MACHINE LEARNING
& DATA ANALYTICS

FAU
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Benchmarking of Sleep/Wake Detection Algorithms using Wearable Sensors and Machine Learning

Master's Thesis in Medical Engineering

submitted
by

Daniel Krauß

born 02.10.1996 in Schwäbisch Hall

Written at

Machine Learning and Data Analytics Lab
Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Advisors: Robert Richer, M.Sc., Arne Küderle, M.Sc., Prof. Dr. Bjoern Eskofier
(*Machine Learning and Data Analytics Lab*, FAU Erlangen-Nürnberg)
Prof. Dr. Nicolas Rohleder (*Chair of Health Psychology*, FAU Erlangen-Nürnberg)

Started: 15.06.2021

Finished: 15.12.2021

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 15. Dezember 2021

Übersicht

Schlaf ist eine wichtige physiologische Funktion, die sich nicht nur auf eine Vielzahl täglicher Aktivitäten wie Lernen, Produktivität und Aufmerksamkeit auswirkt, sondern auch mit zahlreichen Krankheiten wie Bluthochdruck, Schlaganfall und Herzerkrankungen in Verbindung gebracht wird. Daher ist eine genaue Schlafüberwachung für viele Anwendungsszenarien in Medizin und Psychologie von entscheidender Bedeutung. Da Schlaflabore sehr kosten- und ressourcenintensiv sind, stellen tragbare Sensoren eine vielversprechende Alternative für eine nicht-invasive Schlafüberwachung zu Hause dar. Ziel dieser Arbeit war es, mehrere Machine Learning- und Deep Learning-Algorithmen systematisch mit etablierten, heuristischen Algorithmen auf einem Benchmark-Datensatz zu vergleichen. Außerdem wurde untersucht, ob zusätzliche Biosignale, wie Herzratenvariabilität, die Klassifizierung verbessern können. Die Ergebnisse zeigen, dass ein multimodaler Ansatz, der Bewegungs- und Herzinformationen kombiniert, die beiden monomodalen Ansätze übertrifft. Die beste Klassifizierung wurde für ein multimodales LSTM gefunden, das eine Genauigkeit von $83.7 \pm 9.6\%$ erreichte, während die machine learning Algorithmen etwas schlechter abschnitten. Die heuristischen Algorithmen schnitten am schlechtesten ab.
Um diese Algorithmen in einer realen Umgebung zu bewerten, wurde ein neuer Datensatz mit IMU- und EKG-Daten und einer klinisch validierten Schlafmatte als Referenz von 42 Teilnehmern (85 Nächte) aufgenommen. Anhand dieses Datensatzes wurde untersucht, ob die Verwendung von inertialen Messeinheiten (**Inertial Measurement Unit (IMU)**) anstelle von aggregierten Aktigraphiedaten die Schlaf/Wach-Erkennung weiter verbessern kann. Dabei übertraf der ***IMU***-basierte Ansatz mit ***XGBoost (XGB)***-Klassifikator den auf Aktigraphie basierenden Ansatz und erreichte ein Cohen's κ von 0.59 ± 0.27 (vs. $\kappa = 0.38 \pm 0.24$). Außerdem wurde die Erkennungsrate von ***IMU***- und ***Heart Rate Variability (HRV)***-basierter Schlaf/Wach-Erkennung verglichen, wobei der multimodale Ansatz im Vergleich zum monomodalen, bewegungsbasierten Ansatz nicht von zusätzlichen kardialen Informationen profitieren konnte. Da teilnehmerspezifische Einflüsse auf die Klassifizierungsgenauigkeit angenommen wurden, wurden verschiedene demografische Merkmale und Eigenschaften untersucht, die statistisch signifikante ($p < 0.05$) Einflüsse auf die Erkennungsrate ergaben, beispielsweise für das Geschlecht und die Aufnahmequalität.
Die Ergebnisse zeigen, dass die Verwendung von ***IMU***-Rohdaten anstelle der Aktigraphie für eine bessere Schlaf-/Wach-Erkennung mit tragbaren Sensoren vorteilhaft sein könnte. Jedoch müssen weitere Forschungen durchgeführt werden, um mehr Daten aus der realen Welt zu sammeln, die zum Vergleich verschiedener Algorithmen zur Schlaf-/Wach-erkennung in einem realistischeren Szenario verwendet werden können. Des Weiteren sollte der Nutzen weiterer Biosignale, wie die aus EKG-Daten, oder der Bewegung des Brustkorbs extrahierte Atmung, untersucht werden.

