (IHR) data in the MESA dataset. The purpose was to understand whether the type of intermediate feature would cause a difference in results.

The subject-level evaluation is shown in Table 5.3. In the early-stage fusion, similar to the ACT-HRV feature setting, the results of the two backbone networks were comparable. The ResDeepCNN, using the Attention-on-Mov fusion method, achieved the highest accuracy, Cohen's κ, and the mean F_1 score of 80.3%, 65.6, and 72.9%, respectively. However, the Attention-on-Mov model based on the ACT-HRS feature set highly overestimated the NREM sleep time and underestimated the wake minutes. Again, the models with window lengths 51 and 21 achieved lower performance than 101. Therefore, these results were listed in Appendix B.3.1

Fusion Specifics			Performance Metrics			Time Deviation(min.)		
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy(%)**	**Cohen's κ**	**Mean F_1(%)**	**Non-REM sleep**	**REM sleep**	**Wake**
Early-Stage Fusion	DeepCNN	Concatenation	78.0 ± 1.0	63.4 ± 1.8	72.0 ± 1.2	2.7 ± 7.0	15.6 ± 4.0	-18.2 ± 6.4
	ResDeepCNN	Concatenation	76.9 ± 1.1	61.9 ± 1.9	71.1 ± 1.3	-36.7 ± 7.8	12.1 ± 4.2	24.5 ± 7.5
Late-Stage Fusion	DeepCNN	Concatenation	79.1 ± 1.0	64.7 ± 1.7	72.8 ± 1.3	0.6 ± 6.9	7.5 ± 3.7	-8.1 ± 6.4
		Addition	78.1 ± 0.9	62.8 ± 1.6	70.6 ± 1.2	23.1 ± 6.6	-4.5 ± 3.7	-18.5 ± 6.3
	ResDeepCNN	Concatenation	77.8 ± 1.0	62.5 ± 1.7	71.0 ± 1.3	-1.1 ± 7.3	-0.7 ± 3.5	1.8 ± 6.8
		Addition	77.7 ± 1.0	62.6 ± 1.7	70.7 ± 1.2	24.3 ± 7.0	3.7 ± 3.9	-28.0 ± 6.4
Hybrid Fusion	DeepCNN	Concatenation	78.2 ± 0.9	64.4 ± 1.7	70.2 ± 1.2	18.5 ± 7.1	-14.6 ± 3.3	-3.9 ± 6.5
		Addition	78.1 ± 0.9	62.2 ± 1.8	71.2 ± 1.2	14.7 ± 7.4	1.4 ± 3.6	-16.1 ± 6.8
		Attention-on-Mov	79.2 ± 0.9	63.8 ± 1.8	71.8 ± 1.3	28.1 ± 7.3	-4.3 ± 3.5	-23.9 ± 6.5
		Attention-on-Car	76.6 ± 1.0	61.4 ± 1.7	70.4 ± 1.2	-7.4 ± 7.9	17.7 ± 4.7	-10.3 ± 6.7
		Bilinear	75.6 ± 0.9	58.0 ± 1.8	67.7 ± 1.2	6.3 ± 7.6	-8.1 ± 3.7	1.8 ± 7.1
	ResDeepCNN	Concatenation	79.4 ± 1.0	64.4 ± 1.7	72.7 ± 1.2	31.7 ± 6.9	4.9 ± 3.6	-36.6 ± 6.3
		Addition	78.9 ± 0.9	63.6 ± 1.8	72.2 ± 1.2	24.6 ± 7.0	4.9 ± 3.5	-29.5 ± 6.5
		Attention-on-Mov	**80.3 ± 0.9**	**65.6 ± 1.7**	**72.9 ± 1.3**	35.5 ± 6.9	0.8 ± 3.6	-36.3 ± 6.2
		Attention-on-Car	79.3 ± 0.9	62.8 ± 1.7	71.1 ± 1.2	29.1 ± 7.1	-2.4 ± 3.7	-26.7 ± 6.3
		Bilinear	74.1 ± 0.9	56.8 ± 1.7	66.9 ± 1.2	6.2 ± 7.1	11.7 ± 4.2	-17.9 ± 6.5

Table 5.3 Three-stage sleep classification results (mean ± standard error at 95% confidence interval) for each combination of fusion strategies and methods with the MESA test dataset using the ACT-HRS feature set based on a window length of 101.

5.5.3. Inference Efficiency

Fusion Strategy	Network	Fusion Method	Total Parameters (M)	Inference Time (ms per sample)
Early-Stage	DeepCNN	Concatenation	**9.44**	**3.52±0.08**
	ResDeepCNN	Concatenation	9.44	3.54±0.08
Late-Stage Fusion	DeepCNN	Concatenation	48.75	22.9±0.17
		Addition	9.43	32.25±5.53
	ResDeepCNN	Concatenation	48.75	31.16±5.06
		Addition	9.43	22.11±0.25
Hybrid Fusion	DeepCNN	Concatenation	18.80	7.13±0.12
		Addition	12.24	7.01±0.12
		Attention-on-Act	19.07	7.32±0.12
		Bilinear	274.65	10.09±0.11
	ResDeepCNN	Concatenation	18.80	7.02±0.16
		Addition	12.24	7.0±0.14
		Attention-on-Act	19.07	7.27±0.17
		Bilinear	274.65	10.22±0.17

Table 5.4 The number of model parameters and inference time of each combination of fusion strategies and methods evaluated in millions of parameters and milliseconds respectively with the *Apple Watch* dataset, using the *ACT-HRS* feature sets based on a window length of 101.

In the mobile computing scenario of three-sleep stage classification, the model based on the deep learning architecture may require sufficient computing resources. This may be a challenge for many inexpensive or low-end smartwatches and smartphones. Table 5.4 shows the model parameter size and inference time of each combination of fusion strategies and methods. In addition, the number of trainable parameters was counted and the time required for forward propagation (running on CPU) is also calculated. All experiments were conducted using Pytorch 1.6 and the hardware platform consisting of 8 cores AMD-7 3700X with 4.4GHz and 64GB DDR4 memory. This chapter independently ran each model 10 times on the Apple Watch dataset. Each time, 500 samples (sleep epochs) were inferred using the Pytorch profiling module to calculate the statistical summary of the inference time.

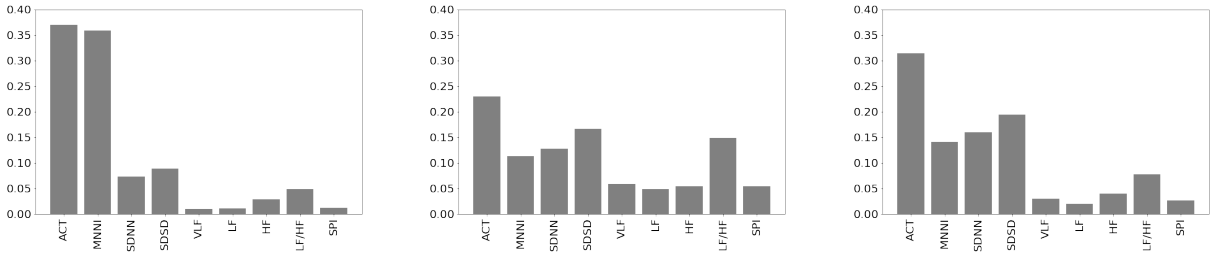
Overall, the models using the addition method in late-stage fusion, hybrid fusion, and concatenation in early-stage fusion had the least model parameters. As a result, the models in early-stage fusion achieved the shortest inference time. The addition method in the late-

stage fusion had the same number of parameters as the models in early-stage fusion, but the inference time was increased by 7-10 times. This was because the late-stage fusion calculated the feature maps of each input channel separately and fused them before the classifier module (fully connected layers). Consequently, the feature matrix extracted by the convolutional module was a 3D tensor (e.g., the number of input feature dimensions $\times$ number of feature maps $\times$ temporal steps) in the late-stage fusion. In contrast, the early-stage fusion generated a 2D tensor (e.g., number of feature maps $\times$ temporal steps). The convolution operation requires additional time to calculate the feature maps of each input channel.

The bilinear model had the largest model parameters. Most model parameters belonged to the feature representation module, which contained a fully connected layer to reduce the dimension of feature representation at an order of two magnitudes. Since the calculation speed of the fully connected (FC) layer was much faster than the convolutional layer, the inference time did not increase as much as the model parameter size.

5.5.4. Visualisation and Interpretation

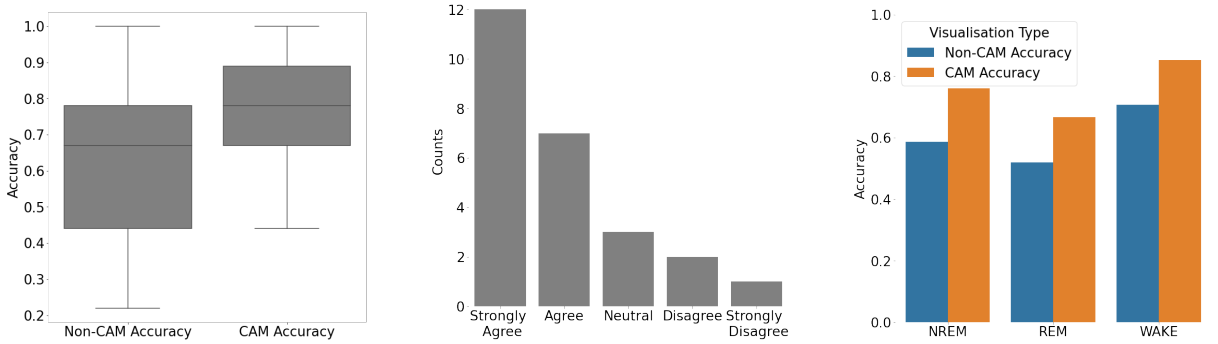
The top three channels of each sleep stage were retained to simplify the visualisation as shown in Figure 5.5. To obtain a clear graph, the highlighted areas are the activation values over the threshold of 0.8 and the light colour areas represent activation values below the threshold of 0.8. To test whether this visualisation is useful and can be understood by humans, A game system[2] was designed that could conduct the exploratory study with users. The game system serves the purpose of engaging users to read and understand these visualisations. The study consists of two phases, which investigate the accuracy of sleep stage classification by humans based solely on the input signals in the cases of Non-CAM and CAM visualisation, respectively. For each phase, a continuous period of sleep data (intermediate feature data and hypnogram) was encoded for each sleep stage into videos to speed up the training process.



(a) Wake **(b)** NREMS **(c)** REMS

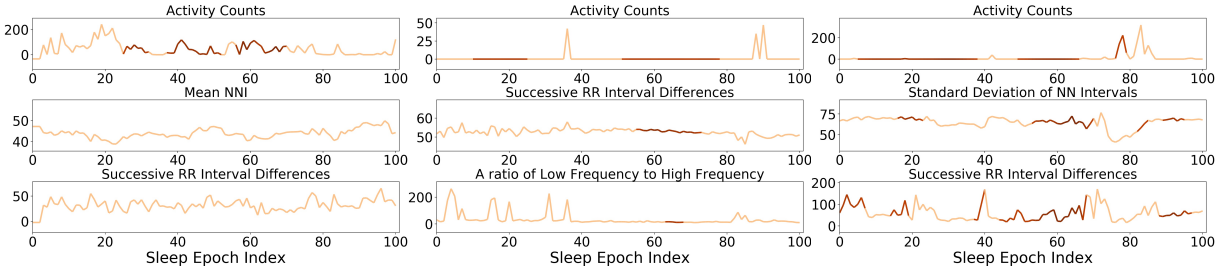
Figure 5.3 The mean of total class activation value for each sleep stage from MESA dataset (ACT-HRV feature) using ResDeepCNN (Addition) in the late-stage fusion.
*Note: ACT : Activity Counts, MNNI : Mean NNI, SDNN: Standard Deviation of NNI, SDSD : Successive RR Interval Differences,VLF: Very-Low-Frequency Band, LF: Low-Frequency Band, HF: High-Frequency Band, LF/HF: The ratio of Low Frequency to High Frequency, SPI: The Signal Power Intensity

[2]https://gradcamvisual1.azurewebsites.net/



(a) **(b)** **(c)**

Figure 5.4 **(a)** A user study of three-stage classification accuracy calculated per participant-wise for Non-CAM and CAM assisted visualization. **(b)** The answer of *The machine-assisted visualization helped me to understand the difference between each sleep stage*. **(c)** The breakdown of classification accuracy calculated for each sleep stage.



(a) Wake **(b)** NREMS **(c)** REMS

Figure 5.5 The Grad-CAM plot of three selected examples from MESA dataset (ACT-HRV feature) using ResDeepCNN (Addition) in the late-stage fusion. Each row is the activation map for the input clinical features.

In each phase, users would first watch the training videos; then they would be asked to recognise nine randomly selected sleep epochs that did not belong to the training videos. At the end of the test, users will be informed of their sleep stage classification accuracy.

There are no established rules for sleep stage classification using movement and cardiac sensing data. To conduct a pilot study, 25 participants were recruited from Amazon's Mechanical Turk. The task took, on average, 20-45 minutes to complete and participants were compensated USD 7.00. All procedures received ethical approval from the University's ethical review board and the Research Ethics Committees (RECs). A total of 25 subjects responded to the questionnaire. Figure 5.4 (a) shows the classification accuracy of CAM-assisted sleep stage classification is higher than the Non-CAM sleep stage classification. This chapter further analysed the results in detail by sleep stages in Figure 5.4 (c). As can be seen, CAM-assisted visualisations improved human recognition accuracy in all sleep stages. This chapter also designed a five-point Likert scale question to test whether the system can improve the user's understanding of visualisation. The results showed that the majority of participants either *Strongly Agree* or *Agree* that the machine-assisted visualisation helped them to understand the difference between each sleep stage.

5.6. Discussion

5.6.1. Simple Fusion Method and Fusion Strategy

All three fusion strategies adopted the concatenation method. In the early-stage fusion, the backbone network fused the latent features from each modality at every convolutional layer. All models studied in this paper surpassed previous studies, except for the models using the bilinear method. With the Apple Watch dataset, the skip connection numerically improved the performance for models in early-stage fusion, but the performance with the MESA dataset decreased. A possible explanation for this might be that simply adding a skip connection may not benefit the model prediction performance in early-stage fusion when the training data is sufficient.

In the late-stage fusion, the concatenation method numerically improved the prediction performance of all metrics for ResDeepCNN compared with the early-stage fusion. In terms of the concatenation method, parameter size and reasoning time increased by five times and six times, respectively, but the performance did not increase by that much. The addition method in late-stage fusion achieved the highest performance with the Apple Watch dataset. This may indicate the benefit of keeping latent features separate, and fusing them at a higher level might produce better results. Another possible explanation for this is the increase in network parameters.

For the hybrid fusion, the intermediate cardiac features are first fused using the early stage and then fused with movement sensing representation at a later stage. With the Apple Watch dataset, the simple operation method produced the same or better results compared with early-stage fusion. The methods in this category increased model parameters and inference time, but the classification performance hardly improved. The addition method aggregated the modal representation at the later stage, and similarly, it only obtained comparable results. A similar pattern was observed with the MESA dataset. These findings suggest that the effectiveness of simple operation methods may be affected by fusion strategies.

5.6.2. Complex Fusion Method and Fusion Strategy

Bilinear pooling turned out to be the weakest fusion method in the hybrid fusion strategy. This is a particularly interesting result because the tensor-based methods showed improvements in the multimodal fusion literature, such as with the task of visual question answering [193]. It is possible that these results could be due to the failure to use the CNN network to learn about a post-bilinear latent feature. The model parameters were too large to be suitable for mobile computing scenarios. The cost of exploring the potential solutions exceeded the benefits.

With the attention mechanism, with the MESA dataset using the ACT-HRV feature set, the mean F1 and Cohen's κ of ResDeepCNN using the Attention-on-Mov method were statistically higher than those in early-stage fusion. With the Apple Watch dataset, the same method can also produce comparable results to the highest-performing method in the late-stage fusion, and the

inference speed is three times faster. Moreover, the time deviation value was balanced in the prediction of each sleep stage.

Compared to the Attention-on-Car method, the results demonstrated a similar pattern with the MESA dataset, that is, applying attention weights to the latent features of movement produced higher results. It can therefore be assumed that the attention method improved the network's ability to learn better representations that can benefit three-stage sleep classification by adjusting the weights of movement sensing representations, as it was difficult to discern REM sleep and NREM sleep using movement sensing alone.

5.6.3. Model Selection

The model parameters, model architecture, model inference time and model performance were the key considerations for the model selection in ubiquitous computing. In addition to these factors, the time deviation should also be considered. It reflected the bias of the model prediction for each sleep stage. With imbalanced sleep data sets, biased models may constantly overestimate the duration of certain sleep stages and may lead to unreasonable health decisions. The mean and standard error of time deviation should be as close to zero as possible. In terms of inference time, predicting sleep data for a whole night using the late fusion strategy was 6.2 times slower than when using the early-stage fusion strategy. As the model calculates the feature maps of each input channel individually using the convolution method, the time consumption of this method was related to the number of intermediate input features. In addition, for the design of the fusion strategy, the ratio of model parameters to inference time is also a consideration for performance evaluation. For instance, the parameters of the bilinear model were 77 times larger than the early-stage fusion models, but the speed was only three times slower. because most of the model parameters belonged to a fully connected layer in the bilinear module, which reduced the feature dimension of the feature matrix (the results of the outer product).

Since sleep is generally considered a nocturnal behaviour, most people are used to charging their phones at night. In addition, sleep analysis does not need to generate real-time feedback, so sleep analysis algorithms are less constrained by computing resources. But the trade-off between performance improvement and computing resource consumption is still an important consideration. From the results in Appendix B.2.2, using raw accelerometer data yields comparable results but increases model parameters and inference time compared to using handcrafted features. It is considered a sub-optimal solution for long-term sleep stage monitoring.

In summary, if the movement and cardiac sensing data can be transmitted to the smartphone, the ResDeepCNN with the addition method in the late-stage fusion may be a feasible model for everyday use. Since it achieved the highest F1 score and with a balanced time deviation on each sleep stage. In scenarios with limited computing resources, such as smartwatches, using the ResDeepCNN model in the early-stage fusion may be a practical choice. The cost of using the late-stage fusion has increased by nine times in inference time. In mobile computing scenarios, power consumption is also a key consideration. In real-world deployments, sleep mostly occurs

at night, and data collected via smart wristbands or smartwatches can be processed on the phone. Power consumption can be easily solved by plugging in the charging cable.

5.6.4. Cross Dataset Comparison

Based on activity counts and HRS features, the highest-performing model for the MESA dataset achieved higher performance than the highest such model for the Apple Watch dataset. Not only were the classification metrics of the MESA dataset higher than these of the Apple Watch dataset, but the standard error on the MESA dataset was smaller. Furthermore, in terms of fusion strategies and methods, none of the methods outperformed the others by large margins on both datasets. But some improvements are still statistically significant. It seems possible that these differences were due to two reasons. The first reason was that the Apple Watch dataset contained far fewer subjects compared with the MESA dataset. The second possible reason was the differences in data acquisition equipment and data pre-processing methods. The HR sensing module of Apple Watch dynamically calculated the HR data within two-five seconds, whereas the cardiac sensing used in the MESA dataset is IHR. The higher resolution IHR data might provide more discriminant information in order to discern three sleep stages.

5.6.5. Exploratory Research of Visualisation

One of the objectives of this work is to better understand in what way the decision-making process of neural networks using multimodal fusion techniques on sleep stage classification can be understood by humans. Based on Grad-CAM scores, a simplified visualisation method was adopted in this study and an exploratory study with users was performed. The visualisation tool can highlight both the key temporal signal segments and the most discriminant feature channels, and by providing essential clues/patterns for users it can serve as an assistant tool in understanding different sleep stage signals.

Compared with highlighting the temporal steps, the channel dimension reduction retained the minimum number of discriminant channels for sleep stage classification, which showed a combined reduction that could improve user understanding. Based on the repeated patterns of activity count and HRV features during a continuous sleep period, the results demonstrated that reducing information overload could improve human understanding of three-stage sleep recognition performance. The visualisation increased users' understanding in terms of the neural network decision-making process to some extent.

Wake recognition accuracy was higher than in the other two sleep stages, which indicates wrist movement is obvious to classify wake. This result is consistent with the known capability of actigraphy can be used to distinguish between wake and sleep. The highlighted patterns that appeared in the continuous sleep stage agreed with previous sleep physiology findings to some extent.

The modelling process has window bias and Grad-CAM modelling bias. For example, this study adopted a window length of 101 (50.5 minutes). The highlighted areas will move backwards as the window moves forwards. The network is capable of locating signatures in a

window that are meaningful for the current sleep stage recognition. This is very different to the annotation process using high-resolution (over 100Hz) PSG data. An interesting question for future work is to investigate whether these patterns have physiological meaning.

Many existing interpretation and visualisation techniques have been developed for visual data, yet it is unclear whether these methods are suitable for explaining sleep time-series data. This is a pilot study to investigate whether Grad-CAM applied to sleep time-series wearable data may be useful to humans. It is one of the mainstream methods used in visual and text data but is subject to the network structure. For instance, it is difficult to highlight the important areas in channels in early-stage fusion models without substantial changes to the method. On the other hand, it is difficult for humans to understand the highlighted patterns in the time dimension.

The results demonstrated an interesting phenomenon, e.g., three out of 25 users experienced negative impacts on their understanding. In addition, the questionnaire data also highlighted that not everyone agreed that the system helped them to understand the difference in the patterns between sleep stages. A possible explanation for this might be that the visualisation may not be understood by every person, or the training and testing process may be problematic. Future studies may consider conducting experiments with more detailed personalised questions.

The visualisation was designed as a pilot study to understand the decision-making process of multimodal fusion for sleep stage classification. So, this chapter did not investigate other mainstream interpretation methods, nor did it conduct large-scale user research. Future work may consider investigating the other interpretation methods such as SHAP, Anchor, etc. or even create a special algorithm for time-series data to facilitate intuitive understanding by humans.

5.6.6. Comparison with Previous Work and Implications

Fusion Specifics			Performance Metrics			Time Deviation (min.)		
Dataset and Feature Set	**Fusion Stage**	**Model**	**Accuracy (%)**	**Cohen's κ**	**Mean F1 (%)**	**Non-REM sleep**	**REM sleep**	**Wake**
MESA (ACT-HRV)	Early-stage Fusion	CNN (101) (Zhai, 2020)	76.0 ± 1.0	58.6 ± 1.9	68.1 ± 1.3	14.9 ± 6.7	-0.5 ± 4.3	-14.4 ± 5.8
	Hybrid Fusion	ResDeepCNN (Attention-on-Mov)	**79.8 ± 0.9**	**65.5 ± 1.8**	**73.3 ± 1.3**	-0.9 ± 6.4	6.1 ± 3.7	-5.1 ± 6.3
Apple Watch (Activity Counts, HR, Time)	Early-stage Fusion	MLP (Walch, 2019)	72.1 ± 2.4	23.7 ± 4.4	47.8 ± 3.6	84.2 ± 17.2	-65.4 ± 15.0	-18.8 ± 6.1
Apple Watch (Activity Counts, HRS)	Late-stage Fusion	ResDeepCNN Addition	**78.2 ± 2.3**	**52.0 ± 5.8**	**66.5 ± 3.8**	1.9 ± 11.7	4.7 ± 12.6	-6.5 ± 7.3

Table 5.5 Three-stage sleep classification prediction results compared with previous work evaluated at subject level (mean ± standard error at 95% confidence interval) during the recording period.

Table 5.5 and Figure 5.6 show the results compared with previous works. The results demonstrated that the use of multimodal fusion strategies and fusion methods can improve model prediction performance. With the three-stage sleep classification dataset, the class imbalance issue causes the classifier to be biased towards the majority class, which is NREM sleep.

To compare with the works in previous chapters, this chapter conducted ten runs of the model with the highest mean F1 and the baseline model for each dataset, respectively. Each run used a different random number seed. Compared with the work completed in chapter 3 for the MESA dataset, the accuracy (p<0.001), Cohen's κ (p<0.001) and mean F1 score (p<0.001) improved statistically significantly based on the ACT-HRV feature set. There were also statistically signifi-

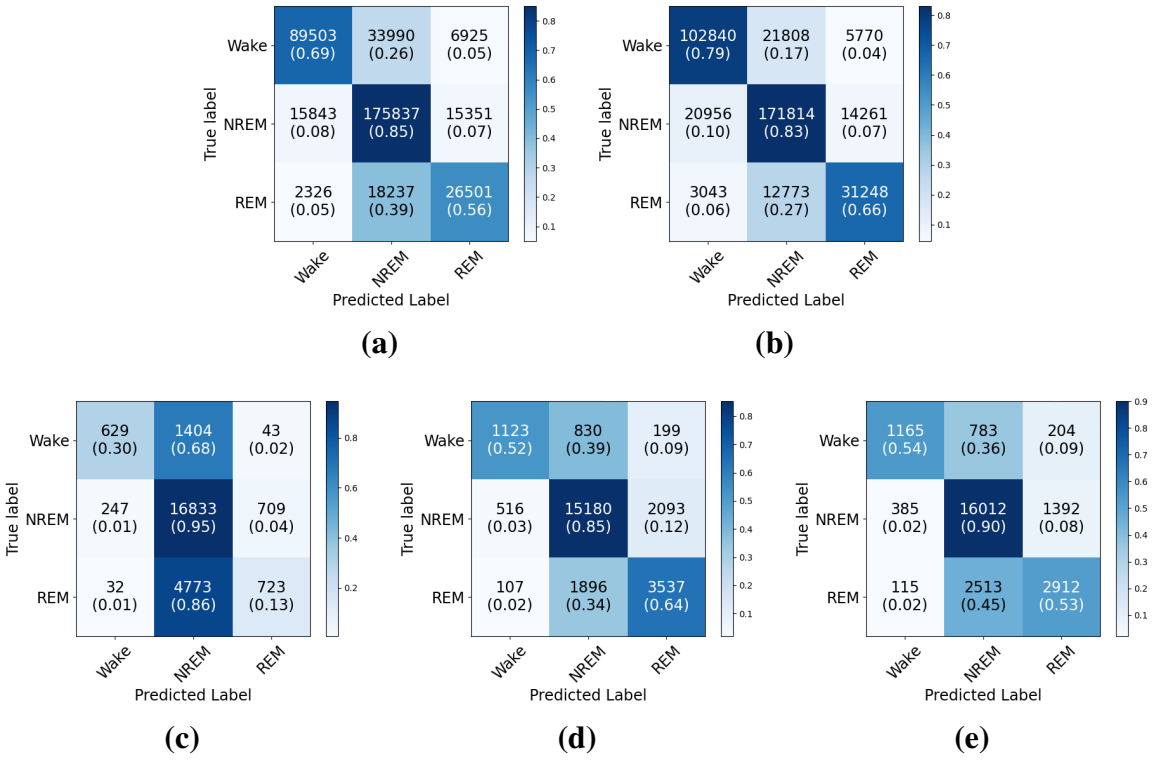


Figure 5.6 (a) The CNN(101) and early-stage fusion based on the MESA (ACT-HRV) dataset used in [5]. **(b)** Hybrid fusion using ResDeepCNN (Attention-on-Act) based on the MESA (ACT-HRV) dataset **(c)** Walch et al. using multiple layer perception based on activity counts, HR and circadian time [6] **(d)** Late-stage fusion using ResDeepCNN (Addition) based on the Apple Watch Dataset **(e)** Late-stage fusion using ResDeepMixCNN (Concatenation) using raw accelerometer data and HRS based on the Apple Watch dataset.

cant improvements in accuracy (p<0.001), mean F1 score (p<0.001), and Cohen's κ (p<0.001) for the Apple Watch dataset [6]. These improvements suggest that the proper multimodal fusion strategy and method can improve the robustness of the model, which is a step towards automated three-stage sleep classification. The findings reported here suggest that reasonable performance may be achieved using the movement and cardiac features derived from consumer/research-grade wearable devices.