Abstract

Sleep is an important physiological function that does not only affect a variety of daily activities like learning, productivity, and attention, but is also linked to multiple diseases such as hypertension, strokes, and heart disorders. For that reason, accurate sleep monitoring is crucial for many application scenarios in medicine and psychology. As sleep laboratories are very cost- and resource-intensive, wearable sensors are a promising alternative for unobtrusive sleep monitoring at home. The aim of this work was to systematically compare several state-of-the-art machine and deep learning algorithms with traditional heuristic algorithms on a large benchmark dataset that was collected in a controlled laboratory environment. It was further assessed if additional data modalities, such as ***HRV***, are able to boost the classification. The results demonstrate that a multimodal approach combining movement and cardiac information outperforms both monomodal approaches. The best classification performance was found for a multimodal LSTM, which achieved $83.7 \pm 9.6\%$ accuracy, while the machine learning algorithms performed slightly worse. The heuristic algorithms performed worst.

To evaluate these algorithms in a real-world setting, a new sleep dataset containing IMU and ECG data and data from a clinically validated sleep mat as ground truth of 42 participants (85 nights) was collected. Using this dataset, it was examined whether using raw ***IMU*** data instead of aggregated actigraphy data can further improve sleep/wake classification performance. Thereby, the ***IMU***-based sleep/wake detection using ***XGB*** outperformed the actigraphy-based approach reaching a Cohen's κ of 0.59 ± 0.27 (vs. $\kappa = 0.38 \pm 0.24$). Furthermore, the performance of ***IMU***- and ***Electrocardiogram (ECG)***-based sleep/wake detection were compared in mono- and multimodal approaches. Here, the multimodal approach was not able to benefit from additional cardiac information compared to the monomodal, motion-based approach. Because subject-specific influences on classification performance were hypothesized, different demographics and characteristics were examined and yielded statistically significant ($p < 0.05$) influences on classification performance, for instance, for gender and recording quality.

The results of this thesis show that the usage of raw ***IMU*** data instead of actigraphy might be advantageous for better sleep/wake detection using wearable sensors. However, further research needs to be conducted by collecting more real-world data resulting in larger datasets that can be used to benchmark different sleep/wake detection algorithms in a more realistic scenario to obtain more generalizable results. Furthermore, the benefit of adding further unobtrusive biosignals, such as respiration, extracted from ***ECG*** data or chest movement, should be examined.

Contents

Chapter 1

Introduction

Sleep is an important physiological function that affects a variety of daily activities like learning, productivity, attention, or memorizing [Cho10]. Insufficient sleep is directly linked to a series of diseases like diabetes or hypertension and causes a higher risk of strokes and heart disorders [Ban07, Cho10]. For that reason, sleep monitoring is crucial for many application scenarios in medicine and psychology because it can help to identify the causes for sleep disorders, thus enabling to initiate adequate therapies. The gold standard approach for sleep monitoring and the detection of sleep disorders is ***Polysomnography (PSG)***, which is typically performed in a sleep laboratory [Ban07, Che20b]. In a ***PSG*** examination, different physiological signals like ***Electroencephalogram (EEG)***, ***ECG***, pulse, and respiration are assessed during sleep, as well as body position and muscle activities of limbs [Bar02]. Usually, ***PSG*** recordings are divided into 30 s or 1 min epochs which are then labeled by a trained professional that assigns a sleep or wake stage to each epoch. The classification of sleep stages can follow different conventions, but the most common classification includes five different stages: Wakefulness, ***Rapid Eye Movement (REM)***, and three different categories of ***Non-Rapid Eye Movement (NREM)***, known as N1, N2, and N3 [Daf18]. The high precision of ***PSG*** allows a good and reliable diagnosis. However, it also suffers from several drawbacks. For instance, longitudinal measurements are not feasible since ***PSG*** is very cost- and resource-intensive. Furthermore, the unfamiliar laboratory setting can influence the sleep quality of patients [Ibe04].

In contrast, sleep diaries are inexpensive and easy to acquire for large-scale datasets. However, sleep diaries lack accuracy because of imprecise data during the night, poor awareness of exact sleep and wake times, as well as a subject-specific bias [Ber97].

A promising alternative to the gold standard method for sleep/wake detection is provided by wearable sensors. The advantage of this approach is that wearable sensors experience wide popularity since they are broadly accepted in the population, unobtrusive, and low-cost. The advantages enable longitudinal studies, thus potentially allowing to develop better diagnostic approaches. Moreover, individuals can follow their regular daily habits and, most importantly sleep in their own bed, which makes sleep monitoring in a more realistic setting possible.

During sleep, body movements decrease compared to a wakeful state [WF83]. Hence, assessing human activity is a promising candidate for unobtrusive sleep detection. An established approach to assess body activity is via actigraphy [Sad11]. Actigraphy is an accelerometer-based aggregation of movement in time windows of 30 s or 1 min. In medical applications, Actigraphy is used since the 1950s. Since then it rapidly developed into a valuable asset for sleep medicine clinicians [Mar11].

Algorithms for actigraphy-based sleep/wake detection have been developed over the past four decades. In the beginning, most algorithms were of heuristic nature, defining activity rules when individuals are likely to be sleeping or not. In the past years, the rise of machine learning and deep learning techniques has also led to researchers developing more advanced, data-driven algorithms for sleep/wake detection.

An alternative to perform movement-based sleep/wake detection is to use ***IMU*** sensors. In contrast to Actigraphy, ***IMU*** sensors are higher sampled and offer different sensing modalities like acceleration and angular velocity. This opens up the potential for finer detection of movements, and thus, for a more precise classification [Bor14, Pal19, Che20b].

However, not only sleep but also cardiac activity changes during sleep. During ***NREM***-phases, blood pressure decreases by 10 % concurrently with a lower heart rate [Sil13]. Therefore, combining movement and cardiac information allows to perform sleep staging, i.e., identifying different sleep stages, instead of solely detecting sleep and wake phases [Pal19, Che20b, Zha20, Hag21].

In order to determine the best set of algorithms for sleep/wake detection, it is important to systematically compare different algorithms with different input modalities on a large-scale, diverse, and publicly available dataset. This ensures comparability between different approaches and avoids the risk of study-specific influences. One dataset allowing the benchmarking of different algorithms and different input modalities is the ***Multi-Ethnic Study of Atherosclerosis (MESA)***. ***MESA*** was a multi-centric collaborative longitudinal investigation of factors associated with the development of subclinical cardiovascular disease between 2000 and 2012. It included 6,814 men and women of different age and ethnicities. Between 2010 and 2012, 2,237 participants of this study were also enrolled in a sleep study, containing actigraphy and heart rate data with synchronized full overnight unattended ***PSG*** recordings as well as sleep diaries [Che15, Zha18].

However, the dataset only contains actigraphy data and no raw ***IMU*** data, thus allowing no conclusions on whether raw ***IMU*** data might increase sleep/wake detection accuracy.
The goal of this master's thesis is therefore to implement state-of-the-art sleep/wake detection algorithms of different types and input modalities and compare these to newly developed algorithms using the "MESA Sleep" benchmark dataset [Che15, Zha18]. Furthermore, this thesis will evaluate the developed algorithms on real-world sleep data by collecting a new dataset with wearable ***IMU*** and ***ECG*** sensors while using a clinically validated sleep mat (Withings Sleep Analyzer, Issy-les-Moulineaux, France) as reference. Concurrently, this thesis will identify whether sleep/wake detection performance can be further improved by different input modalities, e.g, by using features computed from raw ***IMU*** data instead of aggregated actigraph data. Furthermore, the performance of ***IMU***- and ***ECG***-based sleep/wake detection will be compared in mono- and multimodal approaches.
The structure of this thesis is organized as follows: Chapter 2 presents relevant work of recent research in the field of sleep/wake detection, while Chapter 3 outlines the medical background necessary for this thesis. This Chapter includes especially the assessment of body parameters during sleep as well as the connection of sleep disorders with widespread diseases. In Chapter 4, the technical fundamentals are presented. Chapter 5 introduces the MESA Dataset, as well as all methods which were used to benchmark the algorithms. Chapter 6 focuses on the study conducted to acquire real-world data and its processing. The evaluation metrics are presented in Chapter 7, followed by the results in Chapter 8. The discussion of the outcomes of this thesis will be presented in Chapter 9 while the conclusion and outlook conclude this thesis in Chapter 10.