5.7. Summary

Using actigraphy to monitor sleep-wake has existed for many decades. But, for sleep stage classification, the sleep study has relied on the PSG equipment, which is an expensive, burdensome, laboratory sleep monitoring method. This limits many research advances in sleep and health. In recent years, more and more new products using ubiquitous computing technology have passed FDA clearance, such as the Apple Watch irregular heart rate detection function [290]. The achievements of these wearables provide important instrumental tools for the study of longitudinal sleep and health. The core contribution of this chapter lies in the systematic study on how to better integrate multi-modal data to monitor three-stage sleep that may use consumer/research-grade wearables. Through the study, the performance of several new models

exceeded the previous benchmark studies significantly. This chapter provided a new multimodal fusion benchmark for the ubiquitous computing community, which has provided the potential for the use of consumer wearables to study the sleep health of large-scale populations in the future. One of the motivations for this work was to respond to previous research that called for more accurate and transparent sleep stage sensing algorithms on consumer wearable devices [6]. The implication of this chapter's work is to encourage more researchers, consumers, and application developers to use consumer/research-grade wearables to study and understand sleep and health.

Chapter 6. Conclusion

6.1. Discussion

Long-term, low-cost, large-scale sleep monitoring has profound implications for health and medical research. Chronic non-communicable diseases such as neurodegenerative diseases will impair social productivity and limit the growth of the economy, and elderly care will increase the financial burden on public finances as well as families. The early onset of several neurodegenerative diseases may be accompanied by abnormal sleep behaviour such as Parkinson's [34, 291]. REM sleep disorder has been found to be one of the early symptoms of Parkinson's disease [35, 291]. Using wearable sensors to detect sleep stages could enable early-onset detection, treatment, or ongoing care that may reduce the severity or delay disease onset.

Moreover, in genetic epidemiology studies, assessing genetic susceptibility to complex human diseases and locating disease markers in genetic variation has profound implications for medicine [292, 293]. Population-wide genome-wide association study (GWAS) is one of the commonly used methods for conducting analyses [294]. This approach typically requires a large number of cases and controls to obtain sufficient statistical power, which is critical in the design phase of genetic associations [295, 293, 294]. Therefore, a user-acceptable solution for large-scale sleep architecture assessment is an important tool to address research gaps to explain genetic differences in sleep. To accurately assess sleep architecture, wearable sensing technologies may play a key role. The outputs of this paper could serve as a starting point for revolutionising sleep monitoring at the population level.

One of the main goals of digital health is to develop an automated monitoring system for long-term non-intrusive sleep stage monitoring, which is sufficient and robust to be deployed in a free-living environment. In most cases, the target application needs to be integrated into people's daily lives, where systems have to abide by practical usability and privacy constraints. In recent years, the development of consumer/commercial wearable devices (e.g., Huawei Watch and Fitbit Band) has provided unparalleled opportunities for large-scale measurement of physiological parameters and human activities. Many of these devices use undisclosed algorithms and the data processing pipeline is not under user control. These obstacles make clinical validation of these devices difficult. The lack of clinical validation limits their application in studying the impact of lifestyle and sleep stages in free-living scenarios as well as in clinical and occupational settings. Throughout the work, this thesis systematically demonstrates the feasibility of using commercial and research-grade multi-modal wearable devices for sleep stage inference. The work conducted in this thesis can be summarised in Figure 6.1. The details are given as follows.

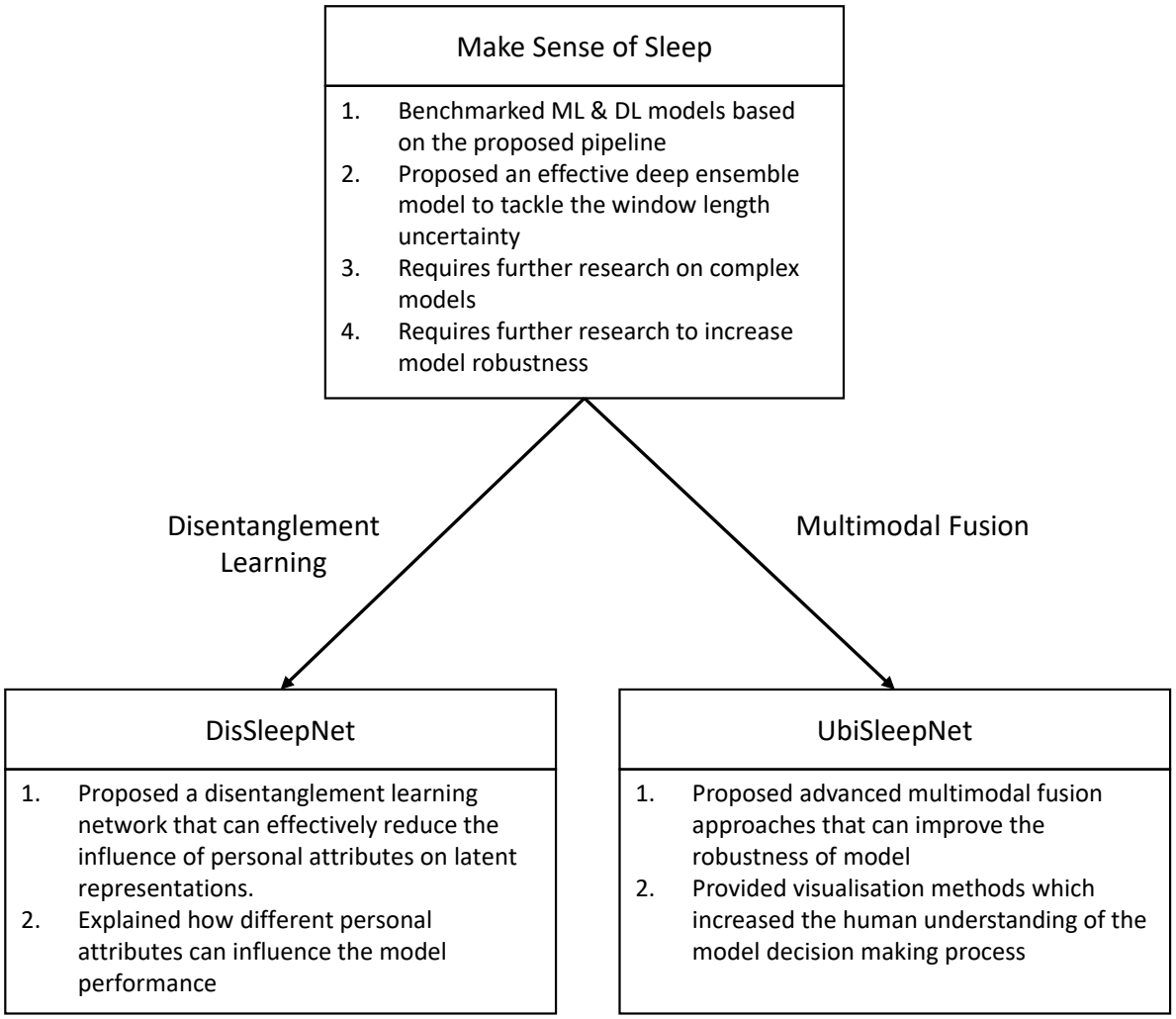


Figure 6.1 Thesis Road Map

In **chapter 2**, a comprehensive review covered the state-of-the-art sleep tracking technologies, ranging from consumer/commercial products to the latest research prototypes that can monitor sleep outside the sleep laboratory. This chapter also analyses the strengths and limitations of each technology. Then, the author reviewed state-of-the-art machine learning and deep learning algorithms that may be used for sleep stage classification. **Chapter 3** systematically investigated the classical machine learning and deep learning models in the largest sleep stage dataset, which includes the movement and cardiac sensing data to date. This chapter demonstrated the feasibility of using multimodal sensing data to classify sleep stages at different levels of granularity. The experimental results also revealed the important score of each clinical feature for different stages. This chapter also proposed an ensemble model that can significantly improve classification performance by combining neural network models.

Chapter 4 investigated how personal attributes (PAs) may affect model performance. However, it is impractical to collect data covering all sleep disorders and health conditions of all severity with ground-truth annotations. Compared with healthy adults, PAs such as age, obesity,

and sleep-related breathing disorders may cause sensor data to have different patterns. A big challenge for supervised machine learning models is that the model's performance can degrade in populations with unseen health conditions. To alleviate the impacts of PAs, this chapter proposed a novel disentangled representation learning network, namely DisSleepNet (detailed information is shown in **Chapter 4**). The model can learn representations that are invariant to specific PAs, thereby reducing model performance degradation when applying the trained model to subjects with completely different ranges of PAs. The DisSleepNet is developed on the framework of variational autoencoders, which can encourage independence between different latent factors in the representation space. This reduces the influence of personal attributes on features. To learn PA free features, two disentanglers were introduced to separate the representations into PA-specific features and PA-free features. As the independence of two features is not always guaranteed, This chapter then proposed to use the modified independence excitation mechanism to further maximise the independence of these two features. Compared with the use of the backbone neural network in **chapter 3**, DisSleepNet can significantly reduce the PA effects on model performance.

For long-term sleep monitoring, ubiquitous sensing may be a solution, since the cardiac and movement sensing data can be easily acquired from research-grade or consumer-grade devices (e.g., Apple Watch). In **chapter 3**, the results revealed that among all granularities of sleep stages, the three-stage sleep classification was considered the most promising task. The performance of representation learning can be further improved with the use of multimodal fusion techniques. However, how best to fuse the data for the greatest accuracy remains an open question. In **chapter 5**, this thesis comprehensively studied DL-based advanced fusion techniques consisting of three fusion strategies alongside three fusion methods for three-stage sleep classification based on two datasets where the evaluation was done with respect to PSG, including a study where the evaluation was carried using a commercial device (Apple Watch). Two deep convolutional neural networks were introduced that could capture high-level representations and outperform the networks used in the benchmark study. Secondly, this chapter evaluated three fusion strategies including the *early-stage, mid-stage* and *late-stage* fusion, and three fusion methods, including the simple operations, attention-based methods and tensor-based methods. The attention method proposed in the hybrid fusion strategy can capture complementary inter-modality information and filter out less important information. Its performance is higher or comparable to other baseline methods. In terms of transparency, there are no established rules for three-stage sleep classification using cardiac and motion-sensing data. This chapter also investigated the visualisation methods to explore the decision-making process of the multimodal fusion model for three-stage sleep classification. The exploratory user research demonstrated that the Grad-CAM based sleep data visualisation can be understood and used by humans, which facilitates the transparency of using DL in sleep health research.

This chapter systemically demonstrated that the three-stage sleep classification is the most realistic and promising task by using the cardiac and movement sensing data that are generously available on many profound smart wearables, such as the Apple Watch. The experimental results

demonstrated important evidence indicating that the three-stage sleep can be reliably classified by fusing cardiac/movement sensing modalities, which may potentially become a practical tool for large-scale sleep stage assessment studies or long-term self-tracking on sleep.

6.2. Limitation and Future Work

The main contribution of this thesis is to show that with novel computational methods, commercial/consumer-grade wearables could be used for sleep stage monitoring with a certain degree of accuracy and robustness. This work unlocked the potential of using ubiquitous computing technologies to conduct large-scale sleep and health studies with transparent open-source sleep stage monitoring algorithms. However, more work is yet to be done in order to bring the sleep monitoring system to a greater level.

Given the scope of objectives for this thesis, in **chapter 3**, the experimental design did not explore in detail how different models perform in participants with sleep disorders versus healthy populations or highlight the differences between them. Similarly, the model design did not exploit the well-known reciprocal interaction model which was introduced by McCarley, and Hobson [296], which describes ultradian periodicity, the approximately 90-min sleep cycle, which indicates that NREM-REM stage transitions are regulated by both cholinergic and monoaminergic neuronal structures. This inherent sleep architecture shall be explored in future work to improve model performance. Furthermore, the ensemble model used in this paper is just an example of how advanced machine-learning approaches can be improved by using a novel approach tool.

Furthermore, the data processing pipeline for the entire thesis did not enforce strict quality control on the polysomnography data, which could have led to the models performing more poorly than if those practices and more stringent exclusion criteria had been applied. For instance, the empirical results found that a total of 30 subjects (about 2% of the total cohort) did not have any REM epochs at all. Similarly, on a small percentage of participants (less than 1%), accuracy scores were very low ($< 45\%$). After the post-hoc visual inspection of those cases, the visualisation suggests their sleep patterns were abnormal, and five of them had a reduced number of sleep transitions. However, for the purposes of this chapter, those participant results were included in the final performance metrics. To exclude non-wear time and activity measurement failure from actigraphy data, the human expert annotated tags are considered as the selection criteria. Full processing pipelines shall integrate automated non-wear time and data corruption detection algorithms in the preprocessing phase. In this chapter, the Grad-CAM based visualisation increases the user's understanding of the neural network decision-making process to a certain extent. However, this thesis didn't perform a sleep physiological analysis of highlighted regions because these regions change dynamically over time. Future work should investigate the physiological significance behind these patterns.

A number of limitations need to be noted regarding the **chapter 4**. In addition to the VAE framework, there are tools and models that can disentangle features in the representation space, such as GANs which typically require large amounts of training data to ensure the convergence

of the training process. In this work, the model is limited to VAE-based disentanglement frameworks, and future research can consider investigating GAN-based models to disentangle PA-specific and PA-free features. Another limitation is that this thesis did not test other variants of CNNs or recurrent neural network based models. Since the purpose of this work is to understand whether the use of disentanglement learning can reduce the influence of personal attributes on the model. In the future work, these models will be investigated and the unsupervised domain adaptation methods will be investigated too. Furthermore, aside from factors such as age, obesity, and sleep apnoea, REM sleep disorders and heart diseases may also negatively influence model performance, which should also be investigated.

In spite of its limitations, the proposed model performed well on a single PA, but the effect was not significant when jointly disentangling multiple PAs. One of the reasons is that age may negatively affect the learning of other PAs during multi-label learning. In addition, the distribution of individual PAs in the training samples is non-uniform, and this work did not explore whether the influence of the values of PAs on the model is linear. In future work, the author plans to investigate new loss functions to ensure that those trivial PAs do not affect the overall performance during the training. In addition, the sample size based on the final experimental setting of this work is small. Future work should consider conducting experiments on other datasets, if available.

6.2.1. Domain Adaptation for Wearable Sensing Based Sleep Monitoring

One of the most interesting and important challenges still to be solved in the field of ubiquitous computing for health monitoring is addressing domain discrepancy. This challenge means that models trained on a specific device experience performance degradation when applied to other devices that may have different sensing mechanisms or slightly different data processing pipelines.

This is especially important considering that current models learned using a specific device or a specific population may not generalise well when the target data distribution is too different. As shown in **chapter 4**, machine learning models derived from healthy adult data may not be suitable for diseased populations (different distributions), which may produce misleading results. Personal attribute-invariant feature learning via disentanglement approaches is just one way to alleviate the problem of domain discrepancy. Future work might consider using appropriate domain adaptation methods to address this issue by leveraging multiple datasets collected from different populations. For example, several previous works on human activity recognition have applied methods to reduce the distributional differences between source and target domains while respecting the learned discriminant information by establishing maximum mean difference (MMD) and discriminant distance [297, 298].

6.2.2. Data Driven Approaches

Another limitation of this thesis is that the MESA and the Apple Watch dataset only include adult participants. Thus the results cannot be generalised to teenagers or children. Further, as with all

studies in this area to date, all inferences are derived in laboratory settings, whilst the potential applications are in a free-living environment. Good quality 'ground truth' data collection in free-living environments is complex and expensive, although it is interesting to explore the possibility of larger studies using ambulatory PSG or wireless EEG for this purpose. Thus, this thesis encourages these companies to be more transparent about the way they collect data and the algorithms they use. In this thesis, the author only tested handcrafted features that are derived using sliding window methods based on sensing signals. The greatest advantage of deep learning approaches is that they can automatically extract the most effective representation for the task from raw data through an end-to-end learning method. The use of handcrafted features may not maximise the advantages of multi-modal fusion. Since these handcrafted features are limited to human knowledge, they may not be the best features to represent different sleep stages. The additional experiments conducted in **chapter 5** demonstrated that three-stage sleep classification can be further improved using raw accelerometer data and HRV features which demonstrates the potential of end-to-end approaches. Although the improvement has not yet reached statistical significance. Future work may consider the possibility of learning deep representations from ECG and accelerometer raw data in an end-to-end learning manner. Using raw sensing data may reduce the window length and thus reduce ambiguity in the learning process.

Appendix A. Benchmark Study Performance By Modalities And Methods

A.1. Epoch By Epoch Performance Metrics

Task 1: Wake, Sleep								
Method Specifics			**Performance Metrics**					
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's k
Multimodality	♥ 🏃 *[HR/HRV, Actigraphy]*	LSTM (51)	84.2	67.2	84.7	93.0	88.6	63.1
		LSTM (101)	84.2	67.1	84.6	93.0	88.6	63.1
		CNN (101)	84.1	66.2	84.3	93.3	88.6	62.7
		Maximum selection	**85.2**	68.2	85.2	**94.0**	**89.4**	65.3
		Mean over classifiers	85.2	**69.3**	**85.6**	93.4	89.3	**65.5**
Single Modality	♥ *[HR/HRV]*	LSTM (101)	**79.4**	**60.2**	**81.4**	89.2	**85.1**	**51.8**
		LSTM (51)	78.8	54.3	79.6	**91.4**	85.1	49.2
		CNN (101)	78.4	59.0	80.8	88.4	84.4	49.6
	🏃 *[Actigraphy]*	LSTM (101)	**84.7**	66.0	84.4	**94.3**	**89.1**	**63.9**
		CNN (101)	84.3	66.4	84.4	93.5	88.7	63.1
		LSTM (51)	84.3	**69.7**	**85.5**	91.7	88.5	63.6
Task 2: Wake, NREM, REM								
Method Specifics			**Performance Metrics**					
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's k
Multimodality	♥ 🏃 *[HR/HRV, Actigraphy]*	LSTM (51)	76.3	85.7	72.3	69.3	70.6	60.9
		LSTM (101)	76.1	85.4	72.9	67.3	69.3	60.2
		CNN (101)	75.9	85.6	71.5	70.0	70.4	61.0
		Mean over classifiers	**78.2**	**86.6**	**75.1**	**70.8**	**72.6**	**64.4**
		Maximum selection	77.9	86.6	74.2	70.8	72.2	63.9
Single Modality	♥ *[HR/HRV]*	LSTM (101)	**73.7**	**84.1**	**69.1**	**66.1**	**67.3**	**51.2**
		LSTM (51)	72.7	83.6	68.3	64.3	65.9	47.3
		CNN (101)	71.0	83.3	65.3	65.2	65.2	47.7
	🏃 *[Actigraphy]*	LSTM (101)	**71.3**	**80.8**	58.5	**52.9**	**50.5**	**52.5**
		CNN (101)	70.9	80.2	**76.5**	52.0	49.8	51.0
		LSTM (51)	70.8	80.4	48.4	52.3	49.8	51.0
Task 3: Wake, Light Sleep, Deep Sleep, REM								
Method Specifics			**Performance Metrics**					
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's k
Multimodality	♥ 🏃 *[HR/HRV, Actigraphy]*	LSTM (51)	70.4	87.7	65.5	54	54	56.8
		LSTM (101)	70.4	87.2	66	52.1	**54.2**	54.4
		CNN (101)	69.1	87.3	63.2	53.6	53.6	54.4
		Mean over classifiers	**71.7**	87.9	67.5	53.8	53.5	**58.8**
		Maximum selection	71.3	**88**	**68**	**54.4**	54.1	58.1
Single Modality	♥ *[HR/HRV]*	LSTM (101)	**67.4**	**86.3**	**62.9**	**50.9**	**52.2**	**46.9**
		LSTM (51)	66.2	85.8	61.9	48.9	49.4	43.4
		CNN (101)	64.4	85.5	59.2	49.8	49.6	43.4
	🏃 *[Actigraphy]*	LSTM (101)	**64.1**	**83.7**	34.4	**38.8**	**35.6**	**35.1**
		CNN (101)	64.0	83.7	**42.5**	38.8	35.6	35.1
		LSTM (51)	63.6	83.4	36.3	38.4	35.3	34.6
Task 4: Wake, REM, N1,N2,N3								
Method Specifics			**Performance Metrics**					
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's k
Multimodality	♥ 🏃 *[HR/HRV, Actigraphy]*	LSTM (51)	63.9	89.0	55.9	43.4	42.5	59.0
		LSTM (101)	63.8	89.0	55.6	43.3	43.2	59.7
		CNN (101)	63.2	89.0	55.6	44.8	**44.1**	58.4
		Mean over classifiers	**65.6**	**89.6**	**59.9**	44.7	42.8	**62.3**
		Maximum selection	65.3	89.6	58.4	**45.1**	43.0	61.7
Single Modality	♥ *[HR/HRV]*	CNN (101)	**55.6**	**86.7**	**48.3**	**38.4**	**38.8**	**38.1**
		CNN (21)	55.5	86.6	47.0	36.8	35.0	37.6
		CNN(50)	54.3	86.1	47.1	35.0	34.6	32.9
	🏃 *[Actigraphy]*	LSTM (101)	**57.0**	**86.4**	24.4	31.7	27.0	**50.0**
		CNN (101)	57.0	86.4	**32.1**	**31.8**	**27.2**	49.5
		LSTM (51)	57.0	86.3	29.8	31.6	27.0	49.9

Table A.1 Task 1-4 classification results by multimodal and single modality approaches, using epoch-by-epoch performance metrics. This table complements what was found on the main text and reported in Figures 6 and 7; Actigraphy modality: 🏃, HR/HRV modality: ♥; Three different tasks: Task 2: 3 stages, Task 3: 4 stages, Task 4: 5 Stages

A.2. Sleep Stage Classification Results Measured In Sleep Period

Task 1: Wake, Sleep

Method Specifics			Performance Metrics						Time Deviation*	
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	Sleep
Multimodality	♥ 🏃	CNN (51)	**85.1 ± 1.1**	50.1 ± 2.5	88.7 ± 1.0	92.8 ± 1.2	90.0 ± 1.0	44.7 ± 2.2	-19.6 ± 5.8	19.6 ± 5.8
		CNN (101)	**85.1 ± 1.1**	**50.9 ± 2.5**	**88.9 ± 1.0**	92.5 ± 1.2	90.0 ± 1.0	**44.8 ± 2.2**	**-17.4 ± 5.9**	**17.4 ± 5.9**
		CNN (21)	84.9 ± 1.0	48.9 ± 2.2	88.3 ± 1.0	**93.1 ± 1.0**	**90.1 ± 0.9**	44.0 ± 2.0	-22.2 ± 5.4	22.2 ± 5.4
Single Modality	♥	CNN (51)	**81.7 ± 1.2**	43.0 ± 2.3	86.3 ± 1.2	**91.3 ± 1.3**	**87.9 ± 1.1**	35.3 ± 1.9	23.2 ± 7.3	-23.2 ± 7.3
		CNN (101)	81.7 ± 1.3	43.7 ± 2.3	86.5 ± 1.1	91.0 ± 1.4	87.8 ± 1.2	**35.7 ± 2.0**	21.4 ± 7.4	-21.4 ± 7.4
		LSTM (21)	80.9 ± 1.1	**48.4 ± 2.2**	**87.1 ± 1.2**	89.0 ± 1.2	87.3 ± 1.0	35.4 ± 1.8	**8.9 ± 6.8**	**-8.9 ± 6.8**
	🏃	CNN (101)	**85.5 ± 1.0**	46.9 ± 2.5	88.3 ± 1.0	**93.9 ± 0.9**	**90.5 ± 0.8**	43.5 ± 2.2	25.5 ± 5.5	-25.5 ± 5.5
		CNN (51)	85.3 ± 1.1	**50.7 ± 2.4**	**88.8 ± 1.0**	92.9 ± 1.1	90.2 ± 0.9	**45.4 ± 2.2**	19.7 ± 5.7	-19.7 ± 5.7
		LSTM (101)	83.9 ± 1.0	46.2 ± 2.5	88.5 ± 0.9	91.3 ± 1.1	89.4 ± 0.9	38.6 ± 2.3	**13.3 ± 5.4**	**-13.3 ± 5.4**

Task 2: Wake, NREM, REM

Method Specifics			Performance Metrics						Time Deviation*		
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	NREM
Multimodality	♥ 🏃	CNN (101)	**74.9 ± 1.1**	**82.1 ± 0.8**	**68.7 ± 1.3**	64.5 ± 1.3	**62.9 ± 1.3**	**47.0 ± 2.0**	-14.4 ± 5.8	**-0.5 ± 4.3**	14.9 ± 6.7
		LSTM (51)	74.2 ± 1.1	81.7 ± 0.8	66.5 ± 1.4	62.8 ± 1.2	61.5 ± 1.3	45.1 ± 2.1	**0.3 ± 5.6**	-12.5 ± 3.7	12.2 ± 6.2
		CNN (51)	73.6 ± 1.2	82.5 ± 0.8	67.7 ± 1.3	**65.5 ± 1.3**	62.6 ± 1.4	47.0 ± 2.1	-4.7 ± 5.7	7.5 ± 5.8	**-2.9 ± 7.5**
Single Modality	♥	LSTM (101)	**72.9 ± 1.3**	81.0 ± 0.8	**64.7 ± 1.5**	**60.1 ± 1.3**	**58.7 ± 1.5**	**37.2 ± 2.4**	-6.6 ± 7.0	**-11.8 ± 4.0**	18.4 ± 6.9
		LSTM (51)	72.4 ± 1.2	**81.1 ± 0.8**	63.0 ± 1.5	59.4 ± 1.3	57.6 ± 1.4	32.9 ± 2.3	5.7 ± 6.9	-19.1 ± 4.1	**13.4 ± 6.8**
		CNN (51)	71.1 ± 1.2	80.3 ± 0.8	63.6 ± 1.5	59.3 ± 1.3	56.6 ± 1.4	33.2 ± 2.1	**-5.5 ± 7.3**	-12.0 ± 5.5	17.5 ± 7.9
	🏃	Linear SVM	**68.7 ± 1.0**	71.5 ± 0.7	**43.3 ± 1.1**	**43.9 ± 0.9**	**40.8 ± 0.9**	**25.1 ± 1.8**	-24.3 ± 6.1	-67.4 ± 3.0	91.7 ± 6.7
		CNN (51)	68.4 ± 1.0	**71.6 ± 0.7**	42.2 ± 1.1	43.8 ± 0.9	40.4 ± 1.0	24.5 ± 2.0	**-17.4 ± 6.2**	**-67.3 ± 3.0**	**84.7 ± 7.0**
		CNN (101)	68.4 ± 1.0	71.4 ± 0.7	42.5 ± 1.1	43.5 ± 0.9	40.1 ± 1.0	24.1 ± 2.0	-19.4 ± 6.3	**-67.3 ± 3.0**	86.7 ± 7.1