Chapter 2

Related Work

Over the past decades, many researchers worked in the field of sleep/wake detection. To get a good overview of the state-of-the-art in research, it is practical to group the publications according to the algorithm type and the input modalities that were used to predict sleep and wake states.
Grouping sleep/wake detection algorithms according to their algorithm type yields three categories: heuristic, rule-based algorithms, traditional machine learning-based algorithms, and deep learning-based algorithms. Prominent heuristic algorithms have been developed by Webster et al. [Web82], Kripke et al. [Kri10], and Cole et al. [Col92]. These threshold-based algorithms work with actigraph data convolved with a windowed kernel to gain time dependency. Another rule-based algorithm for distinguishing sleep and wake phases was developed by Sadeh et al. [Sad94]. In their paper they presented an approach to classify sleep and wakefulness using wrist-worn actigraphy and compared the results of the device worn on the dominant and non-dominant hand. Sazonov et al. [Saz02, Saz04] published an accelerometer-based algorithm to detect sleep/wake states of infants as well as their current position in the crib. The estimation of sleep/wake is based on logistic regression and was evaluated against ***PSG***. These traditional, rule-based approaches produced promising results with accuracies greater than 80%. Except of Sazonov's work, all presented heuristic algorithms suffer from a considerable overprediction of sleep that is expressed in high precisions of over 90% and low recalls of about 70%.
However, these heuristic algorithms are static and can, therefore, barely be adapted to different subject-specific sleep patterns. To get a better representation of different sleep characteristics, many researchers applied classical machine learning with multiple input features to sleep data. Tillmanne et al. compared an ***Artificial Neural Network (ANN)*** and decision trees with the approaches of Sazonov et al. and Sadeh et al. in a study of 354 young children who wore ankle actimeters. Their results show that ***ANN***s and decision trees are able to improve the quality

of sleep/wake estimation [Til09]. Another publication of Orellana et al. aimed to address the problem of overestimating sleep by focusing on balancing sensitivity and specificity, rather than accuracy, with an ***ANN*** and a balanced dataset [Ore14]. To address this imbalance, Domingues et al. [Dom14] also developed an approach that combines two linear discriminant classifiers refined by a ***Hidden Markov Model (HMM)***. To overcome situations where ***PSG*** is hard to acquire and supervised methods are difficult to train, Li et al. developed an unsupervised sleep/wake detection approach using ***HMM***. They trained an ***HMM*** algorithm that was tested on 43 individuals and compared the results to the Actiwatch and its proprietary algorithm, as well as to the ***University of California San Diego (UCSD)*** algorithm that is of heuristic nature. The results showed that they were able to improve the Cohen's κ statistic to 0.446 compared to the Actiwatch algorithm (0.399) and the UCSD algorithm (0.311) [Li20].