Task 3: Wake, Light Sleep, Deep Sleep, REM

Method Specifics			Performance Metrics						Time Deviation*			
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	Deep Sleep	Light Sleep
Multimodality	♥ 🏃	LSTM (51)	**66.5 ± 1.1**	**84.1 ± 0.7**	53.8 ± 1.4	50.2 ± 1.0	**47.4 ± 1.1**	**44.8 ± 2.2**	10.5 ± 5.8	-7.5 ± 3.9	-36.2 ± 3.5	**33.2 ± 6.7**
		LSTM (101)	**66.5 ± 1.1**	82.9 ± 0.6	**55.5 ± 1.6**	47.5 ± 1.1	46.2 ± 1.2	40.6 ± 2.1	-7.3 ± 5.5	-25.7 ± 3.6	**-32.5 ± 3.5**	65.5 ± 6.8
		CNN (101)	65.6 ± 1.2	83.7 ± 0.7	54.8 ± 1.5	**49.8 ± 1.1**	47.1 ± 1.1	42.6 ± 2.0	**-1.7 ± 6.3**	**1.5 ± 4.7**	-34.6 ± 3.5	34.7 ± 7.5
Single Modality	♥	LSTM (101)	**64.5 ± 1.3**	**83.2 ± 0.6**	**51.8 ± 1.6**	**47.0 ± 1.1**	**44.7 ± 1.3**	**35.0 ± 2.3**	**4.7 ± 7.0**	**-16.1 ± 3.6**	**-33.7 ± 3.5**	45.2 ± 7.2
		LSTM (51)	64.3 ± 1.2	83.1 ± 0.6	51.0 ± 1.5	46.5 ± 1.0	43.7 ± 1.2	33.2 ± 2.3	8.9 ± 6.6	-18.1 ± 3.9	-36.3 ± 3.5	45.5 ± 6.9
		LSTM (21)	62.8 ± 1.2	82.7 ± 0.6	47.9 ± 1.4	45.5 ± 0.9	42.0 ± 1.0	30.5 ± 2.0	16.7 ± 6.9	-18.8 ± 4.0	-38.5 ± 3.6	**40.6 ± 7.2**
	ENMO	LSTM (101)	**60.1 ± 1.1**	**77.3 ± 0.6**	30.1 ± 1.0	33.5 ± 0.8	29.5 ± 0.9	15.9 ± 1.3	-21.6 ± 5.5	-67.3 ± 3.0	**-39.2 ± 3.6**	**128.1 ± 7.3**
		CNN (101)	60.0 ± 1.1	**77.3 ± 0.6**	29.9 ± 1.0	33.4 ± 0.8	29.3 ± 0.9	16.2 ± 1.4	**-18.2 ± 5.7**	**-67.2 ± 3.0**	**-39.2 ± 3.6**	124.6 ± 7.5
		LSTM (51)	60.0 ± 1.1	**77.3 ± 0.6**	**30.3 ± 1.0**	**33.8 ± 0.8**	**29.7 ± 0.9**	**16.7 ± 1.4**	-18.2 ± 6.2	-67.3 ± 3.0	**-39.2 ± 3.6**	124.7 ± 7.7

Task 4: Wake, REM, N1, N2, N3

Method Specifics			Performance Metrics						Time Deviation*				
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	N3 Sleep	N2 Sleep	N1 Sleep
Multimodality	♥ 🏃	CNN (101)	**59.2 ± 1.2**	**86.5 ± 0.6**	**49.7 ± 1.4**	**42.0 ± 1.0**	**39.1 ± 1.0**	**45.1 ± 1.9**	-8.6 ± 5.8	**5.0 ± 4.8**	-34.3 ± 3.5	78.2 ± 7.4	**-40.3 ± 3.1**
		CNN (51)	58.8 ± 1.2	86.4 ± 0.6	46.6 ± 1.4	41.5 ± 0.9	37.8 ± 1.0	44.7 ± 2.0	**-4.7 ± 5.7**	11.5 ± 5.7	-37.9 ± 3.5	**73.1 ± 7.9**	-42.0 ± 3.1
		LSTM (101)	58.4 ± 1.1	85.9 ± 0.6	44.3 ± 1.4	39.9 ± 0.9	36.7 ± 1.0	43.5 ± 2.0	7.7 ± 5.7	-17.5 ± 3.8	**-32.3 ± 3.5**	88.0 ± 7.1	-45.9 ± 3.0
Single Modality	♥	CNN (21)	**53.8 ± 1.3**	85.0 ± 0.6	38.9 ± 1.3	36.7 ± 0.9	32.1 ± 1.0	**28.7 ± 1.8**	**18.8 ± 8.2**	**-9.4 ± 5.4**	-39.0 ± 3.6	74.0 ± 8.7	-44.4 ± 3.0
		CNN (101)	52.4 ± 1.3	**85.1 ± 0.6**	**42.8 ± 1.4**	**37.6 ± 1.0**	**33.6 ± 1.1**	26.8 ± 1.9	28.9 ± 9.0	-15.7 ± 4.5	**-29.8 ± 3.5**	57.6 ± 8.4	-41.0 ± 3.1
		CNN (51)	50.4 ± 1.3	84.6 ± 0.6	39.2 ± 1.4	35.2 ± 0.9	30.0 ± 1.1	22.9 ± 1.9	62.0 ± 10.4	-30.8 ± 4.8	-36.4 ± 3.5	**46.2 ± 9.8**	**-40.9 ± 3.0**
	🏃	LSTM (51)	**51.3 ± 1.1**	**82.0 ± 0.6**	**21.8 ± 0.9**	28.2 ± 0.8	**22.7 ± 0.8**	27.8 ± 1.9	**2.3 ± 6.2**	-67.3 ± 3.0	-39.2 ± 3.6	153.3 ± 8.1	**-49.1 ± 3.2**
		CNN (21)	50.9 ± 1.2	**82.0 ± 0.6**	21.5 ± 0.9	28.1 ± 0.7	22.5 ± 0.8	27.4 ± 1.8	4.6 ± 6.5	**-67.1 ± 3.0**	-39.2 ± 3.6	150.7 ± 8.1	**-49.1 ± 3.2**
		LSTM (21)	50.9 ± 1.1	**82.0 ± 0.6**	21.1 ± 0.9	**28.3 ± 0.8**	**22.7 ± 0.8**	**27.8 ± 1.8**	5.1 ± 6.1	-67.4 ± 3.0	**-39.1 ± 3.6**	**150.5 ± 7.9**	**-49.1 ± 3.2**

Table A.2 Sleep stage classification results during sleep period (mean ± standard error at 95% confidence interval and predicted minutes by multimodal and single modality approaches; Actigraphy modality: 🏃, HR/HRV modality: ♥; Three different tasks: Task 2: 3 stages, Task 3: 4 stages, Task 4: 5 Stages (*Average time deviation from ground truth across all subjects ± standard error)

A.3. Hyperparameters Tuning And Results

Tree Approaches						
Algorithms	**Task**	**Number of trees**	**Number of Features For the Best Split**	**Criterion**	**Use Out-of-bag Samples**	
Random Forest	All tasks	100, 200, 300	20	Gini	No	
Best Hyperparameters						
Random Forest	All Tasks	300	20	Gini	No	
Shallow Machine Learning Approaches						
Algorithms	**Task**	**Alpha (Regularisation's multiplier)**	**Fit Intercept**	**Max Iteration (aka epochs)**	**Classifier Penalty**	**Sliding Window Length**
Linear SVM	All tasks	1.e-01, 1.e-02, 1.e-03, 1.e-04, 1.e-05, 1.e-06	True, False	5, 10 , 20	L1, L2	Actigraphy = 20 sleep epochs HR/HRV = 1 sleep epoch
Perceptron					L1, L2, Elastic net	
Logistic Regression					L1, L2, Elastic net	
Best Hyperparameters Used In Study						
Linear SVM	All tasks	1.e-3	True	5	L2	Actigraphy = 20 sleep epochs HR/HRV = 1 sleep epoch
Perceptron	All tasks	1.e-1	False	5	L2	Actigraphy = 20 sleep epochs HR/HRV = 1 sleep epoch
Logistic Regression	All tasks	1.e-4	True	20	L2	Actigraphy = 20 sleep epochs HR/HRV = 1 sleep epoch
Deep Neural Network Approaches						
Algorithms	**Task**	**Number of Layers**	**Number of Kernels (CNN) Hidden Units (LSTM)**	**Kernel Length**	**Optimiser**	**Window Length**
CNN (1D-Conv)	All tasks	1, 2, 3	32, 64, 128	3, 5, 7	RMSprop	21, 51, 101
LSTM (Many-to-one)	All tasks	1, 2, 3	64, 128, 256	N/A	RMSprop	21, 51, 101
Best Hyperparameters						
CNN (1D-Conv)	Task1	1	128	7	RMSprop	101
LSTM (Many-to-one)		3	128	N/A	RMSprop	51
CNN (1D-Conv)	Task2	3	64	5	RMSprop	101
LSTM (Many-to-one)		3	64	N/A	RMSprop	101
CNN (1D-Conv)	Task3	3	64	3	RMSprop	101
LSTM (Many-to-one)		3	128	N/A	RMSprop	101
CNN (1D-Conv)	Task4	3	64	5	RMSprop	101
LSTM (Many-to-one)		3	64	N/A	RMSprop	101
Hyperparameters Used In Study						
CNN (1D-Conv)	All tasks	1	64	2	RMSprop	21, 51, 101
LSTM (Many-to-one)		1	32	N/A	RMSprop	21, 51, 101

Table A.3 Hyper-parameters for ML and DL algorithms

A.4. Sleep Disorders within MESA

Total subjects	Healthy	Sleep apnea	Insomnia	Restless Legs Syndrome
1743	1469(84.3%)	132(7.6%)	109(6.2%)	78(4.5%)

Table A.4 Sleep Disorder Population details

A.5. Benchmark of Different Combinations of Modalities By Tasks

algorithms	**accuracy**	**specificity**	**precision**	**recall**	F_1	**Cohen's κ**	**Sleep**	**Wake**
CNN (101)	84.4 ± 1.0	67.9 ± 2.0	84.8 ± 1.3	92.4 ± 1.2	87.6 ± 1.1	62.0 ± 2.0	36.2 ± 7.3	-36.2 ± 7.3
LSTM (101)	84.4 ± 1.0	67.4 ± 1.9	84.7 ± 1.2	92.5 ± 1.1	87.8 ± 1.0	61.6 ± 2.1	36.0 ± 6.7	-36.0 ± 6.7
CNN (51)	84.3 ± 1.0	67.3 ± 2.0	84.5 ± 1.3	92.7 ± 1.2	87.6 ± 1.1	61.7 ± 2.1	39.0 ± 7.2	-39.0 ± 7.2
LSTM (51)	84.0 ± 1.0	72.7 ± 1.7	86.3 ± 1.2	89.2 ± 1.2	87.0 ± 1.1	62.0 ± 2.0	14.8 ± 6.7	-14.8 ± 6.7
CNN (21)	83.2 ± 1.0	63.8 ± 2.0	83.1 ± 1.3	93.0 ± 1.0	87.1 ± 1.0	58.8 ± 2.0	46.8 ± 7.1	-46.8 ± 7.1
LSTM (21)	83.2 ± 1.0	71.0 ± 1.8	85.4 ± 1.2	89.3 ± 1.1	86.6 ± 1.0	60.2 ± 1.9	18.8 ± 7.0	-18.8 ± 7.0
Random Forest	82.3 ± 1.0	65.7 ± 2.1	83.7 ± 1.3	90.6 ± 1.1	86.2 ± 1.0	57.1 ± 2.1	32.9 ± 7.3	-32.9 ± 7.3
Logistic Regression	82.1 ± 1.1	64.1 ± 2.1	83.3 ± 1.3	91.0 ± 1.2	86.0 ± 1.1	56.4 ± 2.2	37.3 ± 7.9	-37.3 ± 7.9
Linear SVM	81.8 ± 1.1	60.2 ± 2.1	82.1 ± 1.3	92.5 ± 1.1	86.2 ± 1.0	54.9 ± 2.2	49.0 ± 7.6	-49.0 ± 7.6
Perception	78.6 ± 1.1	65.2 ± 1.9	82.5 ± 1.3	85.1 ± 1.4	82.7 ± 1.2	49.8 ± 2.0	13.8 ± 8.2	-13.8 ± 8.2
Always sleep	66.5 ± 1.4	0.0 ± 0.0	66.5 ± 1.4	100.0 ± 0.0	79.1 ± 1.1	0.0 ± 0.0	187.4 ± 8.6	-187.4 ± 8.6
Always wake	33.5 ± 1.4	100.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	-365.1 ± 8.7	365.1 ± 8.7
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	365.1 ± 8.7	187.4 ± 8.6

Table A.5 Sleep wake classifiers performance for combined modality sensing using Actigraphy and HR/HRV

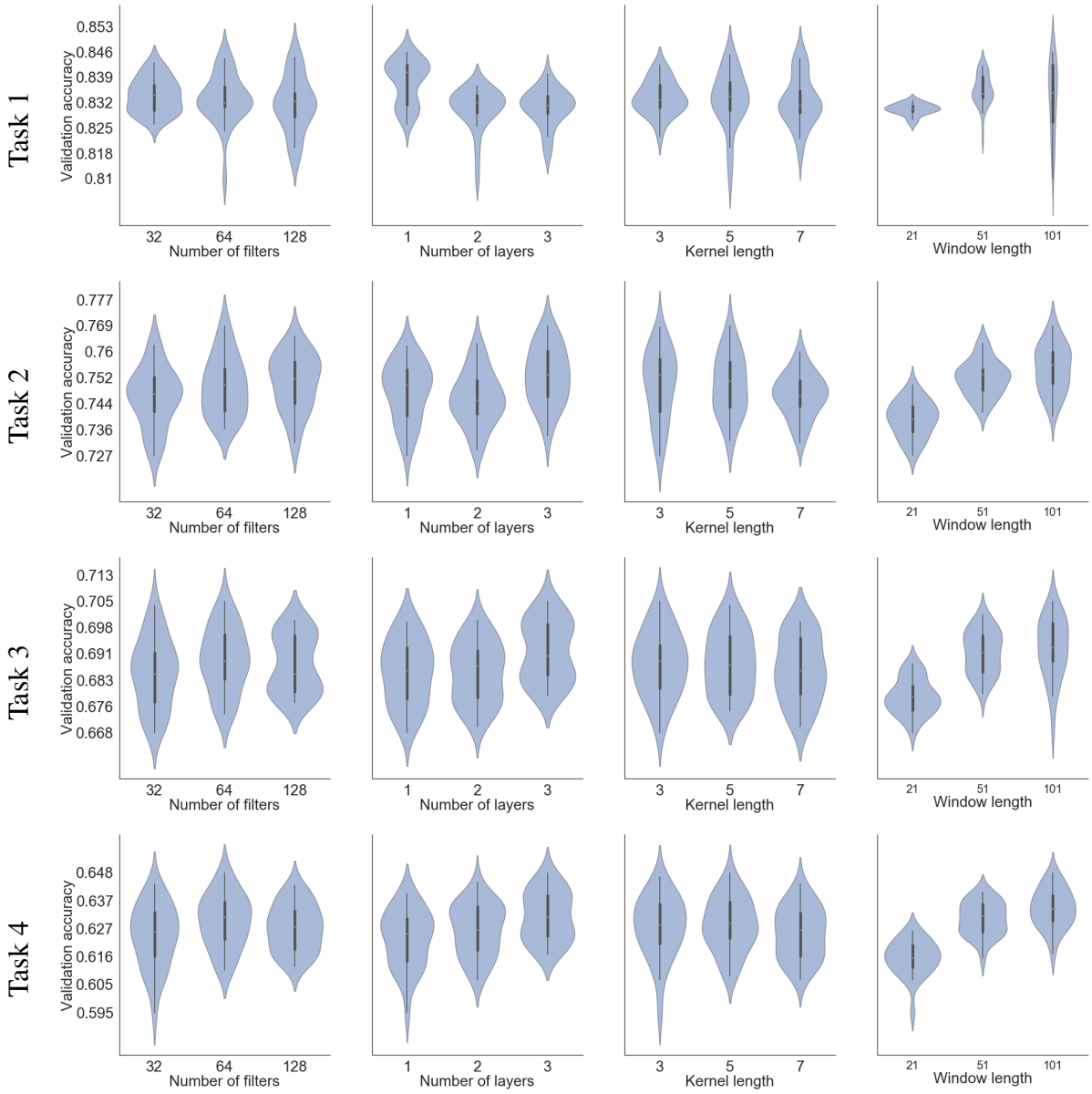


Figure A.1 CNN hyper-parameters tuning results

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	Sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	365.1 ± 8.7	187.4 ± 8.6
CNN (101)	84.9 ± 1.0	67.1 ± 2.0	84.7 ± 1.3	93.8 ± 1.0	88.3 ± 1.0	63.0 ± 2.0	43.0 ± 6.9	-43.0 ± 6.9
CNN (51)	84.4 ± 1.0	67.6 ± 2.0	84.6 ± 1.3	92.9 ± 1.1	87.8 ± 1.1	62.2 ± 2.1	39.0 ± 7.1	-39.0 ± 7.1
LSTM (101)	84.3 ± 1.0	69.7 ± 1.8	85.5 ± 1.2	91.2 ± 1.1	87.6 ± 1.0	62.0 ± 2.0	26.5 ± 6.6	-26.5 ± 6.6
LSTM (51)	83.9 ± 1.0	72.5 ± 1.7	86.2 ± 1.2	89.4 ± 1.2	87.0 ± 1.0	62.0 ± 2.0	15.6 ± 6.7	-15.6 ± 6.7
LSTM (21)	81.6 ± 1.0	63.7 ± 2.0	83.0 ± 1.2	90.2 ± 1.1	85.7 ± 1.0	55.1 ± 2.1	34.1 ± 7.2	-34.1 ± 7.2
Logistic Regression	81.6 ± 1.1	60.6 ± 2.1	82.2 ± 1.3	91.8 ± 1.2	85.9 ± 1.1	54.5 ± 2.3	45.3 ± 7.7	-45.3 ± 7.7
Linear SVM	81.5 ± 1.1	57.9 ± 2.1	81.4 ± 1.3	93.0 ± 1.0	86.1 ± 1.0	53.6 ± 2.2	54.7 ± 7.5	-54.7 ± 7.5
CNN (21)	81.4 ± 1.1	58.7 ± 2.1	81.5 ± 1.3	92.5 ± 1.1	85.9 ± 1.0	53.6 ± 2.2	51.8 ± 7.3	-51.8 ± 7.3
Random Forest	81.2 ± 1.0	63.4 ± 2.0	82.9 ± 1.3	89.7 ± 1.1	85.4 ± 1.0	54.1 ± 2.1	32.6 ± 7.2	-32.6 ± 7.2
Always sleep	66.5 ± 1.4	0.0 ± 0.0	66.5 ± 1.4	100.0 ± 0.0	79.1 ± 1.1	0.0 ± 0.0	187.4 ± 8.6	-187.4 ± 8.6
Perception	66.0 ± 1.0	73.5 ± 1.3	81.3 ± 1.3	62.2 ± 1.6	69.2 ± 1.3	30.7 ± 1.6	-85.9 ± 8.3	85.9 ± 8.3
Always wake	33.5 ± 1.4	100.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	-365.1 ± 8.7	365.1 ± 8.7

Table A.6 Task 1: Sleep wake classifiers performance for Actigraphy

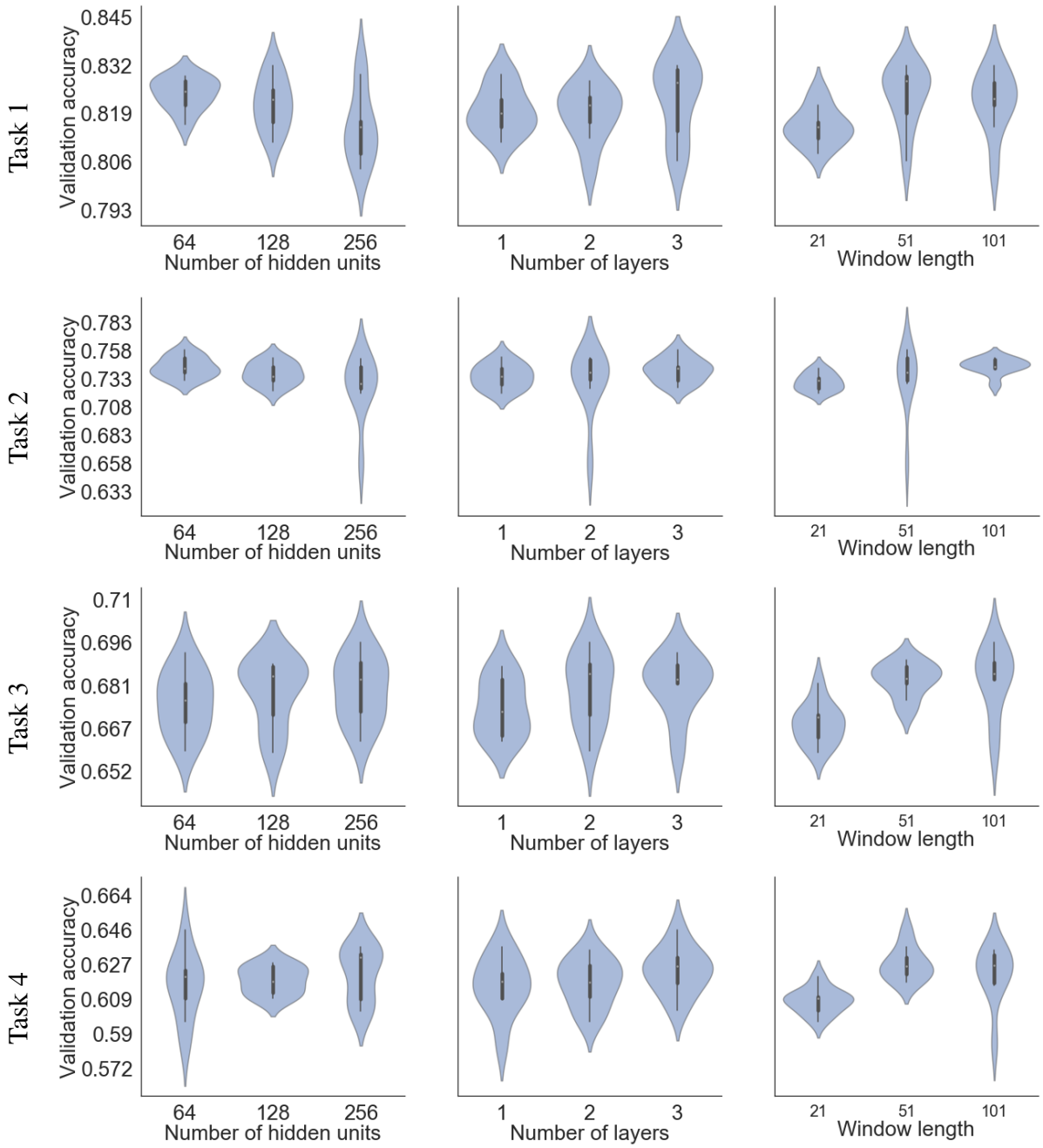


Figure A.2 LSTM hyper-parameters tuning results

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	Sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	365.1 ± 8.7	187.4 ± 8.6
LSTM (101)	79.5 ± 1.2	62.2 ± 2.1	81.8 ± 1.4	88.9 ± 1.3	84.1 ± 1.1	51.5 ± 2.2	35.3 ± 8.7	-35.3 ± 8.7
CNN (101)	79.1 ± 1.2	57.0 ± 2.1	79.9 ± 1.5	91.0 ± 1.4	83.9 ± 1.3	49.8 ± 2.1	54.4 ± 9.2	-54.4 ± 9.2
LSTM (51)	78.6 ± 1.2	61.1 ± 2.0	81.2 ± 1.4	88.2 ± 1.3	83.4 ± 1.2	49.5 ± 2.1	34.5 ± 8.9	-34.5 ± 8.9
CNN (51)	78.2 ± 1.2	54.0 ± 2.1	78.9 ± 1.5	91.3 ± 1.3	83.5 ± 1.2	47.1 ± 2.1	61.2 ± 9.2	-61.2 ± 9.2
LSTM (21)	77.7 ± 1.1	57.0 ± 1.8	79.5 ± 1.4	88.9 ± 1.2	83.0 ± 1.2	46.7 ± 1.9	45.2 ± 8.4	-45.2 ± 8.4
CNN (21)	75.7 ± 1.2	50.3 ± 2.2	77.4 ± 1.5	89.4 ± 1.4	81.7 ± 1.2	41.0 ± 2.0	61.0 ± 9.8	-61.0 ± 9.8
Logistic Regression	70.3 ± 1.3	34.7 ± 2.6	73.0 ± 1.5	89.1 ± 2.0	78.1 ± 1.5	25.5 ± 1.9	87.9 ± 12.9	-87.9 ± 12.9
Random Forest	70.3 ± 1.2	39.2 ± 2.3	73.6 ± 1.4	86.7 ± 1.9	77.6 ± 1.5	27.1 ± 1.7	70.4 ± 12.5	-70.4 ± 12.5
Linear SVM	70.1 ± 1.3	18.6 ± 1.9	70.1 ± 1.4	96.5 ± 1.0	80.1 ± 1.1	17.6 ± 1.5	142.8 ± 10.1	-142.8 ± 10.1
Always sleep	66.5 ± 1.4	0.0 ± 0.0	66.5 ± 1.4	100.0 ± 0.0	79.1 ± 1.1	0.0 ± 0.0	187.4 ± 8.6	-187.4 ± 8.6
Perception	49.5 ± 1.5	59.7 ± 2.1	65.8 ± 1.7	44.2 ± 2.8	49.1 ± 2.3	4.8 ± 1.6	-128.1 ± 15.3	128.1 ± 15.3
Always wake	33.5 ± 1.4	100.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	-365.1 ± 8.7	365.1 ± 8.7

Table A.7 Task 1: Sleep wake classifiers performance for single modality sensing using HR/HRV

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	Non-REM sleep	REM sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	297.5 ± 7.1	67.6 ± 3.0	187.4 ± 8.6
LSTM (51)	76.2 ± 1.0	85.6 ± 0.5	72.2 ± 1.3	68.8 ± 1.2	67.9 ± 1.3	58.4 ± 1.8	23.9 ± 7.1	-10.7 ± 3.8	-13.2 ± 6.8
LSTM (101)	76.1 ± 0.9	85.1 ± 0.5	71.9 ± 1.4	66.8 ± 1.2	66.4 ± 1.3	57.4 ± 1.9	26.5 ± 7.0	-23.3 ± 3.4	-3.2 ± 6.8
CNN (101)	76.0 ± 1.0	85.6 ± 0.6	72.2 ± 1.2	69.7 ± 1.3	68.1 ± 1.3	58.6 ± 1.9	30.2 ± 7.7	2.5 ± 4.5	-32.7 ± 7.2
CNN (51)	75.2 ± 1.1	85.7 ± 0.6	71.7 ± 1.3	70.2 ± 1.3	67.6 ± 1.3	58.4 ± 1.9	10.2 ± 8.4	10.8 ± 6.0	-21.0 ± 7.1
LSTM (21)	75.0 ± 0.9	84.9 ± 0.5	70.7 ± 1.2	67.2 ± 1.1	66.2 ± 1.2	55.6 ± 1.8	25.6 ± 7.2	-12.9 ± 4.1	-12.7 ± 6.9
CNN (21)	73.5 ± 1.0	83.9 ± 0.5	70.1 ± 1.2	66.0 ± 1.2	64.7 ± 1.2	54.4 ± 1.8	43.7 ± 8.3	-1.3 ± 5.4	-42.5 ± 7.0
Random Forest	70.5 ± 0.9	79.9 ± 0.5	59.2 ± 1.5	53.0 ± 0.7	50.3 ± 0.8	47.6 ± 1.7	83.5 ± 7.8	-63.6 ± 2.9	-20.0 ± 7.6
Logistic Regression	70.3 ± 1.0	79.5 ± 0.6	49.0 ± 0.8	52.2 ± 0.7	48.8 ± 0.8	46.9 ± 1.7	89.4 ± 8.6	-67.6 ± 3.0	-21.8 ± 8.3
Linear SVM	70.2 ± 1.0	79.1 ± 0.6	49.5 ± 0.8	51.5 ± 0.8	48.5 ± 0.8	46.7 ± 1.8	105.3 ± 8.1	-67.6 ± 3.0	-37.7 ± 7.8
Perception	65.8 ± 0.9	77.9 ± 0.5	49.8 ± 0.6	48.2 ± 0.7	46.6 ± 0.7	31.2 ± 1.6	85.2 ± 8.0	-36.8 ± 3.8	-48.4 ± 7.7
Always Non-REM sleep	54.2 ± 1.1	66.5 ± 0.2	18.2 ± 0.4	33.5 ± 0.2	23.3 ± 0.3	0.0 ± 0.0	255.0 ± 8.1	-67.6 ± 3.0	-187.4 ± 8.6
Always wake	33.5 ± 1.4	66.5 ± 0.2	11.3 ± 0.5	33.5 ± 0.2	16.4 ± 0.5	0.0 ± 0.0	-297.5 ± 7.1	-67.6 ± 3.0	365.1 ± 8.7
Always REM sleep	12.3 ± 0.5	66.7 ± 0.0	4.1 ± 0.2	33.0 ± 0.4	7.2 ± 0.3	0.0 ± 0.0	-297.5 ± 7.1	484.8 ± 8.5	-187.4 ± 8.6

Table A.8 Task 2: Non-REM, REM sleep and wake for combined sensing sensing using Actigraphy and HR/HRV.