One recent innovative approach to further improve sleep/wake detection performance is to use ensemble deep learning with accelerometer and ***HRV*** data. Chen et al. [Che20b] proposed a ***Local Feature-Based LSTM (LF-LSTM)*** network that was able to outperform all benchmark approaches. Another work by the same research group proposes the implementation of a multimodal attention-based Convolutional-Neural-Network-***Long Short-Term Memory (LSTM)*** approach using accelerometer and ***HRV*** data [Che20a]. To consider the distinct contributions of both modalities, the attention-based algorithm dynamically adjusts the feature importance of both sensors and enables the full usage of all information. With this particular network architecture, the authors were able to outperform all benchmark approaches including the ***LF-LSTM*** that was presented in their preceding publication [Che20a, Che20b]. Although the results of the presented deep learning approaches are promising, they were just validated against a sleep-monitor headband as well as a MotionWatch, but not validated against the gold standard ***PSG***. In contrast, the work of Haghayegh et al. presents a ***PSG*** validated comparison of different deep learning approaches including ***Convolutional Neural Network (CNN)*** and Deep Convolutional ***LSTM***. The best model they examined was the ***CNN*** with batch normalization [Hag20]. Li et al. published a sleep stage algorithm that is based on ***ECG*** measurements and validated against ***PSG***. They trained a ***CNN*** with ***ECG***-derived respiration and ***HRV*** signals and achieved results of 75.4% accuracy which is one of the highest non-electroencephalographic outcomes in sleep staging [Li18].

Another way to group sleep/wake detection algorithms is to divide them according to their input modalities. A recent systematic review of different wearable sensing technologies for sleep-stage and sleep/wake detection was published by Imtiaz et al. [Imt21]. Based on a review of 90 papers, they identified 13 different input modalities used in clinical and home monitoring environments. Due to the highest agreement rates, the most common sensing modality is ***EEG***. However, in the specific case of home monitoring, the most prominent input modalities are motion and ***HRV*** data. Lauteslager et al. developed a radar-based sleep staging system that extracts both body movement and respiratory features to distinguish different sleep phases. They found that their approach performed better in terms of accuracy than two commercially available wrist-worn sleep monitoring devices [Lau20]. A non-contact biomotion sensor for the detection of sleep/wake patterns was used by De Chazal et al. [DC11]. The principle of this approach was to measure the reflection of radio waves to detect body and respiratory movements. The major advantage of that approach is that it is completely unobtrusive and contactless since no sensors are worn directly on the body. The results obtained with the sensor system are comparable with the performance of actigraph-based approaches although they were not validated against each other in the same study. Sano et al. presented a multimodal sleep/wake detection approach that combined features of biosignals with phone usage statistics. They found that time was one of the most important features, though irregular naps and unusual day routines degraded the performance [San19].

One of the most widely used input modality in home monitoring is actigraphy [Imt21]. Jean-Louis et al. compared the performance of sleep estimation using movement data acquired by two different actigraphs. In their study, both actigraphs worked comparably, suggesting that the usage of different actigraph devices does not have a major impact [JL01]. Because three-axial, linear accelerometry is sampled at a high rate, sleep estimation reliability might improve in comparison to low-sampled actigraphy which cannot exploit full movement information. To provide ongoing studies with backward comparability, te Lindert et al. examined if it is feasible to acquire movement data via high sampled three-axial accelerometry and transform it into traditional activity counts. They found a good agreement between the estimated and the measured movement counts. Furthermore, they stated that the usage of accelerometers avoids the risk of brand-specific incompatibility and showed that the agreement of two accelerometers is higher than the agreement of two actigraphs [tL13].

Another widely spread input modality is cardiac-based. According to the review of Itmaz et al., most of the cardiac data is acquired via ***Photoplethysmography (PPG)*** [Imt21]. A sleep stage detection system based on ***ECG*** signal measurements was proposed by Widasari et al. The decision-tree-based Support Vector Machine that only used features of the ***HRV*** spectrum gained a remarkable accuracy of 89.2% [Wid18]. Lewicke et al. compared the prediction performance of a Learning Vector Quantization (LVQ) neural network using only actigraphy or only ***HRV***. They found better overall agreement using actigraphy, however, being awake was better recognized with the heart-based approach [Lew04].