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	Non-REM sleep	REM sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	297.5 ± 7.1	67.6 ± 3.0	187.4 ± 8.6
LSTM (101)	71.4 ± 0.9	80.1 ± 0.6	51.7 ± 1.1	52.9 ± 0.7	49.8 ± 0.8	49.7 ± 1.7	83.8 ± 7.7	-67.0 ± 3.0	-16.8 ± 7.3
CNN (101)	71.0 ± 1.0	79.5 ± 0.6	50.1 ± 0.8	52.1 ± 0.8	49.1 ± 0.8	48.0 ± 1.8	102.5 ± 7.9	-67.6 ± 3.0	-34.9 ± 7.5
LSTM (51)	70.9 ± 0.9	79.7 ± 0.6	49.0 ± 0.8	52.4 ± 0.7	49.2 ± 0.8	48.3 ± 1.7	87.9 ± 7.8	-67.6 ± 3.0	-20.3 ± 7.4
CNN (51)	70.6 ± 1.0	79.3 ± 0.6	49.6 ± 0.8	51.7 ± 0.8	48.7 ± 0.8	47.0 ± 1.8	103.1 ± 8.0	-67.6 ± 3.0	-35.5 ± 7.6
Logistic Regression	69.9 ± 1.0	79.0 ± 0.6	48.7 ± 0.8	51.3 ± 0.8	48.2 ± 0.8	46.2 ± 1.8	98.9 ± 8.0	-67.6 ± 3.0	-31.3 ± 7.7
LSTM (21)	69.7 ± 0.9	79.0 ± 0.6	48.2 ± 0.8	51.5 ± 0.7	48.3 ± 0.8	46.0 ± 1.7	91.1 ± 7.7	-67.6 ± 3.0	-23.5 ± 7.4
Linear SVM	69.6 ± 1.0	78.5 ± 0.6	49.5 ± 0.8	50.6 ± 0.8	47.7 ± 0.8	45.1 ± 1.8	116.8 ± 7.8	-67.6 ± 3.0	-49.2 ± 7.6
CNN (21)	69.5 ± 0.9	78.6 ± 0.6	48.9 ± 0.8	50.8 ± 0.7	47.8 ± 0.8	44.8 ± 1.8	107.8 ± 7.8	-67.6 ± 3.0	-40.2 ± 7.5
Random Forest	68.6 ± 0.9	78.8 ± 0.5	53.4 ± 1.1	51.0 ± 0.7	48.5 ± 0.8	44.3 ± 1.7	81.2 ± 7.4	-60.7 ± 2.9	-20.5 ± 7.2
Perception	56.7 ± 1.1	76.7 ± 0.5	54.8 ± 0.9	48.2 ± 0.9	47.2 ± 1.0	26.3 ± 1.7	6.1 ± 9.4	73.3 ± 7.6	-79.4 ± 7.1
Always wake	33.5 ± 1.4	66.5 ± 0.2	11.3 ± 0.5	33.5 ± 0.2	16.4 ± 0.5	0.0 ± 0.0	-297.5 ± 7.1	-67.6 ± 3.0	365.1 ± 8.7
Always REM sleep	12.3 ± 0.5	66.7 ± 0.0	4.1 ± 0.2	33.0 ± 0.4	7.2 ± 0.3	0.0 ± 0.0	-297.5 ± 7.1	484.8 ± 8.5	-187.4 ± 8.6

Table A.9 Task 2: Non-REM, REM sleep and wake for single modality sensing using Actigraphy

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	Non-REM sleep	REM sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	297.5 ± 7.1	67.6 ± 3.0	187.4 ± 8.6
LSTM (101)	73.8 ± 1.2	84.3 ± 0.6	69.8 ± 1.5	66.1 ± 1.3	64.9 ± 1.5	50.0 ± 2.2	36.4 ± 8.1	-8.6 ± 4.2	-27.8 ± 8.5
LSTM (51)	72.9 ± 1.1	83.8 ± 0.6	67.9 ± 1.4	64.1 ± 1.3	62.9 ± 1.4	45.5 ± 2.1	34.1 ± 8.3	-16.2 ± 4.3	-17.9 ± 8.5
CNN (101)	71.0 ± 1.2	83.6 ± 0.6	66.3 ± 1.4	65.4 ± 1.4	62.7 ± 1.4	46.1 ± 2.0	10.6 ± 9.1	2.3 ± 5.0	-12.9 ± 9.4
LSTM (21)	70.4 ± 1.0	82.3 ± 0.6	65.3 ± 1.4	60.8 ± 1.2	59.6 ± 1.3	39.3 ± 1.9	36.9 ± 8.0	-20.4 ± 4.1	-16.4 ± 8.6
CNN (51)	70.3 ± 1.1	82.2 ± 0.6	66.6 ± 1.4	62.2 ± 1.3	60.1 ± 1.4	42.9 ± 2.0	46.0 ± 9.5	-7.9 ± 5.8	-38.2 ± 9.2
CNN (21)	67.9 ± 1.1	80.8 ± 0.6	63.5 ± 1.4	59.0 ± 1.2	57.0 ± 1.3	37.3 ± 1.9	41.9 ± 9.9	-14.4 ± 5.3	-27.5 ± 9.9
Logistic Regression	59.8 ± 1.1	73.1 ± 0.5	44.3 ± 0.9	42.7 ± 0.7	37.7 ± 0.9	19.4 ± 1.6	123.2 ± 13.9	-67.6 ± 3.0	-55.5 ± 14.2
Random Forest	59.6 ± 1.0	73.8 ± 0.4	48.4 ± 1.4	43.4 ± 0.6	39.2 ± 0.8	19.7 ± 1.4	92.0 ± 13.2	-65.3 ± 3.0	-26.8 ± 13.5
Linear SVM	59.4 ± 1.1	71.8 ± 0.5	45.6 ± 0.8	41.1 ± 0.6	35.8 ± 0.8	18.0 ± 1.5	166.7 ± 12.3	-67.6 ± 3.0	-99.0 ± 12.6
Always Non-REM sleep	54.2 ± 1.1	66.5 ± 0.2	18.2 ± 0.4	33.5 ± 0.2	23.3 ± 0.3	0.0 ± 0.0	255.0 ± 8.1	-67.6 ± 3.0	-187.4 ± 8.6
Perception	39.6 ± 1.5	69.6 ± 0.4	40.9 ± 0.8	36.0 ± 0.6	30.9 ± 0.8	2.5 ± 1.1	-36.9 ± 16.7	137.3 ± 16.1	-100.4 ± 9.6
Always wake	33.5 ± 1.4	66.5 ± 0.2	11.3 ± 0.5	33.5 ± 0.2	16.4 ± 0.5	0.0 ± 0.0	-297.5 ± 7.1	-67.6 ± 3.0	365.1 ± 8.7
Always REM sleep	12.3 ± 0.5	66.7 ± 0.0	4.1 ± 0.2	33.0 ± 0.4	7.2 ± 0.3	0.0 ± 0.0	-297.5 ± 7.1	484.8 ± 8.5	-187.4 ± 8.6

Table A.10 Task 2: Non-REM, REM sleep and wake for single modality sensing of HR/HRV

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	Deep sleep	Light sleep	REM sleep	Wake
ground truth	100.0 +- 0.0	100.0 +- 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	39.3 ± 3.6	258.2 ± 7.0	67.6 ± 3.0	187.4 ± 8.6
LSTM (51)	70.3 ± 1.0	87.4 ± 0.4	57.9 ± 1.3	54.0 ± 1.0	51.9 ± 1.0	53.8 ± 1.9	-36.2 ± 3.5	42.8 ± 7.4	-5.6 ± 4.0	-1.0 ± 6.9
LSTM (101)	70.2 ± 1.0	86.9 ± 0.4	59.9 ± 1.5	52.4 ± 1.0	51.3 ± 1.1	51.7 ± 1.8	-32.4 ± 3.5	76.0 ± 7.3	-24.7 ± 3.7	-18.9 ± 6.6
CNN (101)	69.0 ± 1.0	87.0 ± 0.4	58.0 ± 1.4	53.7 ± 1.0	51.2 ± 1.1	51.6 ± 1.8	-34.5 ± 3.5	46.1 ± 8.1	4.4 ± 4.8	-15.9 ± 7.5
LSTM (21)	68.3 ± 1.0	86.5 ± 0.4	55.0 ± 1.2	51.6 ± 0.9	49.5 ± 0.9	50.0 ± 1.7	-37.8 ± 3.5	55.7 ± 7.7	-6.6 ± 4.1	-11.2 ± 7.0
CNN (51)	68.0 ± 1.1	86.9 ± 0.4	54.8 ± 1.3	53.5 ± 1.0	49.9 ± 1.0	51.6 ± 1.9	-37.8 ± 3.6	32.5 ± 8.8	18.5 ± 6.2	-13.2 ± 7.1
CNN (21)	67.0 ± 1.0	85.9 ± 0.4	52.2 ± 1.1	50.5 ± 0.9	47.8 ± 0.9	48.3 ± 1.7	-38.9 ± 3.6	69.9 ± 8.5	1.1 ± 5.4	-32.1 ± 6.9
Random Forest	63.6 ± 1.0	83.3 ± 0.4	44.7 ± 1.3	40.1 ± 0.6	36.7 ± 0.6	34.4 ± 1.3	-38.9 ± 3.6	115.3 ± 8.3	-61.3 ± 2.9	-15.2 ± 7.6
Logistic Regression	63.5 ± 1.0	82.8 ± 0.4	35.6 ± 0.8	39.5 ± 0.7	35.5 ± 0.7	32.9 ± 1.3	-39.2 ± 3.6	132.6 ± 8.8	-67.6 ± 3.0	-25.9 ± 8.1
Linear SVM	63.4 ± 1.1	82.7 ± 0.4	35.2 ± 0.8	39.4 ± 0.7	35.4 ± 0.7	32.9 ± 1.3	-39.3 ± 3.6	137.9 ± 8.6	-67.6 ± 3.0	-30.9 ± 7.9
Perception	50.6 ± 1.1	80.7 ± 0.3	36.8 ± 0.5	31.3 ± 0.5	30.0 ± 0.6	0.6 ± 1.4	35.6 ± 6.9	96.5 ± 9.6	-28.6 ± 4.3	-103.5 ± 7.7
Always light sleep	47.0 ± 1.1	74.3 ± 0.3	12.1 ± 0.3	25.7 ± 0.3	16.2 ± 0.4	0.0 ± 0.0	-39.3 ± 3.6	294.3 ± 8.2	-67.6 ± 3.0	-187.4 ± 8.6
Always wake	33.5 ± 1.4	74.3 ± 0.3	8.7 ± 0.4	25.7 ± 0.3	12.6 ± 0.5	0.0 ± 0.0	-39.3 ± 3.6	-258.2 ± 7.0	-67.6 ± 3.0	365.1 ± 8.7
Always REM sleep	12.3 ± 0.5	74.5 ± 0.2	3.1 ± 0.1	25.1 ± 0.3	5.4 ± 0.2	0.0 ± 0.0	-39.3 ± 3.6	-258.2 ± 7.0	484.8 ± 8.5	-187.4 ± 8.6
Always deep sleep	7.2 ± 0.7	74.9 ± 0.1	1.8 ± 0.2	23.6 ± 0.6	3.2 ± 0.3	0.0 ± 0.0	513.2 ± 9.4	-258.2 ± 7.0	-67.6 ± 3.0	-187.4 ± 8.6

Table A.11 Task 3: Wake, light sleep, deep sleep and REM-sleep classifiers performance for combined modality sensing sensing using Actigraphy and HR/HRV

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	Deep sleep	Light sleep	REM sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	39.3 ± 3.6	258.2 ± 7.0	67.6 ± 3.0	187.4 ± 8.6
LSTM (101)	67.4 ± 1.2	86.2 ± 0.4	56.2 ± 1.6	51.3 ± 1.1	49.5 ± 1.2	44.6 ± 2.2	-33.7 ± 3.5	60.4 ± 8.1	-13.5 ± 3.8	-13.1 ± 8.4
LSTM (51)	66.2 ± 1.1	85.6 ± 0.4	54.4 ± 1.5	49.5 ± 1.1	47.4 ± 1.1	41.2 ± 2.1	-36.4 ± 3.5	65.9 ± 7.9	-14.6 ± 4.1	-15.0 ± 8.1
CNN (101)	64.3 ± 1.1	85.3 ± 0.4	54.4 ± 1.6	50.2 ± 1.1	47.1 ± 1.1	40.9 ± 2.1	-34.8 ± 3.5	49.6 ± 8.9	8.2 ± 5.1	-23.0 ± 9.5
LSTM (21)	63.9 ± 1.0	84.7 ± 0.4	50.3 ± 1.3	47.1 ± 0.9	44.7 ± 1.0	36.3 ± 1.9	-38.6 ± 3.6	65.6 ± 8.3	-13.9 ± 4.3	-13.2 ± 8.4
CNN (51)	63.4 ± 1.2	84.6 ± 0.4	51.6 ± 1.4	48.1 ± 1.0	44.9 ± 1.1	38.4 ± 2.1	-37.5 ± 3.5	70.8 ± 9.5	3.1 ± 6.0	-36.3 ± 9.4
CNN (21)	61.3 ± 1.1	83.5 ± 0.4	46.8 ± 1.1	44.9 ± 0.9	41.8 ± 1.0	33.3 ± 1.8	-39.1 ± 3.6	75.7 ± 10.4	-10.9 ± 5.5	-25.7 ± 10.2
Linear SVM	53.4 ± 1.2	78.6 ± 0.5	31.8 ± 0.8	32.6 ± 0.6	27.1 ± 0.8	11.5 ± 1.3	-39.3 ± 3.6	150.1 ± 15.3	-67.6 ± 3.0	-43.2 ± 15.7
Logistic Regression	53.4 ± 1.2	78.9 ± 0.5	31.2 ± 0.8	33.0 ± 0.6	27.7 ± 0.7	12.3 ± 1.3	-39.3 ± 3.6	133.2 ± 15.2	-67.6 ± 3.0	-26.3 ± 15.6
Random Forest	53.3 ± 1.0	79.2 ± 0.4	35.5 ± 1.1	33.2 ± 0.5	28.6 ± 0.6	12.6 ± 1.1	-39.0 ± 3.6	108.0 ± 13.9	-64.7 ± 3.0	-4.3 ± 14.0
Always light sleep	47.0 ± 1.1	74.3 ± 0.3	12.1 ± 0.3	25.7 ± 0.3	16.2 ± 0.4	0.0 ± 0.0	-39.3 ± 3.6	294.3 ± 8.2	-67.6 ± 3.0	-187.4 ± 8.6
Always wake	33.5 ± 1.4	74.3 ± 0.3	8.7 ± 0.4	25.7 ± 0.3	12.6 ± 0.5	0.0 ± 0.0	-39.3 ± 3.6	-258.2 ± 7.0	-67.6 ± 3.0	365.1 ± 8.7
Perception	24.0 ± 1.3	76.6 ± 0.2	31.2 ± 1.0	27.3 ± 0.7	17.9 ± 0.8	-0.7 ± 1.1	105.1 ± 16.6	-141.8 ± 15.7	172.9 ± 18.4	-136.2 ± 9.0
Always REM sleep	12.3 ± 0.5	74.5 ± 0.2	3.1 ± 0.1	25.1 ± 0.3	5.4 ± 0.2	0.0 ± 0.0	-39.3 ± 3.6	-258.2 ± 7.0	484.8 ± 8.5	-187.4 ± 8.6
Always deep sleep	7.2 ± 0.7	74.9 ± 0.1	1.8 ± 0.2	23.6 ± 0.6	3.2 ± 0.3	0.0 ± 0.0	513.2 ± 9.4	-258.2 ± 7.0	-67.6 ± 3.0	-187.4 ± 8.6

Table A.12 Task 3: Wake, light sleep, deep sleep and REM-sleep classifiers performance for single modality sensing of HR/HRV

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	Deep sleep	Light sleep	REM sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	39.3 ± 3.6	258.2 ± 7.0	67.6 ± 3.0	187.4 ± 8.6
LSTM (101)	64.1 ± 1.0	82.9 ± 0.5	35.6 ± 0.7	39.6 ± 0.7	35.8 ± 0.7	33.5 ± 1.4	-39.3 ± 3.6	139.4 ± 8.5	-67.6 ± 3.0	-32.5 ± 7.4
CNN (101)	63.9 ± 1.0	83.0 ± 0.4	36.3 ± 0.9	39.6 ± 0.7	35.7 ± 0.7	33.5 ± 1.4	-39.3 ± 3.6	133.2 ± 8.7	-67.5 ± 3.0	-26.4 ± 7.6
LSTM (51)	63.6 ± 1.0	82.7 ± 0.4	35.6 ± 0.8	39.3 ± 0.7	35.5 ± 0.7	33.0 ± 1.4	-39.3 ± 3.6	143.0 ± 8.2	-67.3 ± 3.0	-36.3 ± 7.1
CNN (51)	63.4 ± 1.0	82.8 ± 0.4	35.2 ± 0.8	39.3 ± 0.6	35.3 ± 0.7	33.1 ± 1.4	-39.3 ± 3.6	126.3 ± 8.8	-67.5 ± 3.0	-19.4 ± 7.8
Logistic Regression	62.9 ± 1.1	82.5 ± 0.4	35.2 ± 0.8	38.9 ± 0.7	35.0 ± 0.8	32.2 ± 1.4	-39.2 ± 3.6	140.9 ± 8.6	-67.6 ± 3.0	-34.1 ± 7.7
Linear SVM	62.8 ± 1.1	82.3 ± 0.4	35.3 ± 0.8	38.7 ± 0.7	34.9 ± 0.8	32.0 ± 1.4	-39.3 ± 3.6	148.2 ± 8.5	-67.6 ± 3.0	-41.2 ± 7.7
LSTM (21)	62.7 ± 1.0	82.3 ± 0.4	35.4 ± 0.8	38.6 ± 0.7	34.8 ± 0.7	31.3 ± 1.3	-39.3 ± 3.6	152.2 ± 8.0	-67.5 ± 3.0	-45.4 ± 7.0
CNN (21)	62.6 ± 1.0	82.3 ± 0.4	35.1 ± 0.8	38.6 ± 0.7	34.8 ± 0.7	31.3 ± 1.3	-39.3 ± 3.6	148.1 ± 8.4	-67.6 ± 3.0	-41.2 ± 7.4
Random Forest	61.4 ± 1.0	82.6 ± 0.4	39.6 ± 0.9	38.1 ± 0.5	35.0 ± 0.6	31.2 ± 1.3	-37.1 ± 3.6	112.1 ± 8.1	-59.3 ± 2.9	-15.6 ± 7.3
Always light sleep	47.0 ± 1.1	74.3 ± 0.3	12.1 ± 0.3	25.7 ± 0.3	16.2 ± 0.4	0.0 ± 0.0	-39.3 ± 3.6	294.3 ± 8.2	-67.6 ± 3.0	-187.4 ± 8.6
Perception	41.3 ± 0.9	79.7 ± 0.3	37.3 ± 0.6	31.1 ± 0.6	31.0 ± 0.6	11.1 ± 1.1	57.6 ± 5.2	-32.5 ± 8.6	57.2 ± 4.5	-82.3 ± 7.3
Always wake	33.5 ± 1.4	74.3 ± 0.3	8.7 ± 0.4	25.7 ± 0.3	12.6 ± 0.5	0.0 ± 0.0	-39.3 ± 3.6	-258.2 ± 7.0	-67.6 ± 3.0	365.1 ± 8.7
Always REM sleep	12.3 ± 0.5	74.5 ± 0.2	3.1 ± 0.1	25.1 ± 0.3	5.4 ± 0.2	0.0 ± 0.0	-39.3 ± 3.6	-258.2 ± 7.0	484.8 ± 8.5	-187.4 ± 8.6
Always deep sleep	7.2 ± 0.7	74.9 ± 0.1	1.8 ± 0.2	23.6 ± 0.6	3.2 ± 0.3	0.0 ± 0.0	513.2 ± 9.4	-258.2 ± 7.0	-67.6 ± 3.0	-187.4 ± 8.6

Table A.13 Task 3: Wake, light sleep, deep sleep and REM-sleep classifiers performance for single modality sensing of Actigraphy

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	N1 sleep	N2 sleep	N3 sleep	REM sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	49.4 ± 3.2	208.7 ± 6.2	39.3 ± 3.6	67.6 ± 3.0	187.4 ± 8.6
LSTM (51)	63.7 ± 1.0	88.7 ± 0.3	47.1 ± 1.4	43.0 ± 0.8	39.9 ± 0.8	56.3 ± 1.8	-46.0 ± 3.0	71.9 ± 7.5	-35.2 ± 3.5	-12.9 ± 3.9	22.2 ± 7.1
LSTM (101)	63.6 ± 1.0	88.7 ± 0.3	47.8 ± 1.3	43.3 ± 0.8	40.5 ± 0.9	57.0 ± 1.8	-46.2 ± 3.0	97.7 ± 7.5	-32.3 ± 3.5	-15.9 ± 3.9	-3.3 ± 6.8
CNN (101)	63.1 ± 1.1	88.8 ± 0.3	51.5 ± 1.4	44.7 ± 0.9	41.9 ± 0.9	56.2 ± 1.8	-40.2 ± 3.1	92.4 ± 8.0	-34.3 ± 3.5	8.2 ± 5.0	-26.2 ± 7.1
CNN (51)	62.9 ± 1.1	88.8 ± 0.3	48.7 ± 1.3	44.2 ± 0.8	40.6 ± 0.9	56.0 ± 1.9	-42.0 ± 3.1	86.6 ± 8.5	-37.9 ± 3.6	14.8 ± 5.9	-21.5 ± 7.1
LSTM (21)	62.6 ± 1.0	88.5 ± 0.3	43.6 ± 1.1	43.1 ± 0.7	39.2 ± 0.7	54.2 ± 1.7	-45.6 ± 3.0	68.3 ± 7.3	-37.8 ± 3.5	3.8 ± 4.4	11.3 ± 7.0
CNN (21)	61.3 ± 1.1	88.3 ± 0.3	44.7 ± 1.1	42.8 ± 0.7	38.8 ± 0.8	53.0 ± 1.8	-43.3 ± 3.1	80.8 ± 8.4	-39.1 ± 3.6	18.5 ± 6.0	-16.9 ± 7.2
Random Forest	56.9 ± 1.0	86.2 ± 0.3	36.4 ± 1.2	33.1 ± 0.5	28.8 ± 0.5	46.3 ± 1.6	-48.6 ± 3.2	123.6 ± 8.4	-38.7 ± 3.6	-54.9 ± 3.1	18.6 ± 8.1
Logistic Regression	56.7 ± 1.1	85.8 ± 0.4	29.2 ± 0.9	32.4 ± 0.5	27.1 ± 0.6	45.6 ± 1.7	-47.7 ± 3.1	135.8 ± 9.3	-39.2 ± 3.6	-67.2 ± 3.0	18.3 ± 8.9
Linear SVM	56.1 ± 1.1	85.9 ± 0.3	31.4 ± 0.9	32.1 ± 0.5	27.3 ± 0.5	44.2 ± 1.7	-44.8 ± 3.1	113.8 ± 9.6	-38.7 ± 3.6	-63.3 ± 3.0	33.0 ± 9.1
Always N2 sleep	38.0 ± 1.0	79.4 ± 0.4	7.9 ± 0.3	20.6 ± 0.4	11.2 ± 0.3	0.0 ± 0.0	-49.4 ± 3.2	343.7 ± 8.3	-39.3 ± 3.6	-67.6 ± 3.0	-187.4 ± 8.6
Perception	37.6 ± 0.9	83.3 ± 0.2	32.0 ± 0.5	25.7 ± 0.5	23.5 ± 0.5	6.8 ± 1.4	-6.2 ± 4.1	79.3 ± 9.6	33.5 ± 6.5	26.8 ± 6.1	-133.4 ± 7.6
Always wake	33.5 ± 1.4	79.4 ± 0.4	7.0 ± 0.4	20.6 ± 0.4	10.1 ± 0.4	0.0 ± 0.0	-49.4 ± 3.2	-208.7 ± 6.2	-39.3 ± 3.6	-67.6 ± 3.0	365.1 ± 8.7
Always REM sleep	12.3 ± 0.5	79.6 ± 0.2	2.5 ± 0.1	20.0 ± 0.3	4.3 ± 0.2	0.0 ± 0.0	-49.4 ± 3.2	-208.7 ± 6.2	-39.3 ± 3.6	484.8 ± 8.5	-187.4 ± 8.6
Always N1 sleep	9.0 ± 0.6	79.6 ± 0.2	1.8 ± 0.1	20.0 ± 0.3	3.3 ± 0.2	0.0 ± 0.0	503.0 ± 8.8	-208.7 ± 6.2	-39.3 ± 3.6	-67.6 ± 3.0	-187.4 ± 8.6
Always N3 sleep	7.2 ± 0.7	79.8 ± 0.1	1.4 ± 0.1	18.9 ± 0.5	2.6 ± 0.2	0.0 ± 0.0	-49.4 ± 3.2	-208.7 ± 6.2	513.2 ± 9.4	-67.6 ± 3.0	-187.4 ± 8.6

Table A.14 Task 4: Wake, N1 sleep, N2 sleep, N3 and REM-sleep classifiers performance for combined modality sensing sensing using Actigraphy and HR/HRV

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	N1 sleep	N2 sleep	N3 sleep	REM sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	49.4 ± 3.2	208.7 ± 6.2	39.3 ± 3.6	67.6 ± 3.0	187.4 ± 8.6
LSTM (51)	56.9 ± 1.0	85.7 ± 0.4	26.1 ± 0.8	32.2 ± 0.6	27.1 ± 0.7	46.9 ± 1.7	-49.4 ± 3.2	169.2 ± 8.5	-39.3 ± 3.6	-67.6 ± 3.0	-12.9 ± 7.5
LSTM (101)	56.9 ± 1.0	85.7 ± 0.4	25.3 ± 0.7	32.3 ± 0.6	27.1 ± 0.7	47.1 ± 1.7	-49.4 ± 3.2	159.7 ± 8.7	-39.3 ± 3.6	-67.6 ± 3.0	-3.3 ± 7.5
CNN (101)	56.8 ± 1.1	85.8 ± 0.3	27.7 ± 0.9	32.2 ± 0.5	27.2 ± 0.6	46.9 ± 1.7	-49.4 ± 3.2	144.7 ± 9.1	-39.3 ± 3.6	-65.6 ± 3.0	9.6 ± 8.3
CNN (51)	56.5 ± 1.1	85.7 ± 0.3	25.4 ± 0.7	32.0 ± 0.5	26.7 ± 0.6	46.0 ± 1.7	-49.4 ± 3.2	138.2 ± 9.2	-39.3 ± 3.6	-67.3 ± 3.0	17.8 ± 8.5
Logistic Regression	56.1 ± 1.1	85.5 ± 0.4	25.2 ± 0.7	31.8 ± 0.6	26.7 ± 0.7	45.9 ± 1.7	-49.4 ± 3.2	163.3 ± 8.8	-39.2 ± 3.6	-67.6 ± 3.0	-7.1 ± 8.0
LSTM (21)	56.0 ± 1.0	85.5 ± 0.4	25.2 ± 0.7	31.7 ± 0.6	26.6 ± 0.7	45.1 ± 1.7	-49.4 ± 3.2	170.0 ± 8.4	-39.2 ± 3.6	-67.6 ± 3.0	-13.7 ± 7.4
CNN (21)	55.9 ± 1.0	85.5 ± 0.3	25.5 ± 0.7	31.5 ± 0.5	26.5 ± 0.6	44.8 ± 1.7	-49.4 ± 3.2	170.4 ± 8.6	-39.3 ± 3.6	-67.4 ± 3.0	-14.4 ± 7.8
Linear SVM	54.8 ± 1.1	85.6 ± 0.3	31.3 ± 0.7	31.1 ± 0.5	26.8 ± 0.5	43.8 ± 1.7	-44.5 ± 3.2	114.3 ± 9.2	-37.7 ± 3.6	-60.3 ± 3.0	28.2 ± 8.8
Random Forest	54.4 ± 1.0	85.6 ± 0.3	31.7 ± 0.8	31.1 ± 0.4	27.2 ± 0.5	42.7 ± 1.6	-44.3 ± 3.3	120.8 ± 8.0	-36.7 ± 3.5	-56.0 ± 2.8	16.2 ± 7.6
Always N2 sleep	38.0 ± 1.0	79.4 ± 0.4	7.9 ± 0.3	20.6 ± 0.4	11.2 ± 0.3	0.0 ± 0.0	-49.4 ± 3.2	343.7 ± 8.3	-39.3 ± 3.6	-67.6 ± 3.0	-187.4 ± 8.6
Always wake	33.5 ± 1.4	79.4 ± 0.4	7.0 ± 0.4	20.6 ± 0.4	10.1 ± 0.4	0.0 ± 0.0	-49.4 ± 3.2	-208.7 ± 6.2	-39.3 ± 3.6	-67.6 ± 3.0	365.1 ± 8.7
Perception	32.9 ± 0.9	82.9 ± 0.2	31.0 ± 0.5	29.1 ± 0.7	23.2 ± 0.6	21.8 ± 1.4	-13.3 ± 3.5	5.9 ± 8.1	151.6 ± 8.3	-26.4 ± 4.5	-117.8 ± 7.6
Always REM sleep	12.3 ± 0.5	79.6 ± 0.2	2.5 ± 0.1	20.0 ± 0.3	4.3 ± 0.2	0.0 ± 0.0	-49.4 ± 3.2	-208.7 ± 6.2	-39.3 ± 3.6	484.8 ± 8.5	-187.4 ± 8.6
Always N1 sleep	9.0 ± 0.6	79.6 ± 0.2	1.8 ± 0.1	20.0 ± 0.3	3.3 ± 0.2	0.0 ± 0.0	503.0 ± 8.8	-208.7 ± 6.2	-39.3 ± 3.6	-67.6 ± 3.0	-187.4 ± 8.6
Always N3 sleep	7.2 ± 0.7	79.8 ± 0.1	1.4 ± 0.1	18.9 ± 0.5	2.6 ± 0.2	0.0 ± 0.0	-49.4 ± 3.2	-208.7 ± 6.2	513.2 ± 9.4	-67.6 ± 3.0	-187.4 ± 8.6

Table A.15 Task 4: Wake, N1 sleep, N2 sleep, N3 and REM-sleep classifiers performance for single modality sensing using Actigraphy

algorithms	accuracy	specificity	precision	recall	F_1	Cohen's κ	N1 sleep	N2 sleep	N3 sleep	REM sleep	Wake
ground truth	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	49.4 ± 3.2	208.7 ± 6.2	39.3 ± 3.6	67.6 ± 3.0	187.4 ± 8.6
CNN (21)	55.6 ± 1.1	86.4 ± 0.3	40.4 ± 1.2	37.3 ± 0.8	33.6 ± 0.9	36.2 ± 1.8	-44.4 ± 3.0	103.0 ± 9.8	-39.1 ± 3.6	-3.9 ± 5.8	-15.6 ± 10.1
CNN (101)	55.6 ± 1.1	86.7 ± 0.3	44.9 ± 1.4	38.9 ± 0.9	35.9 ± 1.0	37.1 ± 1.8	-40.8 ± 3.1	81.2 ± 9.4	-29.5 ± 3.6	-12.0 ± 4.8	1.1 ± 10.7
CNN (51)	54.2 ± 1.1	86.0 ± 0.3	41.2 ± 1.3	35.6 ± 0.8	32.1 ± 1.0	32.3 ± 1.9	-40.6 ± 3.1	69.5 ± 11.0	-36.4 ± 3.5	-28.2 ± 5.1	35.7 ± 12.1
LSTM (101)	50.6 ± 1.0	83.9 ± 0.3	25.7 ± 0.8	28.9 ± 0.5	23.8 ± 0.6	29.6 ± 1.7	-49.4 ± 3.2	189.7 ± 10.4	-39.3 ± 3.6	-66.0 ± 2.9	-34.9 ± 10.3
LSTM (51)	50.4 ± 1.1	84.1 ± 0.4	23.8 ± 0.9	29.2 ± 0.5	23.4 ± 0.6	27.0 ± 1.7	-49.3 ± 3.2	107.0 ± 13.3	-39.3 ± 3.6	-67.0 ± 3.0	48.6 ± 13.4
LSTM (21)	50.1 ± 1.0	83.8 ± 0.3	30.7 ± 1.0	29.2 ± 0.5	24.6 ± 0.6	28.0 ± 1.6	-48.3 ± 3.1	199.2 ± 10.7	-39.3 ± 3.6	-58.2 ± 3.2	-53.4 ± 10.7
Logistic Regression	46.7 ± 1.1	82.8 ± 0.4	22.6 ± 0.8	26.8 ± 0.5	21.2 ± 0.6	18.4 ± 1.7	-48.5 ± 3.2	128.0 ± 15.8	-39.3 ± 3.6	-67.6 ± 3.0	27.4 ± 16.2
Random Forest	46.6 ± 1.0	83.1 ± 0.3	29.9 ± 1.0	27.1 ± 0.4	22.3 ± 0.5	17.6 ± 1.4	-46.2 ± 3.2	97.8 ± 13.7	-38.9 ± 3.6	-61.3 ± 3.1	48.5 ± 14.3
Linear SVM	43.2 ± 1.3	82.4 ± 0.4	24.4 ± 1.0	25.6 ± 0.6	19.9 ± 0.7	13.5 ± 1.6	-45.9 ± 3.1	-1.7 ± 16.1	-30.8 ± 4.1	-62.1 ± 3.2	140.5 ± 17.5
Always N2 sleep	38.0 ± 1.0	79.4 ± 0.4	7.9 ± 0.3	20.6 ± 0.4	11.2 ± 0.3	0.0 ± 0.0	-49.4 ± 3.2	343.7 ± 8.3	-39.3 ± 3.6	-67.6 ± 3.0	-187.4 ± 8.6
Always wake	33.5 ± 1.4	79.4 ± 0.4	7.0 ± 0.4	20.6 ± 0.4	10.1 ± 0.4	0.0 ± 0.0	-49.4 ± 3.2	-208.7 ± 6.2	-39.3 ± 3.6	-67.6 ± 3.0	365.1 ± 8.7
Perception	19.2 ± 1.1	81.0 ± 0.1	27.6 ± 0.7	23.6 ± 0.6	13.2 ± 0.5	7.6 ± 1.2	60.7 ± 13.2	-87.5 ± 16.0	188.3 ± 15.3	13.3 ± 7.5	-174.8 ± 8.5
Always REM sleep	12.3 ± 0.5	79.6 ± 0.2	2.5 ± 0.1	20.0 ± 0.3	4.3 ± 0.2	0.0 ± 0.0	-49.4 ± 3.2	-208.7 ± 6.2	-39.3 ± 3.6	484.8 ± 8.5	-187.4 ± 8.6
Always N1 sleep	9.0 ± 0.6	79.6 ± 0.2	1.8 ± 0.1	20.0 ± 0.3	3.3 ± 0.2	0.0 ± 0.0	503.0 ± 8.8	-208.7 ± 6.2	-39.3 ± 3.6	-67.6 ± 3.0	-187.4 ± 8.6
Always N3 sleep	7.2 ± 0.7	79.8 ± 0.1	1.4 ± 0.1	18.9 ± 0.5	2.6 ± 0.2	0.0 ± 0.0	-49.4 ± 3.2	-208.7 ± 6.2	513.2 ± 9.4	-67.6 ± 3.0	-187.4 ± 8.6

Table A.16 Task 4: Wake, N1 sleep, N2 sleep, N3 and REM-sleep classifiers performance for single modality sensing of HR/HRV

Appendix B. UbiSleepNet: Appendix

B.1. HYPERPARAMETERS TUNING AND RESULTS FOR UBISLEEPNET

The hyperparameter tuning was performed based on the designed backbone network from three to five convolutional blocks (7-13 convolutional layers). The first two blocks consisted of two convolutional layers. The third, fourth and fifth convolutional blocks consisted of three convolutional layers. The hyperparameter search aimed to reduce the search space and maintain suitable temporal lengths of the latent features. The hyperparameter tuning only focused on the number of kernels for each convolutional block. The convolutional layer kernel length has been investigated in the previous study [5]. We set the kernel length of all convolutional layers to 3.

The number of hidden units in the fully connected layers was all set to the same value during the hyperparameter tuning process to reduce the search space. Furthermore, we performed the hyperparameter tuning based on the MESA dataset - the largest dataset containing the cardiac and activity data to date. Therefore, we expected the hyperparameter tuning could discover robust backbone networks for this study.

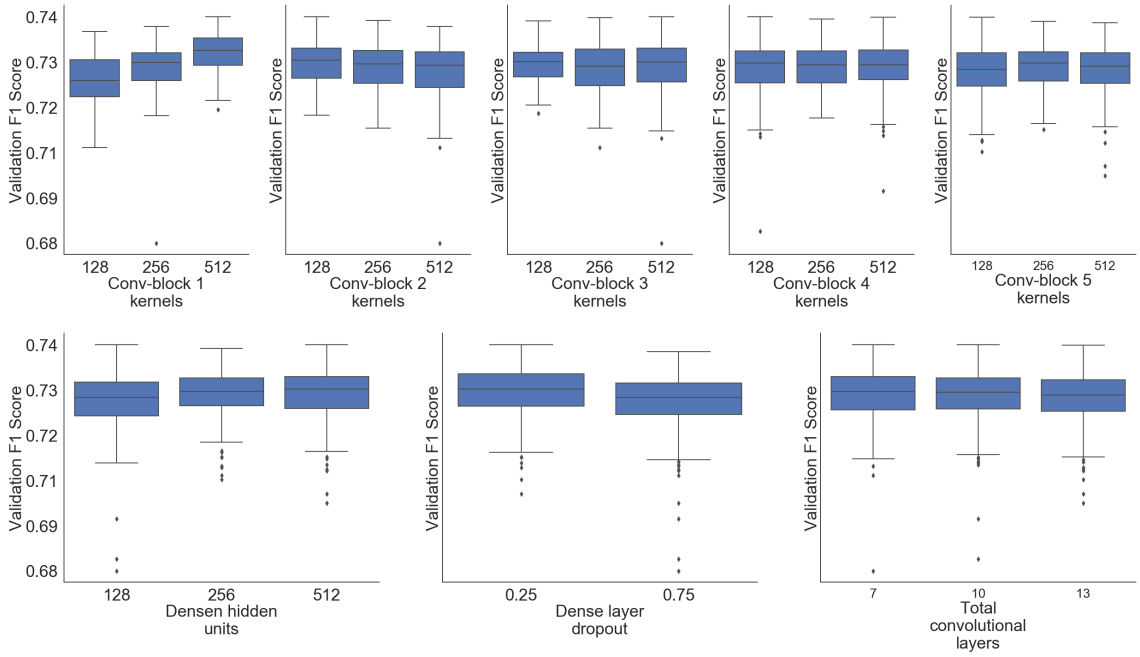


Figure B.1 DeepCNN backbone network hyper-parameters tuning results

Block/Group	Layer	Kernel size	Number of Kernels	Padding	Stride	Hidden Units	Drop out Rate
Convolutional Block 1	Convolutional Layer 1 & 2	3	128, 256, 512	1	1		
	Maxpooling	2			2		
Convolutional Block 2	Convolutional Layer 3 & 4	3	128, 256, 512	1	1		
	Maxpooling	2			2		
Convolutional Block 3	Convolutional Layer 5, 6 & 7	3	128, 256, 512	1	1		
	Maxpooling	2			2		
Convolutional Block 4	Convolutional Layer 8, 9 & 10	3	128, 256, 512	1	1		
	Maxpooling	2			2		
Convolutional Block 4	Convolutional Layer 11, 12 & 13	3	128, 256, 512	1	1		
	Maxpooling	2			2		
FC Block	Fully Connected Layer 1, 2					128, 256, 512	
	Drop Out						0.25, 0.75

Table B.1 Hyper-parameters tuning for backbone networks

B.2. HEART RATE STATISTIC FEATURES COMBINED WITH DEEP MOVEMENT FEATURES

In our study, we also tested whether using the raw accelerometer data could achieve better results. Therefore, we designed two feasible CNNs that could extract the compatible deep features fused with the HR statistic features.The rational behind the network design was to produce a compatible representation of the HR intermediate feature. To match up the dimension of latent feature, we firstly reduced the accelerometer data sampling rate from 50Hz/20Hz to 1Hz. We then designed two CNNs to bridge the sampling gap between movement and cardiac sensing features, one for the early stage fusion and another for the late-stage and hybrid fusion. Each consisted of six convolutional layers (two convolutional blocks) to extract the deep movement feature used for fusion study. For the early-stage fusion, the network was called AccCNN-1. The hybrid and late-stage fusion used the same network to extract the latent representations, and we called it AccCNN-2. We referred to the entire network as DeepMixCNN and ResDeepMixCNN, respectively. The network structure and the experiment setting details can be seen in Figure B.2 and Figure B.3. We adopted the leave-two-subjects-out cross validation experimental setting on Apple Watch dataset. The training, validation and testing process used the same settings as the main content. However, we did not conduct the hyperparameter search together with the backbone network. Therefore, the network designed in the study merely served as feasible networks for the study, yet it might not be the best performing CNN. We focused on the fusion techniques rather than the contribution of network structure.

The MESA dataset contained the activity counts sampled at 1/30 Hz, which technically was no the raw data. In addition, cardiac sensing was acquired via the PSG equipment, which may be difficult to wear everyday. Therefore, the HRV features derived from the RR intervals were most likely to be available from the commercial wearable devices (e.g., photoplethysmogram data), so we did not conduct the experiments on the raw PSG data.The details of this experiment are listed in Table B.2

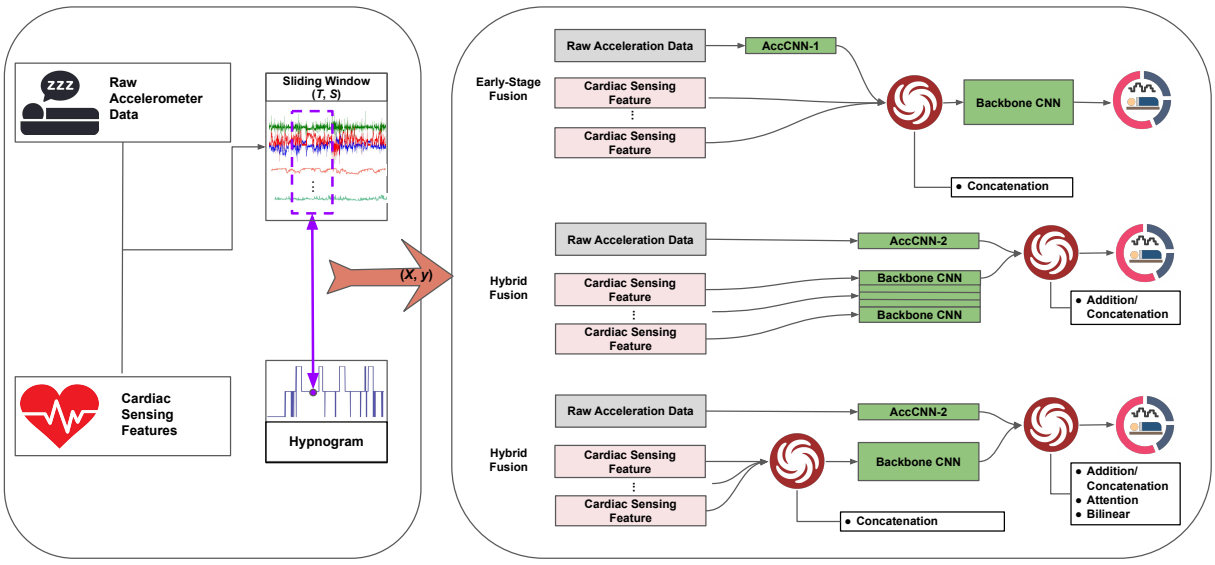


Figure B.2 An overview of the three-stage sleep classification system using the raw accelerometer data with HR statistics features. The raw accelerometer data and HR statistic features were extracted for each sleep epoch (30s). The sliding window method divides the sleep data into multiple segments with window length T and stride S. In this experiment, we have $T = 101$, and $S = 1$. We firstly use the AccCNN to learn deep features then fuse them with HR statistic features. The hypnogram represents the stages of sleep over time. Two fusion strategies and four fusion methods were studied.

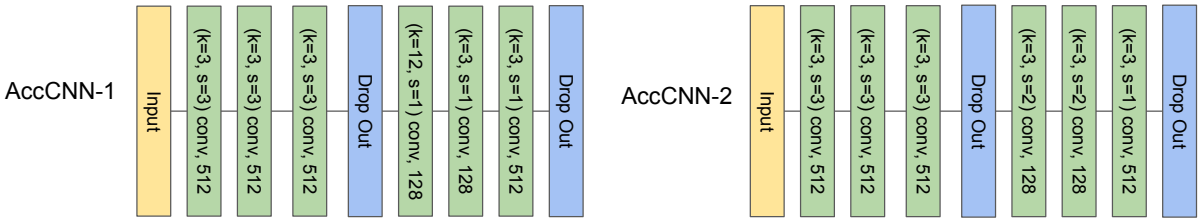


Figure B.3 An overview of the two subnets used to extract the deep features from the raw accelerometer data.

B.2.1. Raw Accelerometer Data and HRS Features

The highest performed model was ResDeepMixCNN in late-stage fusion, using the concatenation method. Its accuracy, the Cohen's κ score and the mean F1 reached 79.1 %, 51.4 and 66.7 % respectively. Thus, the results were comparable to the handcraft features.

B.2.2. Comparison of Raw Data and Intermediate Features

We compared the performance difference between using the raw accelerometer data and using the clinical/handcraft features based on the window length of 101. The ResDeepMixCNN has achieved the comparable performance on the Apple Watch dataset in terms of accuracy, Cohen's κ and mean F1, using the concatenation method in the late-stage fusion. The confusion matrices shown in Figure 5.6 demonstrated the model prediction using raw accelerometer data is biased to NREM sleep. Three reasons might cause the increased bias. The first reason may be the Apple Watch dataset has class imbalance issue. The second reason may be the modality bias of the raw accelerometer data because the wrist movement may not reflect the sleep stage (mainly NREM

Fusion Specifics			Performance Metrics			Deployment Metrics	
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy (%)**	**Cohen's κ**	**Mean F1 (%)**	**Model Size (M)**	**Inference Time (ms)**
Early-Stage	DeepMixCNN	Concatenation	74.8 ± 2.7	34.0 ± 5.9	57.2 ± 3.5	11.9	18.86±1.19
	ResDeepMixCNN	Concatenation	79.8 ± 3.1	48.9 ± 7.1	64.5 ± 4.5	11.9	16.45±0.43
Late-Stage Fusion	DeepMixCNN	Concatenation	79.2 ± 2.8	48.9 ± 7.6	66.0 ± 4.6	50.8	37.03±0.34
		Addition	76.7 ± 2.1	38.7 ± 5.9	58.0 ± 3.7	11.5	32.75±0.19
	ResDeepMixCNN	Concatenation	$\mathbf{79.1 \pm 3.3}$	$\mathbf{51.4 \pm 8.0}$	$\mathbf{66.7 \pm 4.7}$	50.8	31.19±0.72
		Addition	78.9 ± 2.8	48.4 ± 6.4	63.9 ± 4.0	11.5	33.11±0.33
Hybrid	DeepMixCNN	Concatenation	77.3 ± 3.3	43.9 ± 7.3	63.6 ± 4.3	18.0	15.94±0.22
		Addition	75.8 ± 2.6	36.7 ± 7.3	59.0 ± 4.0	11.5	16.28±0.57
		Attention-on-Mov	75.8 ± 3.1	39.4 ± 7.9	60.8 ± 4.8	18.3	15.7±0.18
		Attention-on-Car	71.9 ± 3.1	30.4 ± 8.1	55.7 ± 4.8	18.3	15.7±0.18
		Bilinear	73.1 ± 3.4	31.4 ± 8.2	52.3 ± 5.2	273.9	18.89±0.43
	ResDeepMixCNN	Concatenation	77.7 ± 2.6	42.8 ± 6.0	62.6 ± 3.5	18.0	15.6±0.48
		Addition	80.3 ± 2.9	48.0 ± 7.3	64.4 ± 4.3	11.5	15.65±0.31
		Attention-on-Mov	77.1 ± 2.4	44.8 ± 5.9	62.7 ± 3.6	18.3	16.24±0.32
		Attention-on-Car	75.9 ± 3.0	37.0 ± 7.7	57.7 ± 4.4	18.3	16.24±0.32
		Bilinear	72.8 ± 3.2	29.8 ± 6.5	52.1 ± 4.4	273.9	18.88±0.4

Table B.2 Three-stage sleep classification results (mean $\pm$ standard error at 95% confidence interval) using raw accelerometer data and HRS features based on DeepMixCNN and ResDeepMixCNN with the Apple Watch Dataset for each combination of fusion strategy and method. The experiments were performed using the same experimental setting as in the main content and evaluated at the subject level during recording period based on window length of 101.

and REM) that much. The third reason may be caused by the lacking of hyperparameter search on the network.

Our observations corroborate a study using raw PSG signals for sleep stage classification [156]. That is using intermediate features instead of raw accelerometer data may alleviate the modality bias in the three-sleep stage classification task, while reducing the model parameters.

B.3. THREE SLEEP STAGE CLASSIFICATION PERFORMANCE ON 21, 51 WINDOW LENGTH

B.3.1. The Effects of Sliding Windows Length

In addition to the window length of 101, we also conducted experiments based on the window lengths 51 and 21 followed the previous work [5]. For the Apple Watch dataset, the models with the highest mean F1, accuracy and Cohen's κ score in each fusion strategy were all based on the window length of 101. For the MESA dataset, we observed similar patterns on all feature settings. One possible explanation was when the time step of the input data became shorter, the intermediate features around the time point of the prediction might not contain enough information for three-stage sleep classification. This phenomenon corroborated the previous findings [5].

Fusion Specifics			Performance Metrics			Time Deviation (min.)		
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy (%)**	**Cohen's κ**	**Mean F1 (%)**	**Non-REM sleep**	**REM sleep**	**Wake**
Early-Stage Fusion	DeepCNN	Concatenation	73.1 ± 3.1	40.5 ± 6.7	58.7 ± 4.4	9.1 ± 22.5	-0.1 ± 23.5	-9.0 ± 8.8
	ResDeepCNN	Concatenation	75.4 ± 2.9	43.9 ± 6.4	61.1 ± 3.9	23.0 ± 24.1	-12.5 ± 23.7	-10.5 ± 7.2
Late-Stage Fusion	DeepCNN	Concatenation	74.1 ± 2.7	41.9 ± 6.1	59.8 ± 3.7	12.0 ± 21.6	-1.1 ± 22.1	-10.9 ± 7.9
		Addition	78.0 ± 2.3	50.1 ± 7.0	65.5 ± 3.6	1.3 ± 16.7	11.6 ± 16.8	-12.9 ± 6.6
	ResDeepCNN	Concatenation	76.6 ± 2.5	45.9 ± 6.9	62.6 ± 4.2	20.1 ± 18.8	-13.5 ± 19.6	-6.7 ± 7.8
		Addition	**77.7 ± 2.2**	**48.1 ± 6.8**	**64.6 ± 4.0**	7.6 ± 19.5	3.4 ± 19.6	-11.0 ± 5.8
Hybrid Fusion	DeepCNN	Concatenation	72.5 ± 3.2	39.0 ± 6.2	58.8 ± 3.9	7.5 ± 24.6	5.3 ± 24.6	-12.8 ± 6.8
		Addition	73.3 ± 3.0	39.2 ± 6.3	58.4 ± 3.6	17.1 ± 24.1	-0.8 ± 24.1	-16.3 ± 5.5
		Attention-on-Mov	72.6 ± 3.0	39.0 ± 6.0	59.4 ± 3.4	15.1 ± 23.6	-1.7 ± 23.6	-13.4 ± 6.6
		Attention-on-Car	72.7 ± 2.9	37.2 ± 6.4	58.3 ± 3.7	23.7 ± 21.0	-8.6 ± 20.3	-15.2 ± 6.1
		Bilinear	72.3 ± 2.6	38.2 ± 5.5	58.5 ± 3.2	1.4 ± 20.7	12.3 ± 20.7	-13.7 ± 6.3
	ResDeepCNN	Concatenation	73.6 ± 2.9	41.5 ± 6.5	60.7 ± 4.1	8.8 ± 23.1	1.4 ± 24.9	-10.2 ± 7.4
		Addition	73.1 ± 3.1	40.5 ± 6.5	59.7 ± 3.9	11.5 ± 23.8	-0.8 ± 24.4	-10.7 ± 6.8
		Attention-on-Mov	74.4 ± 3.3	43.8 ± 6.1	61.7 ± 3.8	11.6 ± 20.9	-1.5 ± 21.5	-10.1 ± 6.9
		Attention-on-Car	73.1 ± 3.0	40.6 ± 7.1	60.4 ± 4.1	2.1 ± 24.0	6.9 ± 24.3	-9.0 ± 7.9
		Bilinear	74.9 ± 2.6	42.2 ± 6.5	60.2 ± 3.9	22.2 ± 21.9	-12.2 ± 22.0	-9.9 ± 7.1

Table B.3 Three-stage sleep classification results (mean ± standard error at 95% confidence interval) for each combination of the fusion strategy and method with the Apple Watch dataset using the ACT-HRS feature and evaluated at subject level during the recording period based on the window length of 51.

Fusion Specifics			Performance Metrics			Time Deviation (min.)		
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy (%)**	**Cohen's κ**	**Mean F1 (%)**	**Non-REM sleep**	**REM sleep**	**Wake**
Early-Stage Fusion	DeepCNN	Concatenation	71.9 ± 2.1	35.2 ± 5.5	54.9 ± 3.5	22.9 ± 16.6	-13.9 ± 16.8	-9.0 ± 8.3
	ResDeepCNN	Concatenation	73.4 ± 2.4	38.7 ± 5.8	58.0 ± 3.6	17.3 ± 15.2	-9.6 ± 16.0	-7.7 ± 9.2
Late-Stage Fusion	DeepCNN	Concatenation	72.4 ± 2.5	38.1 ± 6.1	55.9 ± 3.9	21.9 ± 18.2	-7.4 ± 18.0	-14.6 ± 8.6
		Addition	**75.3 ± 2.4**	**39.7 ± 7.7**	**59.2 ± 4.2**	14.6 ± 13.9	-1.4 ± 15.4	-13.1 ± 7.2
	ResDeepCNN	Concatenation	72.7 ± 2.6	38.3 ± 6.8	57.3 ± 4.2	11.9 ± 20.3	-4.5 ± 21.0	-7.4 ± 8.8
		Addition	74.3 ± 2.7	38.9 ± 7.5	58.9 ± 4.2	6.2 ± 16.6	4.5 ± 17.6	-10.7 ± 7.9
Hybrid Fusion	DeepCNN	Concatenation	71.6 ± 2.5	35.8 ± 6.2	55.6 ± 4.0	17.7 ± 23.5	-0.0 ± 23.4	-17.7 ± 6.4
		Addition	71.6 ± 2.6	35.0 ± 5.7	55.8 ± 3.6	22.9 ± 19.7	-5.9 ± 19.7	-17.0 ± 6.7
		Attention-on-Mov	72.2 ± 2.6	37.3 ± 5.4	56.6 ± 3.4	20.5 ± 19.6	-3.6 ± 20.8	-16.9 ± 6.3
		Attention-on-Car	73.2 ± 2.3	35.4 ± 5.9	56.6 ± 3.6	31.1 ± 20.6	-13.3 ± 20.1	-17.8 ± 6.4
		Bilinear	71.5 ± 2.9	37.3 ± 6.0	57.2 ± 3.7	5.0 ± 19.1	6.9 ± 17.7	-11.9 ± 7.7
	ResDeepCNN	Concatenation	72.3 ± 2.7	36.6 ± 6.5	57.3 ± 4.0	21.3 ± 23.8	-5.3 ± 23.3	-15.9 ± 6.5
		Addition	71.6 ± 2.2	35.2 ± 5.2	55.5 ± 3.4	24.7 ± 18.2	-11.6 ± 19.1	-13.2 ± 6.9
		Attention-on-Mov	73.3 ± 2.3	38.0 ± 5.4	57.9 ± 3.3	33.5 ± 17.3	-19.3 ± 17.7	-14.2 ± 6.4
		Attention-on-Car	72.1 ± 2.7	37.8 ± 5.5	57.8 ± 3.6	17.5 ± 19.2	-5.8 ± 18.8	-11.7 ± 7.9
		Bilinear	70.5 ± 2.7	35.7 ± 5.8	55.9 ± 3.6	3.1 ± 19.2	1.2 ± 19.5	-4.3 ± 9.8

Table B.4 Three-stage sleep classification results (mean ± standard error at 95% confidence interval) for each combination of fusion strategy and method with the Apple Watch dataset using the ACT-HRS feature and evaluated at subject level during the recording period based on the window length of 21.

Fusion Specifics			Performance Metrics			Time Deviation (min.)		
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy (%)**	**Cohen's κ**	**Mean F1 (%)**	**Non-REM sleep**	**REM sleep**	**Wake**
Early-Stage Fusion	DeepCNN	Concatenation	78.1 ± 0.9	60.1 ± 1.8	69.3 ± 1.3	14.6 ± 7.3	-18.7 ± 3.6	4.1 ± 6.9
	ResDeepCNN	Concatenation	76.7 ± 1.0	60.2 ± 1.9	70.3 ± 1.3	-15.6 ± 7.4	12.6 ± 5.0	3.0 ± 6.8
Late-Stage Fusion	DeepCNN	Concatenation	**78.4 ± 1.0**	**62.5 ± 1.8**	**71.4 ± 1.3**	9.9 ± 7.3	0.2 ± 4.1	-10.0 ± 6.4
		Addition	76.5 ± 0.9	58.8 ± 1.7	66.7 ± 1.2	62.3 ± 7.4	-21.7 ± 3.7	-40.6 ± 6.9
	ResDeepCNN	Concatenation	77.7 ± 0.9	61.0 ± 1.8	69.9 ± 1.2	17.3 ± 7.6	-4.6 ± 4.2	-12.6 ± 6.6
		Addition	74.8 ± 1.0	55.7 ± 1.8	66.2 ± 1.3	-21.0 ± 7.7	-16.0 ± 4.2	37.0 ± 7.5
Hybrid Fusion	DeepCNN	Concatenation	76.4 ± 1.1	61.1 ± 1.8	70.0 ± 1.3	-9.4 ± 8.0	17.9 ± 5.0	-8.5 ± 6.9
		Addition	77.2 ± 1.0	61.3 ± 1.7	70.6 ± 1.2	-17.9 ± 7.5	2.5 ± 4.3	15.4 ± 7.0
		Attention-on-Mov	74.6 ± 1.1	59.1 ± 1.8	69.0 ± 1.2	-24.0 ± 8.1	39.2 ± 5.7	-15.3 ± 6.6
		Attention-on-Car	77.8 ± 0.9	60.7 ± 1.8	69.8 ± 1.2	5.4 ± 7.4	-9.2 ± 4.1	3.8 ± 6.7
		Bilinear	77.1 ± 0.9	60.1 ± 1.8	70.2 ± 1.2	6.8 ± 8.0	13.9 ± 5.0	-20.8 ± 6.5
	ResDeepCNN	Concatenation	77.7 ± 1.0	62.4 ± 1.7	71.0 ± 1.2	8.3 ± 7.7	15.5 ± 5.0	-23.8 ± 6.3
		Addition	78.5 ± 0.9	62.0 ± 1.7	71.3 ± 1.2	30.4 ± 7.3	4.1 ± 4.3	-34.5 ± 6.5
		Attention-on-Mov	76.1 ± 1.1	60.7 ± 1.8	70.4 ± 1.3	-9.9 ± 8.2	31.3 ± 5.7	-21.4 ± 6.8
		Attention-on-Car	76.6 ± 1.1	61.2 ± 1.8	70.7 ± 1.2	-0.3 ± 7.6	25.8 ± 5.1	-25.6 ± 6.2
		Bilinear	76.7 ± 0.9	60.2 ± 1.7	69.7 ± 1.2	-9.5 ± 7.6	9.4 ± 4.6	0.1 ± 6.6

Table B.5 Three-stage sleep classification results (mean ± standard error at 95% confidence interval) for each combination of fusion strategy and method with the MESA test dataset using the ACT-HRS evaluated at subject level during the recording period based on the window length of 51.

Fusion Specifics			Performance Metrics			Time Deviation (min.)		
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy (%)**	**Cohen's κ**	**Mean F1 (%)**	**Non-REM sleep**	**REM sleep**	**Wake**
Early-Stage Fusion	DeepCNN	Concatenation	75.3 ± 1.0	55.2 ± 1.8	66.7 ± 1.2	61.4 ± 7.7	-4.3 ± 4.2	-57.0 ± 6.8
	ResDeepCNN	Concatenation	75.0 ± 0.9	56.6 ± 1.7	68.3 ± 1.1	13.3 ± 7.4	16.7 ± 4.6	-30.0 ± 6.5
Late-Stage Fusion	DeepCNN	Concatenation	76.6 ± 0.9	57.8 ± 1.6	68.0 ± 1.2	29.8 ± 7.0	-13.9 ± 3.7	-15.9 ± 6.4
		Addition	74.4 ± 0.9	56.9 ± 1.6	66.8 ± 1.1	13.7 ± 7.5	6.9 ± 4.8	-20.7 ± 6.4
	ResDeepCNN	Concatenation	75.9 ± 0.9	58.1 ± 1.6	68.8 ± 1.1	7.8 ± 7.1	6.8 ± 4.1	-14.7 ± 6.4
		Addition	74.7 ± 0.9	56.6 ± 1.6	66.6 ± 1.1	33.4 ± 7.4	2.7 ± 4.7	-36.1 ± 6.4
Hybrid Fusion	DeepCNN	Concatenation	76.0 ± 0.9	57.1 ± 1.7	68.6 ± 1.2	33.8 ± 7.6	8.2 ± 4.4	-42.0 ± 6.7
		Addition	74.8 ± 1.0	55.7 ± 1.8	67.9 ± 1.2	22.6 ± 7.8	21.7 ± 4.9	-44.3 ± 6.6
		Attention-on-Mov	74.2 ± 1.0	56.8 ± 1.7	67.9 ± 1.2	-25.4 ± 7.9	18.4 ± 4.7	7.0 ± 7.0
		Attention-on-Car	76.1 ± 0.9	57.4 ± 1.6	68.3 ± 1.1	13.1 ± 7.5	-1.0 ± 4.3	-12.1 ± 6.7
		Bilinear	76.9 ± 0.9	57.6 ± 1.7	68.0 ± 1.2	26.4 ± 7.2	-20.1 ± 3.6	-6.3 ± 6.8
	ResDeepCNN	Concatenation	76.5 ± 0.9	57.6 ± 1.7	68.6 ± 1.2	47.2 ± 7.4	-0.2 ± 4.1	-47.0 ± 6.6
		Addition	76.2 ± 0.9	57.7 ± 1.7	68.8 ± 1.2	25.0 ± 7.4	7.6 ± 4.3	-32.6 ± 6.6
		Attention-on-Mov	75.9 ± 0.9	58.1 ± 1.7	**69.0 ± 1.1**	-0.2 ± 7.5	10.0 ± 4.3	-9.8 ± 6.7
		Attention-on-Car	75.5 ± 0.9	58.2 ± 1.7	68.8 ± 1.2	-5.1 ± 7.5	13.2 ± 4.6	-8.2 ± 6.6
		Bilinear	**77.0 ± 0.9**	**58.7 ± 1.6**	68.6 ± 1.1	39.9 ± 7.1	-13.1 ± 3.9	-26.8 ± 6.5

Table B.6 Three-stage sleep classification results (mean ± standard error at 95% confidence interval) for each combination of fusion strategy and method with the MESA test dataset using the ACT-HRS evaluated at subject level during the recording period based on the window length of 21.

Fusion Specifics			Performance Metrics			Time Deviation (min.)		
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy (%)**	**Cohen's κ**	**Mean F1 (%)**	**Non-REM sleep**	**REM sleep**	**Wake**
Early-Stage Fusion	DeepCNN	Concatenation	75.7 ± 1.0	56.7 ± 1.8	67.2 ± 1.3	40.0 ± 6.9	-11.1 ± 3.9	-28.9 ± 6.5
	ResDeepCNN	Concatenation	76.0 ± 0.9	56.9 ± 1.8	66.8 ± 1.3	63.1 ± 7.3	-17.3 ± 3.6	-45.9 ± 6.8
Late-Stage Fusion	DeepCNN	Concatenation	78.4 ± 0.9	61.7 ± 1.8	70.2 ± 1.3	20.8 ± 6.9	-10.9 ± 3.7	-9.8 ± 6.4
		Addition	77.6 ± 0.9	60.5 ± 1.8	69.2 ± 1.2	34.3 ± 6.9	-8.6 ± 3.9	-25.7 ± 6.5
	ResDeepCNN	Concatenation	78.0 ± 1.0	**62.4 ± 1.7**	**71.1 ± 1.2**	2.8 ± 7.2	3.3 ± 4.1	-6.1 ± 6.4
		Addition	77.5 ± 0.9	61.1 ± 1.8	70.3 ± 1.2	16.8 ± 6.9	2.4 ± 4.0	-19.2 ± 6.5
Hybrid Fusion	DeepCNN	Concatenation	77.2 ± 1.0	60.3 ± 1.7	69.7 ± 1.3	12.5 ± 7.5	-0.5 ± 4.4	-12.0 ± 6.5
		Addition	76.6 ± 1.0	59.7 ± 1.8	69.7 ± 1.3	14.2 ± 7.4	13.4 ± 4.6	-27.6 ± 6.5
		Attention-on-Mov	77.5 ± 1.0	61.2 ± 1.7	70.5 ± 1.3	35.3 ± 7.3	5.7 ± 4.8	-40.9 ± 6.4
		Attention-on-Car	77.8 ± 0.9	60.4 ± 1.7	68.5 ± 1.3	45.2 ± 7.1	-20.6 ± 3.8	-24.7 ± 6.4
		Bilinear	77.1 ± 1.0	60.8 ± 1.7	70.6 ± 1.2	5.4 ± 6.9	12.2 ± 4.4	-17.7 ± 6.2
	ResDeepCNN	Concatenation	76.7 ± 1.1	60.7 ± 1.9	70.6 ± 1.3	6.2 ± 7.9	23.4 ± 5.4	-29.6 ± 6.3
		Addition	76.7 ± 1.1	61.3 ± 1.8	70.9 ± 1.3	-27.4 ± 7.2	23.5 ± 4.8	3.8 ± 6.6
		Attention-on-Mov	**78.7 ± 0.9**	62.2 ± 1.7	70.0 ± 1.3	43.9 ± 6.9	-17.6 ± 3.8	-26.3 ± 6.4
		Attention-on-Car	78.3 ± 0.9	62.2 ± 1.7	70.6 ± 1.3	37.7 ± 7.1	-4.7 ± 4.4	-33.0 ± 6.2
		Bilinear	76.8 ± 1.0	59.1 ± 1.8	69.6 ± 1.2	24.0 ± 7.0	6.7 ± 4.2	-30.6 ± 6.2

Table B.7 Three-stage sleep classification results (mean ± standard error at 95% confidence interval) for each combination of fusion strategy and method in the MESA test dataset using the ACT-HRV feature set evaluated at subject level during the recording period based on the window length of 51.

Fusion Specifics			Performance Metrics			Time Deviation (min.)		
Fusion Strategy	**Network**	**Fusion Method**	**Accuracy (%)**	**Cohen's κ**	**Mean F1 (%)**	**Non-REM sleep**	**REM sleep**	**Wake**
Early-Stage Fusion	DeepCNN	Concatenation	75.9 ± 0.9	57.0 ± 1.7	67.5 ± 1.2	34.7 ± 7.0	-10.2 ± 3.7	-24.6 ± 6.6
	ResDeepCNN	Concatenation	75.4 ± 0.9	56.0 ± 1.7	67.2 ± 1.2	8.5 ± 7.1	-10.9 ± 3.7	2.4 ± 6.7
Late-Stage Fusion	DeepCNN	Concatenation	76.0 ± 0.9	57.5 ± 1.7	68.0 ± 1.1	12.1 ± 7.1	-4.5 ± 3.9	-7.6 ± 6.5
		Addition	74.4 ± 0.9	54.4 ± 1.7	64.0 ± 1.2	26.7 ± 7.7	-21.6 ± 3.8	-5.1 ± 7.2
	ResDeepCNN	Concatenation	76.2 ± 0.9	58.2 ± 1.7	68.0 ± 1.2	16.7 ± 6.9	-7.1 ± 3.8	-9.6 ± 6.5
		Addition	75.3 ± 0.9	56.0 ± 1.7	65.4 ± 1.2	38.7 ± 7.6	-19.2 ± 3.8	-19.5 ± 7.0
Hybrid Fusion	DeepCNN	Concatenation	75.8 ± 1.0	57.5 ± 1.7	68.5 ± 1.2	21.2 ± 7.7	3.6 ± 4.4	-24.8 ± 7.0
		Addition	75.2 ± 1.0	57.0 ± 1.8	68.5 ± 1.2	11.3 ± 7.4	21.4 ± 5.2	-32.7 ± 6.7
		Attention-on-Mov	75.4 ± 1.0	56.7 ± 1.7	67.7 ± 1.2	15.3 ± 7.5	3.9 ± 4.9	-19.3 ± 6.7
		Attention-on-Car	74.3 ± 1.0	56.6 ± 1.7	67.8 ± 1.2	-14.6 ± 7.5	14.7 ± 4.6	-0.1 ± 6.5
		Bilinear	74.9 ± 1.0	57.4 ± 1.8	68.3 ± 1.2	-13.4 ± 7.9	8.6 ± 4.3	4.7 ± 7.2
	ResDeepCNN	Concatenation	75.5 ± 1.0	57.1 ± 1.8	68.7 ± 1.2	19.8 ± 7.4	20.8 ± 5.0	-40.6 ± 6.6
		Addition	76.2 ± 0.9	58.4 ± 1.7	69.0 ± 1.2	19.6 ± 7.1	-0.9 ± 4.1	-18.7 ± 6.6
		Attention-on-Mov	76.0 ± 1.0	**58.7 ± 1.8**	**69.0 ± 1.2**	18.5 ± 7.4	10.3 ± 4.9	-28.8 ± 6.4
		Attention-on-Car	75.4 ± 1.0	58.0 ± 1.7	68.6 ± 1.2	1.5 ± 7.5	12.9 ± 4.8	-14.4 ± 6.5
		Bilinear	**76.4 ± 0.9**	58.7 ± 1.8	68.8 ± 1.2	37.8 ± 7.1	-0.8 ± 4.1	-36.9 ± 6.5

Table B.8 Three-stage sleep classification results (mean ± standard error at 95% confidence interval) for each combination of fusion strategy and method with the MESA test dataset using the ACT-HRV evaluated at subject level during the recording period based on the window length of 21.

Bibliography

[1] I. Perez-Pozuelo, B. Zhai, J. Palotti, R. Mall, M. Aupetit, J. M. Garcia-Gomez, S. Taheri, Y. Guan, and L. Fernandez-Luque, "The future of sleep health: a data-driven revolution in sleep science and medicine," pp. 1–15, 3 2020. [Online]. Available: https://www.nature.com/articles/s41746-020-0244-4

[2] K. Zarzycki and M. Ławryńczuk, "Lstm and gru neural networks as models of dynamical processes used in predictive control: A comparison of models developed for two chemical reactors," *Sensors*, vol. 21, no. 16, p. 5625, 2021.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020. [Online]. Available: http://gradcam.cloudcv.org

[4] N. Gozzi, D. Perrotta, D. Paolotti, and N. Perra, "Towards a data-driven characterization of behavioral changes induced by the seasonal flu," *PLoS computational biology*, vol. 16, no. 5, p. e1007879, 2020.

[5] B. Zhai, I. Perez-Pozuelo, E. A. Clifton, J. Palotti, and Y. 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*, vol. 4, no. 2, p. 67, 6 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3397325

[6] O. Walch, Y. Huang, D. Forger, and C. Goldstein, "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device," *Sleep*, vol. 42, no. 12, 12 2019. [Online]. Available: https://github.com/

[7] A. Doherty, D. Jackson, N. Hammerla, T. Plötz, P. Olivier, M. H. Granat, T. White, V. T. Van Hees, M. I. Trenell, C. G. Owen *et al.*, "Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study," *PloS one*, vol. 12, no. 2, p. e0169649, 2017.

[8] S. E. Jones, V. T. van Hees, D. R. Mazzotti, P. Marques-Vidal, S. Sabia, A. van der Spek, H. S. Dashti, J. Engmann, D. Kocevska, J. Tyrrell *et al.*, "Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[9] S. J. Jaiswal, G. Quer, M. Galarnyk, S. R. Steinhubl, E. J. Topol, and R. L. Owens, "Association of sleep duration and variability with body mass index: sleep measurements in a large us population of wearable sensor users," *JAMA Internal Medicine*, vol. 180, no. 12, pp. 1694–1696, 2020.

[10] J. R. L. Schwartz and T. Roth, "Neurophysiology of sleep and wakefulness: basic science and clinical implications." *Current neuropharmacology*, vol. 6, no. 4, pp. 367–78, 12 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19587857http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2701283

[11] L. Imeri and M. R. Opp, "How (and why) the immune system makes us sleep," pp. 199–210, 3 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19209176http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2839418

[12] E. B. Simon, A. Rossi, A. G. Harvey, and M. P. Walker, "Overanxious and underslept," *Nature Human Behaviour*, vol. 4, no. 1, pp. 100–110, 2020.

[13] N. E. Fultz, G. Bonmassar, K. Setsompop, R. A. Stickgold, B. R. Rosen, J. R. Polimeni, and L. D. Lewis, "Coupled electrophysiological, hemodynamic, and cerebrospinal fluid oscillations in human sleep," *Science*, vol. 366, no. 6465, pp. 628–631, 11 2019. [Online]. Available: http://science.sciencemag.org/

[14] S. Abdullah, M. Matthews, E. L. Murnane, G. Gay, and T. Choudhury, "Towards circadian computing: "early to bed and early to rise" makes some of us unhealthy and sleep deprived," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 673–684. [Online]. Available: https://doi.org/10.1145/2632048.2632100

[15] I. Perez-Pozuelo, B. Zhai, J. Palotti, R. Mall, M. Aupetit, J. M. Garcia-Gomez, S. Taheri, Y. Guan, and L. Fernandez-Luque, "The future of sleep health: a data-driven revolution in sleep science and medicine," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–15, 2020.

[16] X. Sun, L. Qiu, Y. Wu, Y. Tang, and G. Cao, "SleepMonitor: Monitoring Respiratory Rate and Body Position During Sleep Using Smartwatch," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–22, 9 2017. [Online]. Available: https://doi.org/10.1145/3130969

[17] L. Chang, J. Lu, J. Wang, X. Chen, D. Fang, Z. Tang, P. Nurmi, and Z. Wang, "SleepGuard: Capturing Rich Sleep Information Using Smartwatch Sensing Data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–34, 2018. [Online]. Available: https://doi.org/10.1145/3264908

[18] D. Dawson and K. Reid, "Fatigue, alcohol and performance impairment," *Nature*, vol. 388, no. 6639, p. 235, 1997.

[19] S. M. Bertisch, B. D. Pollock, M. A. Mittleman, D. J. Buysse, L. A. Bazzano, D. J. Gottlieb, and S. Redline, "Insomnia with objective short sleep duration and risk of incident cardiovascular disease and all-cause mortality: Sleep heart health study," *Sleep*, vol. 41, no. 6, p. zsy047, 2018.

[20] M. Hafner, M. Stepanek, J. Taylor, W. M. Troxel, and C. van Stolk, "Why sleep matters—the economic costs of insufficient sleep: a cross-country comparative analysis," *Rand health quarterly*, vol. 6, no. 4, 2017.

[21] D. R. Hillman, A. S. Murphy, R. Antic, and L. Pezzullo, "The economic cost of sleep disorders," *Sleep*, vol. 29, no. 3, pp. 299–305, 2006.

[22] R. J. Ozminkowski, S. Wang, and J. K. Walsh, "The direct and indirect costs of untreated insomnia in adults in the united states," *Sleep*, vol. 30, no. 3, pp. 263–273, 2007.

[23] M.-P. St-Onge, A. Mikic, and C. E. Pietrolungo, "Effects of diet on sleep quality," *Advances in Nutrition*, vol. 7, no. 5, pp. 938–949, 2016.

[24] C. E. Kline, "The bidirectional relationship between exercise and sleep: implications for exercise adherence and sleep improvement," *American journal of lifestyle medicine*, vol. 8, no. 6, pp. 375–379, 2014.

[25] J. R. L. Schwartz and T. Roth, "Neurophysiology of sleep and wakefulness: basic science and clinical implications." *Current neuropharmacology*, vol. 6, no. 4, pp. 367–78, dec 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19587857http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2701283

[26] T. Deboer, "Sleep homeostasis and the circadian clock: do the circadian pacemaker and the sleep homeostat influence each other's functioning?" *Neurobiology of sleep and circadian rhythms*, vol. 5, pp. 68–77, 2018.

[27] M. Ohayon, E. M. Wickwire, M. Hirshkowitz, S. M. Albert, A. Avidan, F. J. Daly, Y. Dauvilliers, R. Ferri, C. Fung, D. Gozal *et al.*, "National sleep foundation's sleep quality recommendations: first report," *Sleep Health*, vol. 3, no. 1, pp. 6–19, 2017.

[28] S. Taheri, "The link between short sleep duration and obesity: we should recommend more sleep to prevent obesity," *Archives of Disease in Childhood*, vol. 91, no. 11, pp. 881–884, nov 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17056861http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2082964http://adc.bmj.com/cgi/doi/10.1136/adc.2005.093013

[29] K. M. Awad, A. Malhotra, J. H. Barnet, S. F. Quan, and P. E. Peppard, "Exercise is associated with a reduced incidence of sleep-disordered breathing," *The American journal of medicine*, vol. 125, no. 5, pp. 485–490, 2012.

[30] M. Walker, *Why we sleep: The new science of sleep and dreams*. Penguin UK, 2017.

[31] J. S. Higgins, J. Michael, R. Austin, T. Åkerstedt, H. P. A. Van Dongen, N. Watson, C. Czeisler, A. I. Pack, and M. R. Rosekind, "Asleep at the Wheel—The Road to Addressing Drowsy Driving," *Sleep*, vol. 40, no. 2, 01 2017, zsx001. [Online]. Available: https://doi.org/10.1093/sleep/zsx001

[32] S. Garbarino, O. Guglielmi, A. Sanna, G. L. Mancardi, and N. Magnavita, "Risk of occupational accidents in workers with obstructive sleep apnea: systematic review and meta-analysis," *Sleep*, vol. 39, no. 6, pp. 1211–1218, 2016.

[33] M. Willetts, S. Hollowell, L. Aslett, C. Holmes, and A. Doherty, "Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants," *Scientific Reports*, vol. 8, no. 1, p. 7961, dec 2018. [Online]. Available: http://www.nature.com/articles/s41598-018-26174-1

[34] A. Stefani and B. Högl, "Sleep in Parkinson's disease," *Neuropsychopharmacology 2019 45:1*, vol. 45, no. 1, pp. 121–128, 6 2019. [Online]. Available: https://www.nature.com/articles/s41386-019-0448-y

[35] R. B. Postuma, J. F. Gagnon, J. A. Bertrand, D. Génier Marchand, and J. Y. Montplaisir, "Parkinson risk in idiopathic REM sleep behavior disorder: Preparing for neuroprotective trials," *Neurology*, vol. 84, no. 11, pp. 1104–1113, 3 2015. [Online]. Available: /pmc/articles/PMC4371408//pmc/articles/PMC4371408/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4371408/

[36] C. H. Schenck, B. F. Boeve, and M. W. Mahowald, "Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder: a 16-year update on a previously reported series," *Sleep medicine*, vol. 14, no. 8, pp. 744–748, 2013.

[37] N. Cooray, F. Andreotti, C. Lo, M. Symmonds, M. T. Hu, and M. De Vos, "Detection of rem sleep behaviour disorder by automated polysomnography analysis," *Clinical Neurophysiology*, vol. 130, no. 4, pp. 505–514, 2019.

[38] E. Topol, *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.

[39] T. I. Morgenthaler, T. Lee-Chiong, C. Alessi, L. Friedman, R. N. Aurora, B. Boehlecke, T. Brown, A. L. Chesson, V. Kapur, R. Maganti, J. Owens, J. Pancer, T. J. Swick, R. Zak, and R. Standards of Practice Committee of the American Academy of Sleep Medicine, "Practice parameters for the clinical evaluation and treatment of circadian rhythm sleep disorders. An American Academy of Sleep Medicine report." *Sleep*, vol. 30, no. 11, pp. 1445–59, nov 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18041479http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2082098

[40] R. Berry, *Fundamentals of Sleep Medicine*. Elsevier Inc., 2012.

[41] IBER and C., "The AASM Manual for the Scoring of Sleep and Associated Events : Rules," *Terminology and Technical Specification*, 2007. [Online]. Available: https://ci.nii.ac.jp/naid/10024500923

[42] T. E, L. R, E. DA, and V. C. E, "Slow-wave sleep and the risk of type 2 diabetes in humans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 3, pp. 1044–1049, 1 2008. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18172212/

[43] S. Reutrakul and E. Van Cauter, "Sleep influences on obesity, insulin resistance, and risk of type 2 diabetes," *Metabolism: clinical and experimental*, vol. 84, pp. 56–66, 7 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29510179/

[44] M. Y. Agargun and R. Cartwright, "REM sleep, dream variables and suicidality in depressed patients," *Psychiatry Research*, vol. 119, no. 1-2, pp. 33–39, 7 2003.

[45] D. J. Kupfer and F. G. Foster, "INTERVAL BETWEEN ONSET OF SLEEP AND RAPID-EYE-MOVEMENT SLEEP AS AN INDICATOR OF DEPRESSION," *The Lancet*, vol. 300, no. 7779, pp. 684–686, 9 1972. [Online]. Available: http://www.thelancet.com/article/S0140673672920909/fulltexthttp://www.thelancet.com/article/S0140673672920909/abstracthttps://www.thelancet.com/journals/lancet/article/PIIS0140-6736(72)92090-9/abstract

[46] L. Palagini, C. Baglioni, A. Ciapparelli, A. Gemignani, and D. Riemann, "Rem sleep dysregulation in depression: state of the art," *Sleep medicine reviews*, vol. 17, no. 5, pp. 377–390, 2013.

[47] Y.-Q. Wang, R. Li, M.-Q. Zhang, Z. Zhang, W.-M. Qu, and Z.-L. Huang, "The Neurobiological Mechanisms and Treatments of REM Sleep Disturbances in Depression," *Current Neuropharmacology*, vol. 13, no. 4, p. 543, 9 2015. [Online]. Available: /pmc/articles/PMC4790401//pmc/articles/PMC4790401/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4790401/

[48] J. X. Teo, S. Davila, C. Yang, A. A. Hii, C. J. Pua, J. Yap, S. Y. Tan, A. Sahlén, C. W.-L. Chin, B. T. Teh, S. G. Rozen, S. A. Cook, K. K. Yeo, P. Tan, and W. K. Lim, "Digital phenotyping by consumer wearables identifies sleep-associated markers of cardiovascular disease risk and biological aging," *Communications Biology 2019 2:1*, vol. 2, no. 1, pp. 1–10, 10 2019. [Online]. Available: https://www.nature.com/articles/s42003-019-0605-1

[49] R. Ravichandran, S. W. Sien, S. N. Patel, J. A. Kientz, and L. R. Pina, "Making sense of sleep sensors: How sleep sensing technologies support and undermine sleep health," *Conference on Human Factors in Computing Systems - Proceedings*, vol. 2017-May, pp. 6864–6875, 5 2017. [Online]. Available: http://dx.doi.org/10.1145/3025453.3025557

[50] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms," pp. 342–392, 5 2003. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/12749557/

[51] V. K. Kapur, D. H. Auckley, S. Chowdhuri, D. C. Kuhlmann, R. Mehra, K. Ramar, and C. G. Harrod, "Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An American academy of sleep medicine clinical practice guideline," pp. 479–504, 9 2017. [Online]. Available: https://thorax.bmj.com/content/70/9/873https://thorax.bmj.com/content/70/9/873.abstract

[52] F. Chouchou and M. Desseilles, "Heart rate variability: A tool to explore the sleeping brain?" p. 402, 12 2014. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2014.00402/abstract

[53] R. Cabiddu, R. Trimer, A. Borghi-Silva, M. Migliorini, R. G. Mendes, A. D. Oliveira, F. S. Costa, and A. M. Bianchi, "Are complexity metrics reliable in assessing HRV control in obese patients during sleep?" *PLoS ONE*, vol. 10, no. 4, p. e0124458, 4 2015. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124458

[54] F. Chouchou and M. Desseilles, "Heart rate variability: A tool to explore the sleeping brain?" p. 402, 12 2014. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2014.00402/abstract

[55] U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: A review," pp. 1031–1051, 12 2006. [Online]. Available: https://link.springer.com/article/10.1007/s11517-006-0119-0

[56] M. Radha, P. Fonseca, M. Ross, A. Cerny, P. Anderer, and R. M. Aarts, "Lstm knowledge transfer for hrv-based sleep staging," *arXiv preprint arXiv:1809.06221*, 2018.

[57] S. Yue, Y. Yang, H. Wang, H. Rahul, and D. Katabi, "BodyCompass: Monitoring Sleep Posture with Wireless Signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–25, 6 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3397311

[58] K. S. Park, S. H. Hwang, D. W. Jung, H. N. Yoon, and W. K. Lee, "Ballistocardiography for nonintrusive sleep structure estimation," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*. Institute of Electrical and Electronics Engineers Inc., 11 2014, pp. 5184–5187.

[59] C.-Y. Hsu, A. Ahuja, S. Yue, R. Hristov, Z. Kabelac, and D. Katabi, "Zero-Effort In-Home Sleep and Insomnia Monitoring using Radio Signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–18, 9 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3130924

[60] S. Khosla, M. C. Deak, D. Gault, C. A. Goldstein, D. Hwang, Y. Kwon, D. O'Hearn, S. Schutte-Rodin, M. Yurcheshen, I. M. Rosen, D. B. Kirsch, R. D. Chervin, K. A. Carden, K. Ramar, R. N. Aurora, D. A. Kristo, R. K. Malhotra, J. L. Martin, E. J. Olson, C. L. Rosen, J. A. Rowley, and American Academy of Sleep Medicine Board of Directors, "Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement." *Journal of Clinical Sleep Medicine*, vol. 14, no. 5, pp. 877–880, 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29734997http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5940440

[61] E. Yuda, Y. Yoshida, R. Sasanabe, H. Tanaka, T. Shiomi, J. Hayano, E. Yuda, Y. Yoshida, R. Sasanabe, H. Tanaka, T. Shiomi, and J. Hayano, "Sleep Stage Classification by a Combination of Actigraphic and Heart Rate Signals," *Journal of Low Power Electronics and Applications*, vol. 7, no. 4, p. 28, 11 2017. [Online]. Available: http://www.mdpi.com/2079-9268/7/4/28

[62] D. F. Dickinson, "The normal ecg in childhood and adolescence," *Heart*, vol. 91, no. 12, pp. 1626–1630, 2005.

[63] T. Taneja, B. Windhagen Mahnert, R. Passman, J. Goldberger, and A. Kadish, "Effects of sex and age on electrocardiographic and cardiac electrophysiological properties in adults," *Pacing and Clinical Electrophysiology*, vol. 24, no. 1, pp. 16–21, 2001.

[64] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, R. M. Lloyd, C. L. Marcus, and B. V. Vaughn, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications: version 2.3*. American Academy of Sleep Medicine, 2016.

[65] Y. Hao and R. Foster, "Wireless body sensor networks for health-monitoring applications," *Physiological Measurement*, vol. 29, no. 11, pp. R27–R56, 2008.

[66] J. W. Shepard, D. J. Buysse, A. L. Chesson, W. C. Dement, R. Goldberg, C. Guilleminault, C. D. Harris, C. Iber, E. Mignot, M. M. Mitler *et al.*, "History of the development of sleep medicine in the united states," *Journal of clinical sleep medicine*, vol. 1, no. 01, pp. 61–82, 2005.

[67] B. Duce, C. Rego, J. Milosavljevic, and C. Hukins, "The aasm recommended and acceptable eeg montages are comparable for the staging of sleep and scoring of eeg arousals," *Journal of Clinical Sleep Medicine*, vol. 10, no. 7, pp. 803–809, 2014.

[68] C.-S. Huang, C.-L. Lin, L.-W. Ko, S.-Y. Liu, T.-P. Su, and C.-T. Lin, "Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels," *Frontiers in neuroscience*, vol. 8, p. 263, 2014.

[69] J. Corral-Peñafiel, J.-L. Pepin, and F. Barbe, "Ambulatory monitoring in the diagnosis and management of obstructive sleep apnoea syndrome," *European Respiratory Review*, vol. 22, no. 129, pp. 312–324, 2013.

[70] V. R. Porter and A. Y. Avidan, "Clinical overview of rem sleep behavior disorder," in *Seminars in neurology*, vol. 37, no. 04. Thieme Medical Publishers, 2017, pp. 461–470.

[71] D. W. Carley and S. S. Farabi, "Physiology of sleep," *Diabetes Spectrum*, vol. 29, no. 1, pp. 5–9, 2016.

[72] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: A conditional adversarial architecture," in *34th International Conference on Machine Learning, ICML 2017*, vol. 8, 2017, pp. 6205–6214. [Online]. Available: http://proceedings.mlr.press/v70/zhao17d.html

[73] Q. Huang, D. Cohen, S. Komarzynski, X.-M. Li, P. Innominato, F. Lévi, and B. Finkenstädt, "Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data." *Journal of the Royal Society, Interface*, vol. 15, no. 139, 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29436510http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5832732

[74] Y. Nam, Y. Kim, and J. Lee, "Sleep Monitoring Based on a Tri-Axial Accelerometer and a Pressure Sensor," *Sensors*, vol. 16, no. 5, p. 750, may 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/5/750

[75] J. Pion-Massicotte, R. Godbout, P. Savard, and J.-F. Roy, "Development and validation of an algorithm for the study of sleep using a biometric shirt in young healthy adults," *Journal of Sleep Research*, p. e12667, feb 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29473243http://doi.wiley.com/10.1111/jsr.12667

[76] B. H. Te Lindert and E. J. Van Someren, "Sleep estimates using microelectromechanical systems (MEMS)," *Sleep*, vol. 36, no. 5, pp. 781–789, 2013. [Online]. Available: https://academic.oup.com/sleep/article-abstract/36/5/781/2559074

[77] V. T. v. Hees, S. Sabia, S. E. Jones, A. R. Wood, K. N. Anderson, M. Kivimaki, T. M. Frayling, A. I. Pack, M. Bucan, D. R. Mazzotti, P. R. Gehrman, A. Singh-Manoux, and M. N. Weedon, "Estimating sleep parameters using an accelerometer without sleep diary," *Scientific Reports*, vol. 8, no. 1, p. 12975, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/02/01/257972

[78] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[79] B. Caulfield, B. Reginatto, and P. Slevin, "Not all sensors are created equal: a framework for evaluating human performance measurement technologies," *npj Digital Medicine*, vol. 2, no. 1, p. 7, dec 2019. [Online]. Available: http://www.nature.com/articles/s41746-019-0082-4

[80] A. Sadeh, "The role and validity of actigraphy in sleep medicine: an update," *Sleep medicine reviews*, vol. 15, no. 4, pp. 259–267, 2011.

[81] J. L. Martin and A. D. Hakim, "Wrist actigraphy." *Chest*, vol. 139, no. 6, pp. 1514–1527, jun 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21652563http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3109647

[82] Y. Moon, R. S. McGinnis, K. Seagers, R. W. Motl, N. Sheth, J. A. W. Jr, R. Ghaffari, and J. J. Sosnoff, "Monitoring gait in multiple sclerosis with novel wearable motion sensors," *PLOS ONE*, vol. 12, no. 2, p. e0171346, 2017.

[83] A. Tal, Z. Shinar, D. Shaki, S. Codish, and A. Goldbart, "Validation of Contact-Free Sleep Monitoring Device with Comparison to Polysomnography," *Journal of Clinical Sleep Medicine*, vol. 13, no. 03, pp. 517–522, mar 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27998378http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5337599http://jcsm.aasm.org/ViewAbstract.aspx?pid=30976

[84] J. Paalasmaa, L. Leppakorpi, and M. Partinen, "Quantifying respiratory variation with force sensor measurements," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2011. IEEE, aug 2011, pp. 3812–3815. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22255170http://ieeexplore.ieee.org/document/6090773/

[85] J. Paalasmaa, H. Toivonen, and M. Partinen, "Adaptive Heartbeat Modeling for Beat-to-Beat Heart Rate Measurement in Ballistocardiograms," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1945–1952, nov 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24691540http://ieeexplore.ieee.org/document/6780577/

[86] M. P. Turakhia, M. Desai, H. Hedlin, A. Rajmane, N. Talati, T. Ferris, S. Desai, D. Nag, M. Patel, P. Kowey *et al.*, "Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The apple heart study," *American heart journal*, vol. 207, pp. 66–75, 2019.

[87] D. Hernando, S. Roca, J. Sancho, A. Alesanco, and R. Bailon, "Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects," *Sensors*, vol. 18, no. 8, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/8/2619

[88] M. Erkinjuntti, K. Vaahtoranta, J. Alihanka, and P. Kero, "Use of the SCSB method for monitoring of respiration, body movements and ballistocardiogram in infants." *Early human development*, vol. 9, no. 2, pp. 119–26, feb 1984. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6714132

[89] P. Chow, G. Nagendra, J. Abisheganaden, and Y. Wang, "Respiratory monitoring using an air-mattress system," *Physiological measurement*, vol. 21, no. 3, p. 345, 2000.

[90] Y. Chee, J. Han, J. Youn, and K. Park, "Air mattress sensor system with balancing tube for unconstrained measurement of respiration and heart beat movements," *Physiological measurement*, vol. 26, no. 4, p. 413, 2005.

[91] P. Arlotto, M. Grimaldi, R. Naeck, and J.-M. Ginoux, "An ultrasonic contactless sensor for breathing monitoring," *Sensors*, vol. 14, no. 8, pp. 15 371–15 386, 2014.

[92] I. Sadek, J. Bellmunt, M. Kodyš, B. Abdulrazak, and M. Mokhtari, "Novel unobtrusive approach for sleep monitoring using fiber optics in an ambient assisted living platform," in *International Conference on Smart Homes and Health Telematics*. Springer, 2017, pp. 48–60.

[93] J. W. Kam, S. Griffin, A. Shen, S. Patel, H. Hinrichs, H.-J. Heinze, L. Y. Deouell, and R. T. Knight, "Systematic comparison between a wireless eeg system with dry electrodes and a wired eeg system with wet electrodes," *NeuroImage*, vol. 184, pp. 119–129, 2019.

[94] P. H. Finan, J. M. Richards, C. E. Gamaldo, D. Han, J. M. Leoutsakos, R. Salas, M. R. Irwin, and M. T. Smith, "Validation of a Wireless, Self-Application, Ambulatory Electroencephalographic Sleep Monitoring Device in Healthy Volunteers," *Journal of Clinical Sleep Medicine*, vol. 12, no. 11, pp. 1443–1451, nov 2016. [Online]. Available: http://jcsm.aasm.org/ViewAbstract.aspx?pid=30850

[95] M. G. Bleichner, M. Lundbeck, M. Selisky, F. Minow, M. Jäger, R. Emkes, S. Debener, and M. De Vos, "Exploring miniaturized EEG electrodes for brain-computer interfaces. An EEG you do not see?" *Physiological Reports*, vol. 3, no. 4, 2015. [Online]. Available: /pmc/articles/PMC4425967//pmc/articles/PMC4425967/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4425967/

[96] D. Looney, P. Kidmose, C. Park, M. Ungstrup, M. Rank, K. Rosenkranz, and D. Mandic, "The in-the-ear recording concept: User-centered and wearable brain monitoring," *IEEE Pulse*, vol. 3, no. 6, pp. 32–42, 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23247157/

[97] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Computers in Biology and Medicine*, vol. 42, no. 12, pp. 1186–1195, dec 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482512001588

[98] K. B. Mikkelsen, J. K. Ebajemito, M. A. Bonmati-Carrion, N. Santhi, V. L. Revell, G. Atzori, C. della Monica, S. Debener, D.-J. Dijk, A. Sterr, and M. de Vos, "Machine-learning-derived sleep-wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy," *Journal of Sleep Research*, p. e12786, nov 2018. [Online]. Available: http://doi.wiley.com/10.1111/jsr.12786

[99] M. TajDini, V. Sokolov, I. Kuzminykh, S. Shiaeles, and B. Ghita, "Wireless sensors for brain activity—a survey," *Electronics*, vol. 9, no. 12, p. 2092, 2020.

[100] T. Nakamura, V. Goverdovsky, M. J. Morrell, and D. P. Mandic, "POINT-OF-CARE TECHNOLOGIES Automatic Sleep Monitoring Using Ear-EEG."

[101] K. B. Mikkelsen, Y. R. Tabar, S. L. Kappel, C. B. Christensen, H. O. Toft, M. C. Hemmsen, M. L. Rank, M. Otto, and P. Kidmose, "Accurate whole-night sleep monitoring with dry-contact ear-EEG," *Scientific Reports*, vol. 9, no. 1, pp. 1–12, 12 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-53115-3

[102] A. Henriksen, M. H. Mikalsen, A. Z. Woldaregay, M. Muzny, G. Hartvigsen, L. A. Hopstock, and S. Grimsgaard, "Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables," *Journal of medical Internet research*, vol. 20, no. 3, p. e110, 2018.

[103] T. Hao, G. Xing, and G. Zhou, "iSleep: unobtrusive sleep quality monitoring using smartphones," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2013, p. 4.

[104] S. Bhat, A. Ferraris, D. Gupta, M. Mozafarian, V. A. DeBari, N. Gushway-Henry, S. P. Gowda, P. G. Polos, M. Rubinstein, H. Seidu, and S. Chokroverty, "Is There a Clinical Role For Smartphone Sleep Apps? Comparison of Sleep Cycle Detection by a Smartphone Application to Polysomnography." *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*, vol. 11, no. 7, pp. 709–15, jul 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25766719http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4481053

[105] S. D. Min, D. J. Yoon, S. W. Yoon, Y. H. Yun, and M. Lee, "A study on a non-contacting respiration signal monitoring system using doppler ultrasound," *Medical & biological engineering & computing*, vol. 45, no. 11, pp. 1113–1119, 2007.

[106] A. Shahshahani, S. Bhadra, and Z. Zilic, "A continuous respiratory monitoring system using ultrasound piezo transducer," in *Circuits and Systems (ISCAS), 2018 IEEE International Symposium on*. IEEE, 2018, pp. 1–4.

[107] T. Rahman, A. T. Adams, R. V. Ravichandran, M. Zhang, S. N. Patel, J. A. Kientz, and T. Choudhury, "DoppleSleep: A Contactless Unobtrusive Sleep Sensing System Using Short-Range Doppler Radar," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 39–50. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2750858.2804280

[108] L. Giovangrandi, O. T. Inan, R. M. Wiard, M. Etemadi, and G. T. Kovacs, "Ballistocardiography - A method worth revisiting," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2011. NIH Public Access, 2011, pp. 4279–4282. [Online]. Available: /pmc/articles/PMC4274997//pmc/articles/PMC4274997/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4274997/

[109] O. J. Kaltiokallio, H. Yigitler, R. Jäntti, and N. Patwari, "Non-invasive respiration rate monitoring using a single cots tx-rx pair," in *Proceedings of the 13th international symposium on Information processing in sensor networks*. IEEE Press, 2014, pp. 59–70.

[110] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart homes that monitor breathing and heart rate," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 2015, pp. 837–846.

[111] A. D. Droitcour, O. Boric-Lubecke, and G. T. Kovacs, "Signal-to-noise ratio in doppler radar system for heart and respiratory rate measurements," *IEEE transactions on microwave theory and techniques*, vol. 57, no. 10, pp. 2498–2507, 2009.

[112] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning Sleep Stages from Radio Signals: A Conditional Adversarial Architecture," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 4100–4109. [Online]. Available: http://sleep.csail.mit.edu/

[113] C.-Y. Hsu, A. Ahuja, S. Yue, R. Hristov, Z. Kabelac, and D. Katabi, "Zero-effort in-home sleep and insomnia monitoring using radio signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 59, 2017.

[114] A. Tataraidze, L. Korostovtseva, L. Anishchenko, M. Bochkarev, Y. Sviryaev, and S. Ivashov, "Bioradiolocation-based sleep stage classification," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 2839–2842.

[115] W. Van Drongelen, *Signal processing for neuroscientists*. Academic press, 2018.

[116] S. R. Devasahayam, *Signals and systems in biomedical engineering: signal processing and physiological systems modeling*. Springer Science & Business Media, 2012.

[117] S. Akselrod, D. Gordon, F. A. Ubel, D. C. Shannon, A. C. Berger, and R. J. Cohen, "Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control," *science*, vol. 213, no. 4504, pp. 220–222, 1981.

[118] T. Penzel, N. Wessel, M. Riedl, J. W. Kantelhardt, S. Rostig, M. Glos, A. Suhrbier, H. Malberg, and I. Fietze, "Cardiovascular and respiratory dynamics during normal and pathological sleep," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 17, no. 1, p. 015116, 2007.

[119] A. Y. Schumann, R. P. Bartsch, T. Penzel, P. C. Ivanov, and J. W. Kantelhardt, "Aging effects on cardiac and respiratory dynamics in healthy subjects across sleep stages," *Sleep*, vol. 33, no. 7, pp. 943–955, 2010.

[120] V. K. Somers, M. E. Dyken, A. L. Mark, and F. M. Abboud, "Sympathetic-nerve activity during sleep in normal subjects," *New England Journal of Medicine*, vol. 328, no. 5, pp. 303–307, 1993.

[121] J. Trinder, J. Kleiman, M. Carrington, S. Smith, S. Breen, N. Tan, and Y. Kim, "Autonomic activity during human sleep as a function of time and sleep stage," *Journal of sleep research*, vol. 10, no. 4, pp. 253–264, 2001.

[122] G. Pocock, C. D. Richards, and D. A. Richards, *Human physiology*. Oxford university press, 2013.

[123] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger *et al.*, "Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology," 1996.

[124] A. Ponnusamy, J. L. B. Marques, and M. Reuber, "Comparison of heart rate variability parameters during complex partial seizures and psychogenic nonepileptic seizures." *Epilepsia*, vol. 53, no. 8, pp. 1314–21, 8 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22642646

[125] A. Malliani, M. Pagani, F. Lombardi, and S. Cerutti, "Cardiovascular neural regulation explored in the frequency domain," *Circulation*, vol. 84, no. 2, pp. 482–492, 1991.

[126] N. Montano, A. Porta, C. Cogliati, G. Costantino, E. Tobaldini, K. R. Casali, and F. Iellamo, "Heart rate variability explored in the frequency domain: A tool to investigate the link between heart and behavior," pp. 71–80, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763408001176

[127] E. Tobaldini, L. Nobili, S. Strada, K. R. Casali, A. Braghiroli, and N. Montano, "Heart rate variability in normal and pathological sleep," *Frontiers in Physiology*, vol. 4, pp. 1–11, 10 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fphys.2013.00294/abstract

[128] F. Snyder, J. A. Hobson, D. F. Morrison, and F. Goldfrank, "Changes in respiration, heart rate, and systolic blood pressure in human sleep," *Journal of Applied Physiology*, vol. 19, no. 3, pp. 417–422, 5 1964. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14174589https://www.physiology.org/doi/10.1152/jappl.1964.19.3.417

[129] D. A. Kirby and R. L. Verrier, "Differential effects of sleep stage on coronary hemodynamic function," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 256, no. 5, pp. H1378–H1383, 5 1989. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/2719135https://www.physiology.org/doi/10.1152/ajpheart.1989.256.5.H1378

[130] R. Cabiddu, S. Cerutti, G. Viardot, S. Werner, and A. M. Bianchi, "Modulation of the sympatho-vagal balance during sleep: Frequency domain study of heart rate variability and respiration," *Frontiers in Physiology*, vol. 3 MAR, 2012.

[131] M. Méndez, A. M. Blanchi, O. Villantieri, and S. Cerutti, "Time-varying analysis of the heart rate variability during REM and non REM sleep stages," in *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 2006, pp. 3576–3579. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4462570/

[132] A. Monti, C. Medigue, H. Nedelcoux, and P. Escourrou, "Autonomic control of the cardiovascular system during sleep in normal subjects," *European Journal of Applied Physiology*, vol. 87, no. 2, pp. 174–181, 6 2002. [Online]. Available: https://link.springer.com/article/10.1007/s00421-002-0597-1

[133] J. P. Saul, R. Rea, D. L. Eckberg, R. D. Berger, and R. J. Cohen, "Heart rate and muscle sympathetic nerve variability during reflex changes of autonomic activity," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 258, no. 3, pp. H713–H721, 1990.

[134] M. Ako, T. Kawara, S. Uchida, S. Miyazaki, K. Nishihara, J. Mukai, K. Hirao, J. Ako, and Y. Okubo, "Correlation between electroencephalography and heart rate variability during sleep," *Psychiatry and clinical neurosciences*, vol. 57, no. 1, pp. 59–65, 2003.

[135] A. Baharav, S. Kotagal, V. Gibbons, B. Rubin, G. Pratt, J. Karin, and S. Akselrod, "Fluctuations in autonomic nervous activity during sleep displayed by power spectrum analysis of heart rate variability," *Neurology*, vol. 45, no. 6, pp. 1183–1187, 1995.

[136] E. Vanoli, P. B. Adamson, Ba-Lin, G. D. Pinna, R. Lazzara, and W. C. Orr, "Heart Rate Variability During Specific Sleep Stages," *Circulation*, vol. 91, no. 7, pp. 1918–1922, 4 1995. [Online]. Available: https://www.ahajournals.org/doi/10.1161/01.CIR.91.7.1918

[137] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, p. 44, 2019.

[138] J. S. Bauer, S. Consolvo, B. Greenstein, J. Schooler, E. Wu, N. F. Watson, and J. A. Kientz, "ShutEye: Encouraging Awareness of Healthy Sleep Recommendations with a Mobile, Peripheral Display," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, ACM. ACM Press, 2012, pp. 1401–1410. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2207676.2208600

[139] Y. K. Choi, G. Demiris, S.-Y. Lin, S. J. Iribarren, C. A. Landis, H. J. Thompson, S. M. McCurry, M. M. Heitkemper, and T. M. Ward, "Smartphone applications to support sleep self-management: review and evaluation," *Journal of Clinical Sleep Medicine*, vol. 14, no. 10, pp. 1783–1790, 2018.

[140] A. Malafeev, D. Laptev, S. Bauer, X. Omlin, A. Wierzbicka, A. Wichniak, W. Jernajczyk, R. Riener, J. M. Buhmann, and P. Achermann, "Automatic Human Sleep Stage Scoring Using Deep Neural Networks," *Frontiers in Neuroscience*, vol. 12, p. 781, nov 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00781/full

[141] J. Shawe-Taylor, N. Cristianini *et al.*, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[142] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[143] G. Wahba *et al.*, "Support vector machines, reproducing kernel hilbert spaces and the randomized gacv," *Advances in Kernel Methods-Support Vector Learning*, vol. 6, pp. 69–87, 1999.

[144] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, and K. Jerbi, "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines," *Journal of Neuroscience Methods*, vol. 250, pp. 94–105, jul 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027015000230

[145] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[146] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier," *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 10–19, 2012.

[147] M. Xiao, H. Yan, J. Song, Y. Yang, and X. Yang, "Sleep stages classification based on heart rate variability and random forest," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 624–633, 2013.

[148] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 9 2015. [Online]. Available: http://www.robots.ox.ac.uk/

[149] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep Audio-visual Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[150] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[151] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[152] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2015-Janua, 2015, pp. 3995–4001.

[153] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, 2019.

[154] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, P. Kidmose, and M. De Vos, "Personalized automatic sleep staging with single-night data: A pilot study with Kullback-Leibler divergence regularization," *Physiological Measurement*, vol. 41, no. 6, p. 064004, 6 2020.

[155] H. Phan, O. Y. Chen, P. Koch, Z. Lu, I. McLoughlin, A. Mertins, and M. De Vos, "Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 6, pp. 1787–1798, 6 2021.

[156] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[157] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, no. 1-2, pp. 305–307, 6 2018.

[158] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December. IEEE Computer Society, 12 2016, pp. 770–778.

[159] Y. LeCun *et al.*, "Generalization and network design strategies," *Connectionism in perspective*, vol. 19, pp. 143–155, 1989.

[160] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[161] X. Zhang, W. Kou, I. Eric, C. Chang, H. Gao, Y. Fan, and Y. Xu, "Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device," *Computers in biology and medicine*, vol. 103, pp. 71–81, 2018.

[162] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2015.

[163] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 855–868, 2008.

[164] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[165] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*. PMLR, 2013, pp. 1310–1318.

[166] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," *Advances in neural information processing systems*, vol. 9, 1996.

[167] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[168] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 12 2014.

[169] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "DIVA: Domain invariant variational autoencoder," in *Deep Generative Models for Highly Structured Data, DGS@ICLR 2019 Workshop*, 3 2019.

[170] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "B-VAE: Learning basic visual concepts with a constrained variational framework," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 11 2017.

[171] ——, "beta-vae: Learning basic visual concepts with a constrained variational framework," *Proc. ICLR*, 2016.

[172] G. Wilson and D. J. Cook, "A Survey of Unsupervised Deep Domain Adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, 7 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3400066

[173] J. Williams and S. King, "Disentangling style factors from speaker representations," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe. International Speech Communication Association, 2019, pp. 3945–3949.

[174] G. Wilson, J. R. Doppa, and D. J. Cook, "Multi-Source Deep Domain Adaptation with Weak Supervision for Time-Series Sensor Data," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1768–1778. [Online]. Available: https://doi.org/10.1145/3394486.3403228

[175] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, 2017. [Online]. Available: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/building-machines-that-learn-and-think-like-people/A9535B1D745A0377E16C590E14B94993

[176] L. Zhao, Y. Wang, J. Zhao, L. Yuan, J. J. Sun, F. Schroff, H. Adam, X. Peng, D. Metaxas, and T. Liu, "Learning View-Disentangled Human Pose Representation by Contrastive Cross-View Mutual Information Maximization," pp. 12 793–12 802, 2020. [Online]. Available: http://arxiv.org/abs/2012.01405

[177] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua. Institute of Electrical and Electronics Engineers Inc., 11 2017, pp. 1283–1292.

[178] E. Fuster-Garcia, A. Bresó, J. Martínez-Miranda, J. Rosell-Ferrer, C. Matheson, and J. M. García-Gómez, "Fusing actigraphy signals for outpatient monitoring," *Information Fusion*, vol. 23, pp. 69–80, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S156625351400089X

[179] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December. IEEE Computer Society, 12 2016, pp. 4651–4659.

[180] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December. IEEE Computer Society, 12 2016, pp. 21–29.

[181] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th ACM International Conference on Multimedia, MM 2005*. New York, New York, USA: ACM Press, 2005, pp. 677–682. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1101149.1101300

[182] S. Liu, S. Yao, J. Li, D. Liu, T. Wang, H. Shao, and T. Abdelzaher, "GlobalFusion: A global attentional deep learning framework for multisensor information fusion," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–27, 3 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3380999

[183] Y. Bai, Y. Guan, and W. F. Ng, "Fatigue assessment using ECG and actigraphy sensors," in *Proceedings - International Symposium on Wearable Computers, ISWC*. Association for Computing Machinery, 9 2020, pp. 12–16. [Online]. Available: https://doi.org/10.1145/3410531.3414308

[184] Y. Chen, W. Dong, Y. Gao, X. Liu, and T. Gu, "Rapid," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–27, 9 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3130906

[185] A. Rattani, D. R. Kisku, M. Bicego, and M. Tistarelli, "Feature level fusion of face and fingerprint biometrics," in *IEEE Conference on Biometrics: Theory, Applications and Systems, BTAS'07*, 2007.

[186] Y. Guan and T. Plotz, "Ensembles of deep LSTM learners for activity recognition using wearables," pp. 1–28, 3 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3090076

[187] Y. Guan, C. T. Li, and F. Roli, "On Reducing the Effect of Covariate Factors in Gait Recognition: A Classifier Ensemble Method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1521–1528, 7 2015.

[188] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications," *IEEE Journal on Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 3 2020.

[189] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 8 2016.

[190] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10111 LNCS. Springer Verlag, 11 2017, pp. 213–228. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-54181-5_14

[191] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," pp. 423–443, 2 2019.

[192] V. Radu, C. Tong, S. Bhattacharya, N. D. Lane, C. Mascolo, M. K. Marina, and F. Kawsar, "Multimodal Deep Learning for Activity and Context Recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–27, 2018.

[193] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October. Institute of Electrical and Electronics Engineers Inc., 12 2017, pp. 1839–1848.

[194] J. Hayano, E. Yuda, and Y. Yoshida, "Sleep stage classification by combination of actigraphic and heart rate signals," *2017 IEEE International Conference on Consumer Electronics - Taiwan, ICCE-TW 2017*, pp. 387–388, 2017.

[195] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[196] T. Munzner, *Visualization Analysis and Design*, ser. A.K. Peters visualization series. A K Peters, 2014. [Online]. Available: http://www.cs.ubc.ca/%7Etmm/vadbook/

[197] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *ArXiv*, vol. abs/1802.03888, pp. 1–9, 2018.

[198] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[199] J. Girschik, L. Fritschi, J. Heyworth, and F. Waters, "Validation of self-reported sleep against actigraphy." *Journal of epidemiology*, vol. 22, no. 5, pp. 462–8, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22850546http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3798642

[200] A. Sadeh, K. M. Sharkey, and M. A. Carskadon, "Activity-Based Sleep—Wake Identification: An Empirical Test of Methodological Issues," *Sleep*, vol. 17, no. 3, pp. 201–207, 1994.

[201] A. Sadeh, P. J. Hauri, D. F. Kripke, and P. Lavie, "The role of actigraphy in the evaluation of sleep disorders." *Sleep*, vol. 18, no. 4, pp. 288–302, 5 1995. [Online]. Available: https://academic.oup.com/sleep/article-abstract/18/4/288/2749735http://www.ncbi.nlm.nih.gov/pubmed/7618029

[202] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity." *Sleep*, vol. 15, no. 5, pp. 461–9, 10 1992. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1455130

[203] E. Sazonov, N. Sazonova, S. Schuckers, M. Neuman, and CHIME Study Group, "Activity-based sleep-wake identification in infants." *Physiological measurement*, vol. 25, no. 5, pp. 1291–304, 10 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15535193

[204] J.-M. Lee, W. Byun, A. Keill, D. Dinkel, and Y. Seo, "Comparison of Wearable Trackers' Ability to Estimate Sleep." *International journal of environmental research and public health*, vol. 15, no. 6, p. 1265, 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29914050http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6025478

[205] J. Palotti, R. Mall, M. Aupetit, M. Rueschman, M. Singh, A. Sathyanarayana, S. Taheri, and L. Fernandez-Luque, "Benchmark on a large cohort for sleep-wake classification with machine learning techniques," *npj Digital Medicine*, vol. 2, no. 1, p. 50, dec 2019. [Online]. Available: http://www.nature.com/articles/s41746-019-0126-9

[206] M. Aktaruzzaman, M. Migliorini, M. Tenhunen, S. L. Himanen, A. M. Bianchi, and R. Sassi, "The addition of entropy-based regularity parameters improves sleep stage classification based on heart rate variability," *Medical and Biological Engineering and Computing*, vol. 53, no. 5, pp. 415–425, 5 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25690323http://link.springer.com/10.1007/s11517-015-1249-z

[207] M. H. Bonnet and D. L. Arand, "Heart rate variability: sleep stage, time of night, and arousal influences," *Electroencephalography and Clinical Neurophysiology*, vol. 102, no. 5, pp. 390–396, 5 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921884X96960701

[208] S. Elsenbruch, M. J. Harnish, and W. C. Orr, "Heart Rate Variability During Waking and Sleep in Healthy Males and Females," *Sleep*, vol. 22, no. 8, pp. 1067–1071, 12 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10617167https://academic.oup.com/sleep/article-lookup/doi/10.1093/sleep/22.8.1067

[209] N. Daskalova, B. Lee, J. Huang, C. Ni, and J. Lundin, "Investigating the Effectiveness of Cohort-Based Sleep Recommendations," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–19, 9 2018.

[210] D. A. Dean, A. L. Goldberger, R. Mueller, M. Kim, M. Rueschman, D. Mobley, S. S. Sahoo, C. P. Jayapandian, L. Cui, M. G. Morrical, S. Surovec, G.-Q. Zhang, and S. Redline, "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 5 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27070134http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4835314https://academic.oup.com/sleep/article-lookup/doi/10.5665/sleep.5774

[211] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The National Sleep Research Resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 10 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29860441http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6188513https://academic.oup.com/jamia/article/25/10/1351/5026200

[212] M. Hornyak, M. Cejnar, M. Elam, M. Matousek, and B. G. Wallin, "Sympathetic muscle nerve activity during sleep in man." *Brain*, vol. 114 (Pt 3, no. 3, pp. 1281–95, 6 1991. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/2065250https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/114.3.1281

[213] V. K. Somers, M. E. Dyken, A. L. Mark, and F. M. Abboud, "Sympathetic-Nerve Activity during Sleep in Normal Subjects," *New England Journal of Medicine*, vol. 328, no. 5, pp. 303–307, 2 1993. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8419815http://www.nejm.org/doi/abs/10.1056/NEJM199302043280502

[214] P. Boudreau, W.-H. Yeh, G. A. Dumont, and D. B. Boivin, "Circadian Variation of Heart Rate Variability Across Sleep Stages," *Sleep*, vol. 36, no. 12, pp. 1919–1928, 12 2013. [Online]. Available: https://academic.oup.com/sleep/article/36/12/1919/2709417

[215] M. Radha, P. Fonseca, A. Moreau, M. Ross, A. Cerny, P. Anderer, X. Long, and R. M. Aarts, "Sleep stage classification from heart-rate variability using long short-term memory neural networks," *Scientific Reports*, vol. 9, no. 1, pp. 1–11, 12 2019.

[216] U. Singh, S. Chauhan, A. Krishnamachari, and L. Vig, "Ensemble of deep long short term memory networks for labelling origin of replication sequences," in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Paris, France: IEEE, 2015, pp. 1–7.

[217] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Computers in Biology and Medicine*, vol. 42, no. 12, pp. 1186–1195, 12 2012.

[218] E. Alickovic and A. Subasi, "Ensemble SVM Method for Automatic Sleep Stage Classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1258–1265, jun 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8292946/

[219] J. B. Stephansen, A. N. Olesen, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi,

D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. D. Sorensen, P. Jennum, and E. Mignot, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Communications*, vol. 9, no. 1, p. 5229, 12 2018. [Online]. Available: http://www.nature.com/articles/s41467-018-07229-3

[220] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 8 2004.

[221] D. E. Bild, D. A. Bluemke, G. L. Burke, R. Detrano, A. V. Diez Roux, A. R. Folsom, P. Greenland, D. R. Jacobs, R. Kronmal, K. Liu, J. C. Nelson, D. O'Leary, M. F. Saad, S. Shea, M. Szklo, and R. P. Tracy, "Multi-Ethnic Study of Atherosclerosis: Objectives and design," *American Journal of Epidemiology*, vol. 156, no. 9, pp. 871–881, 11 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12397006https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwf113

[222] A. Varri, B. Kemp, T. Penzel, and A. Schlogl, "Standards for biomedical signal databases," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 33–37, 2001.

[223] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *Journal of the American College of Cardiology*, vol. 37, no. 1, pp. 153–156, 2001.

[224] M. Malik, "Heart Rate Variability." *Annals of Noninvasive Electrocardiology*, vol. 1, no. 2, pp. 151–181, 4 1996. [Online]. Available: http://doi.wiley.com/10.1111/j.1542-474X.1996.tb00275.x

[225] J. Tilmanne, J. Urbain, M. V. Kothare, A. V. Wouwer, and S. V. Kothare, "Algorithms for sleep-wake identification using actigraphy: A comparative study and new results," *Journal of Sleep Research*, vol. 18, no. 1, pp. 85–98, 3 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19250177

[226] D. F. Kripke, E. K. Hahn, A. P. Grizas, K. H. Wadiak, R. T. Loving, J. S. Poceta, F. F. Shadan, J. W. Cronin, and L. E. Kline, "Wrist actigraphic scoring for sleep laboratory patients: Algorithm development," *Journal of Sleep Research*, vol. 19, no. 4, pp. 612–619, 12 2010.

[227] S. R. Patel, J. Weng, M. Rueschman, K. A. Dudley, J. S. Loredo, Y. Mossavar-Rahmani, M. Ramirez, A. R. Ramos, K. Reid, A. N. Seiger, D. Sotres-Alvarez, P. C. Zee, and R. Wang, "Reproducibility of a Standardized Actigraphy Scoring Algorithm for Sleep in a US Hispanic/Latino Population," *Sleep*, vol. 38, no. 9, pp. 1497–1503, 9 2015.

[228] B. M. Altevogt, H. R. Colten *et al.*, *Sleep disorders and sleep deprivation: an unmet public health problem*. Washington (DC): National Academies Press, 2006.

[229] S. Bagaveyev and D. J. Cook, "Designing and Evaluating Active Learning Methods for Activity Recognition," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ser. UbiComp '14 Adjunct. New York, NY, USA: Association for Computing Machinery, 2014, p. 469–478. [Online]. Available: https://doi.org/10.1145/2638728.2641674

[230] H. Guo, L. Chen, L. Peng, and G. Chen, "Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1112–1123. [Online]. Available: https://doi.org/10.1145/2971648.2971708

[231] J. Palotti, R. Mall, M. Aupetit, M. Rueschman, M. Singh, A. Sathyanarayana, S. Taheri, and L. Fernandez-Luque, "Benchmark on a large cohort for sleep-wake classification with machine learning techniques," *npj Digital Medicine*, vol. 2, no. 1, p. 50, 2019.

[232] S. Haghayegh, S. Khoshnevis, M. H. Smolensky, K. R. Diller, and R. J. Castriotta, "Deep Neural Network Sleep Scoring Using Combined Motion and Heart Rate Variability Data," *Sensors*, vol. 21, no. 1, p. 25, 12 2020. [Online]. Available: https://www.mdpi.com/1424-8220/21/1/25

[233] K.-T. CHOU, Y.-T. CHANG, Y.-M. CHEN, K.-C. SU, D.-W. PERNG, S.-C. CHANG, and G.-M. SHIAO, "The minimum period of polysomnography required to confirm a diagnosis of severe obstructive sleep apnoea," *Respirology*, vol. 16, no. 7, pp. 1096–1102, 2011.

[234] L. Menghini, E. Gianfranchi, N. Cellini, E. Patron, M. Tagliabue, and M. Sarlo, "Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions," *Psychophysiology*, vol. 56, no. 11, p. e13441, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/psyp.13441

[235] E. A. Thomson, K. Nuss, A. Comstock, S. Reinwald, S. Blake, R. E. Pimentel, B. L. Tracy, and K. Li, "Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities," *Journal of Sports Sciences*, vol. 37, no. 12, pp. 1411–1419, 6 2019. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/02640414.2018.1560644

[236] N. E. Fultz, G. Bonmassar, K. Setsompop, R. A. Stickgold, B. R. Rosen, J. R. Polimeni, and L. D. Lewis, "Coupled electrophysiological, hemodynamic, and cerebrospinal fluid oscillations in human sleep." *Science (New York, N.Y.)*, vol. 366, no. 6465, pp. 628–631, 11 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/31672896

[237] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in neural information processing systems*. Long Beach, CA, United States: Curran Associates, 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[238] C. Eom and B. Ham, "Learning disentangled representation for robust person re-identification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[239] S. Park, S. Hwang, D. Kim, and H. Byun, "Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2403–2411.

[240] B. Hu, Y. Guan, Y. Gao, Y. Long, N. Lane, and T. Ploetz, "Robust cross-view gait recognition with evidence: A discriminant gait gan (diggan) approach," *arXiv preprint arXiv:1811.10493*, 2018.

[241] C. Qin, Y. Zhang, Y. Liu, S. Coleman, D. Kerr, and G. Lv, "Appearance-invariant place recognition by adversarially learning disentangled representation," *Robotics and Autonomous Systems*, vol. 131, p. 103561, 2020.

[242] S. Sun, H. Shi, and Y. Wu, "A Survey of Multi-Source Domain Adaptation," *Inf. Fusion*, vol. 24, no. C, p. 84–92, 7 2015. [Online]. Available: https://doi.org/10.1016/j.inffus.2014.12.003

[243] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[244] A. Mazankiewicz, K. Bohm, and M. Berges, "Incremental Real-Time Personalization in Human Activity Recognition Using Domain Adaptive Batch Normalization," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, 12 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3432230

[245] N. Banluesombatkul, P. Ouppaphan, P. Leelaarporn, P. Lakhan, B. Chaitusaney, N. Jaimchariyatam, E. Chuangsuwanich, W. Chen, H. Phan, N. Dilokthanakul, and T. Wilaiprasitporn, "MetaSleepLearner: A Pilot Study on Fast Adaptation of Bio-Signals-Based Sleep Stage Classifier to New Individual Subject Using Meta-Learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 1949–1963, 6 2021.

[246] X. Qin, Y. Chen, J. Wang, and C. Yu, "Cross-Dataset Activity Recognition via Adaptive Spatial-Temporal Transfer Learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 4, 12 2019. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3369818

[247] E. M. Lima, A. H. Ribeiro, G. M. Paixão, M. H. Ribeiro, M. M. Pinto-Filho, P. R. Gomes, D. M. Oliveira, E. C. Sabino, B. B. Duncan, L. Giatti, S. M. Barreto, W. Meira, T. B. Schön, and A. L. P. Ribeiro, "Deep neural network-estimated electrocardiographic age as a mortality predictor," *Nature Communications*, vol. 12, no. 1, pp. 1–10, 8 2021. [Online]. Available: https://www.nature.com/articles/s41467-021-25351-7

[248] Z. I. Attia, P. A. Friedman, P. A. Noseworthy, F. Lopez-Jimenez, D. J. Ladewig, G. Satam, P. A. Pellikka, T. M. Munger, S. J. Asirvatham, C. G. Scott, R. E. Carter, and S. Kapa, "Age and Sex Estimation Using Artificial Intelligence from Standard 12-Lead ECGs," *Circulation: Arrhythmia and Electrophysiology*, vol. 12, no. 9, 9 2019.

[249] C. Lombardi, M. F. Pengo, and G. Parati, "Obstructive sleep apnea syndrome and autonomic dysfunction," p. 102563, 11 2019.

[250] Y. Wang, K. Hu, K. Liu, Z. Li, J. Yang, Y. Dong, M. Nie, J. Chen, Y. Ruan, and J. Kang, "Obstructive sleep apnea exacerbates airway inflammation in patients with chronic obstructive pulmonary disease," *Sleep Medicine*, vol. 16, no. 9, pp. 1123–1130, 9 2015.

[251] H. Chen, F. Zhuang, L. Xiao, L. Ma, H. Liu, R. Zhang, H. Jiang, and Q. He, "AMA-GCN: Adaptive Multi-layer Aggregation Graph Convolutional Network for Disease Prediction." International Joint Conferences on Artificial Intelligence, 8 2021, pp. 2235–2241.

[252] H. Qian, S. J. Pan, and C. Miao, "Latent Independent Excitation for Generalizable Sensor-based Cross-Person Activity Recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, pp. 11 921–11 929, 5 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17416

[253] E. Vanoli, P. B. Adamson, Ba-Lin, G. D. Pinna, R. Lazzara, and W. C. Orr, "Heart Rate Variability During Specific Sleep Stages," *Circulation*, vol. 91, no. 7, pp. 1918–1922, 4 1995. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/01.cir.91.7.1918

[254] S. Elsenbruch, M. J. Harnish, and W. C. Orr, "Heart rate variability during waking and sleep in healthy males and females," *Sleep*, vol. 22, no. 8, pp. 1067–1071, 12 1999. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10617167/

[255] T. Penzel, J. W. Kantelhardt, L. Grote, J. H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 10, pp. 1143–1151, 10 2003.

[256] A. O. S. A. T. F. of the American Academy of Sleep Medicine, "Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults," *Journal of clinical sleep medicine*, vol. 5, no. 3, pp. 263–276, 2009.

[257] C. Guilleminault, R. Winkle, S. Connolly, K. Melvin, and A. Tilkian, "Cyclical variation of the heart rate in sleep apnoea syndrome. Mechanisms, and usefulness of 24 h electrocardiography as a screening technique," *Lancet (London, England)*, vol. 1, no. 8369, pp. 126–131, 1 1984. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/6140442/

[258] U. Wiklund, B. O. Olofsson, K. Franklin, H. Blom, P. Bjerle, and U. Niklasson, "Autonomic cardiovascular regulation in patients with obstructive sleep apnoea: a study based on spectral analysis of heart rate variability," *Clinical physiology (Oxford, England)*, vol. 20, no. 3, pp. 234–241, 5 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10792417/

[259] M. A. Lips, G. H. De Groot, M. De Kam, F. J. Berends, R. Wiezer, B. A. Van Wagensveld, D. J. Swank, A. Luijten, H. Pijl, and J. Burggraaf, "Autonomic nervous system activity in diabetic and healthy obese female subjects and the effect of distinct weight loss strategies," *European Journal of Endocrinology*, vol. 169, no. 4, pp. 383–390, 10 2013. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23847327/

[260] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2096–2030, 1 2016.

[261] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4893–4902.

[262] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International conference on machine learning*. PMLR, 2017, pp. 2208–2217.

[263] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[264] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[265] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[266] L. Bai, L. Yao, X. Wang, S. S. Kanhere, B. Guo, and Z. Yu, "Adversarial Multi-View Networks for Activity Recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 2, 6 2020. [Online]. Available: https://doi.org/10.1145/3397323

[267] O. Ozdenizci, Y. Wang, T. Koike-Akino, and D. Erdogmus, "Adversarial Deep Learning in EEG Biometrics," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 710–714, 5 2019.

[268] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled Speech Embeddings Using Cross-Modal Self-Supervision," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2020-May. Institute of Electrical and Electronics Engineers Inc., 5 2020, pp. 6829–6833.

[269] G. Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The National Sleep Research Resource: Towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 10 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29860441/

[270] S. Tufik, R. Santos-Silva, J. A. Taddei, and L. R. A. Bittencourt, "Obstructive Sleep Apnea Syndrome in the Sao Paulo Epidemiologic Sleep Study," *Sleep Medicine*, vol. 11, no. 5, pp. 441–446, 5 2010.

[271] X. Chen, R. Wang, P. Zee, P. L. Lutsey, S. Javaheri, C. Alcántara, C. L. Jackson, M. A. Williams, and S. Redline, "Racial/ethnic differences in sleep disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA)," *Sleep*, vol. 38, no. 6, pp. 877–888, 6 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25409106/

[272] K. M. Flegal, B. K. Kit, H. Orpana, and B. I. Graubard, "Association of All-Cause Mortality With Overweight and Obesity Using Standard Body Mass Index Categories: A Systematic Review and Meta-analysis," *JAMA*, vol. 309, no. 1, pp. 71–82, 1 2013. [Online]. Available: https://jamanetwork.com/journals/jama/fullarticle/1555137

[273] Y. Kwon, S. A. Gharib, M. L. Biggs, D. R. Jacobs, A. Alonso, D. Duprez, J. Lima, G. M. Lin, E. Z. Soliman, R. Mehra, S. Redline, and S. R. Heckbert, "Association of sleep characteristics with atrial fibrillation: the Multi-Ethnic Study of Atherosclerosis," *Thorax*, vol. 70, no. 9, pp. 873–879, 9 2015. [Online]. Available: https://thorax.bmj.com/content/70/9/873https://thorax.bmj.com/content/70/9/873.abstract

[274] B. A. Mander, J. R. Winer, and M. P. Walker, "Sleep and Human Aging," pp. 19–36, 4 2017.

[275] N. Sridhar, A. Shoeb, P. Stephens, A. Kharbouch, D. B. Shimol, J. Burkart, A. Ghoreyshi, and L. Myers, "Deep learning for automated sleep staging using instantaneous heart rate," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10, 12 2020. [Online]. Available: https://doi.org/10.1038/s41746-020-0291-x

[276] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in Neural Information Processing Systems*. Neural information processing systems foundation, 2016, pp. 289–297.

[277] P. Cortelli and C. Lombardi, "Sleep and autonomic nervous system dysfunction," in *Handbook of clinical neurophysiology*. Elsevier, 2005, vol. 6, pp. 343–353.

[278] S. Khosla, M. C. Deak, D. Gault, C. A. Goldstein, D. Hwang, Y. Kwon, D. O'Hearn, S. Schutte-Rodin, M. Yurcheshen, I. M. Rosen, D. B. Kirsch, R. D. Chervin, K. A. Carden, K. Ramar, R. Nisha Aurora, D. A. Kristo, R. K. Malhotra, J. L. Martin, E. J. Olson, C. L. Rosen, and J. A. Rowley, "Consumer sleep technology: An American academy of sleep medicine position statement," *Journal of Clinical Sleep Medicine*, vol. 14, no. 5, pp. 877–880, 5 2018. [Online]. Available: /pmc/articles/PMC5940440/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5940440/

[279] P. Fonseca, N. Den Teuling, X. Long, and R. M. Aarts, "A comparison of probabilistic classifiers for sleep stage classification," *Physiological Measurement*, vol. 39, no. 5, 2018.

[280] ——, "Cardiorespiratory Sleep Stage Detection Using Conditional Random Fields," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 4, pp. 956–966, 2017.

[281] J. Palotti, R. Mall, M. Aupetit, M. Rueschman, M. Singh, A. Sathyanarayana, S. Taheri, and L. Fernandez-Luque, "Benchmark on a large cohort for sleep-wake classification with machine learning techniques," *npj Digital Medicine*, vol. 2, no. 1, pp. 1–9, 12 2019. [Online]. Available: https://doi.org/10.1038/s41746-019-0126-9

[282] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics (ACL), 2016, pp. 457–468.

[283] H. Duan, S. Wang, and Y. Guan, "SOFA-Net: Second-Order and First-order Attention Network for Crowd Counting," 8 2020. [Online]. Available: https://arxiv.org/abs/2008.03723v1

[284] T. Y. Lin, A. Roychowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, 2015, pp. 1449–1457. [Online]. Available: http://vis-www.cs.umass.edu/bcnn

[285] O. Walch, "Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography v1.0.0," 2019. [Online]. Available: https://physionet.org/content/sleep-accel/1.0.0/

[286] E. Tobaldini, L. Nobili, S. Strada, K. R. Casali, A. Braghiroli, and N. Montano, "Heart rate variability in normal and pathological sleep," *Frontiers in Physiology*, vol. 4, pp. 1–11, 10 2013. [Online]. Available: www.frontiersin.org

[287] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[288] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS. Springer Verlag, 2016, pp. 630–645. [Online]. Available: https://github.com/KaimingHe/resnet-1k-layers.

[289] A. E. Orhan and X. Pitkow, "Skip Connections Eliminate Singularities," *arXiv*, 1 2017. [Online]. Available: http://arxiv.org/abs/1701.09175

[290] M. V. Perez, K. W. Mahaffey, H. Hedlin, J. S. Rumsfeld, A. Garcia, T. Ferris, V. Balasubramanian, A. M. Russo, A. Rajmane, L. Cheung, G. Hung, J. Lee, P. Kowey, N. Talati, D. Nag, S. E. Gummidipundi, A. Beatty, M. T. Hills, S. Desai, C. B. Granger, M. Desai, and M. P. Turakhia, "Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation," *New England Journal of Medicine*, vol. 381, no. 20, pp. 1909–1917, 11 2019. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMoa1901183

[291] H. Wilson, G. Dervenoulas, G. Pagano, C. Koros, T. Yousaf, M. Picillo, S. Polychronis, A. Simitsi, B. Giordano, Z. Chappell *et al.*, "Serotonergic pathology and disease burden in the premotor and motor phase of a53t α-synuclein parkinsonism: a cross-sectional study," *The Lancet Neurology*, vol. 18, no. 8, pp. 748–759, 2019.

[292] K. L. Lunetta, "Genetic association studies," *Circulation*, vol. 118, no. 1, pp. 96–101, 2008.

[293] L. R. Cardon and J. I. Bell, "Association study designs for complex diseases," *Nature Reviews Genetics*, vol. 2, no. 2, pp. 91–99, 2001.

[294] N. Risch and J. Teng, "The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases i. dna pooling," *Genome research*, vol. 8, no. 12, pp. 1273–1288, 1998.

[295] D. Gordon, S. J. Finch, M. Nothnagel, and J. Ott, "Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms," *Human heredity*, vol. 54, no. 1, pp. 22–33, 2002.

[296] J. A. Hobson, R. W. McCarley, and P. W. Wyzinski, "Sleep cycle oscillation: Reciprocal discharge by two brainstem neuronal groups," *Science*, vol. 189, no. 4196, pp. 55–58, 1975. [Online]. Available: http://www.jstor.org/stable/1740806

[297] Z. Zhou, Y. Zhang, X. Yu, P. Yang, X.-Y. Li, J. Zhao, and H. Zhou, "Xhar: Deep domain adaptation for human activity recognition with smart devices," in *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2020, pp. 1–9.

[298] Y. Chang, A. Mathur, A. Isopoussu, J. Song, and F. Kawsar, "A systematic study of unsupervised domain adaptation for robust human-activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–30, 2020.