Due to the widespread usage with high prediction rates, it is a promising approach to combine these two modalities. Devot et al. investigated the achievement of adding cardiac and respiratory information to movement recordings. They stated that the multimodal approach may improve the classification result for sleep/wake classification [Dev10]. De Zambotti et al. examined the performance of commercially available fitness trackers that estimate sleep, using activity as well as ***HRV*** in different works. Their results were validated against ***PSG***. The authors found a high sensitivity for detecting sleep, but a low prediction rate of wake epochs during the night [dZ15, dZ16]. Haghayegh et al. evaluated their recently presented algorithm with different combinations of input modalities including activity counts as well as ***HRV*** features in different epoch lengths. They found that the combination of ***HRV*** and actigraphy improves the performance compared to the monomodal approaches. They further stated that a less granular epoch length leads to better agreement rates [Hag21].

Even though the presented results gain high prediction rates, one of the major problems in sleep/wake detection research is the limited comparability between different works due to small studies and differing study conditions. For instance, it is way easier to reach high agreement rates in longitudinal studies that include day and night, rather than only night. Due to considerably increased movement in daily activities, epochs of wakefulness over the day are easy to predict and gain a higher overall accuracy. Furthermore, a difference in study-specific parameters such as epoch length or device can have a large impact on results. For this reason, it is of particular importance to systematically benchmark different algorithms and different input modalities on a large standardized dataset.

Hence, Palotti et al. [Pal19] compared a large quantity of state-of-the-art heuristic, machine learning, and deep learning algorithms for sleep/wake detection on the ***MESA*** Sleep dataset which contains more than 2,200 subjects. They also compared different study designs, including an all-day as well as a night-only approach. Their results show that the all-day approach achieved a better result in all metrics due to its high estimation performance during the day. Furthermore, the deep learning approaches had the highest accuracy, followed by the machine learning algorithms, whereas the traditional heuristic algorithms performed the worst. When evaluated only for nighttime, all algorithms except from Sazonov overestimated sleep, while the ***LSTM*** network performed best [Pal19]. Zhai et al. implemented different state-of-the-art machine- and deep learning algorithms to estimate sleep stages in different subdivisions using ***HRV*** and actigraphy in single- and multimodal approaches on the same benchmark dataset as Palotti et al. They found that the estimation performance decreased from 84.4% for sleep/wake detection to 63.7% for sleep stage prediction, i.e., attempting to distinguish wake, ***REM*** as well as N1-3 stages. For sleep/wake detection, they found no improvement between the actigraphy-only and multimodal approaches, however adding ***HRV*** features considerably improved classification accuracy for sleep staging. The monomodal approach using only ***HRV*** features performed worse in all tasks [Zha20].

Chapter 3

Medical Background

3.1 Physiology of Sleep

For a long time, sleep was considered to be a simple passive state to recover for the next day. Since the second half of the 20th century, it has become clear that sleep is a highly complex physical state involving a huge amount of brain activity. Sleep quantity as well as sleep quality considerably contribute to our physical and mental well-being and thus have a major impact on our quality of life [Sta05].
The underlying mechanism of sleep is determined by genetics. For that, it is different for each individual but stable from one night to another. This mechanism is regulated by cardiac rhythm and the intensity of brain activity [PH13].

Sleep is no homogeneous process [Tra14]. According to the ***American Academy of Sleep Medicine (AASM)***, sleep can be divided into two fundamental types: ***REM***, which is associated with active dreaming, and ***NREM*** that can be further divided into the three stages N1-N3, whereas N1 describes light sleep with conjugate reasonably regular sinusoidal eye movements and a low amplitude of brainwaves from 4 to 7 Hz. To classify sleep as N2, one or more trains of sleep spindles have to occur. A sleep spindle is a train of brain waves with a frequency from 11 to 16 Hz and a duration of more than 0.5 s. N3, which is also refereed as the deep sleep phase, is characterized by slow brain waves of 0.5 to 2 Hz [Ibe07, Mos09]. Throughout the different sleep stages, a lot of body functions are diminished. Especially ***NREM*** sleep phases are characterized by a decrease in heart rate, blood pressure, breathing rate, and body core temperature [Mos09, Cho10, PH13, Sin15, Car16].
The actual gold standard to assess sleep is ***PSG***. It consists of ***EEG*** to assess brain activity, ***Electromyography (EMG)*** to measure muscle activity of the limbs, ***Electrooculography (EOG)***

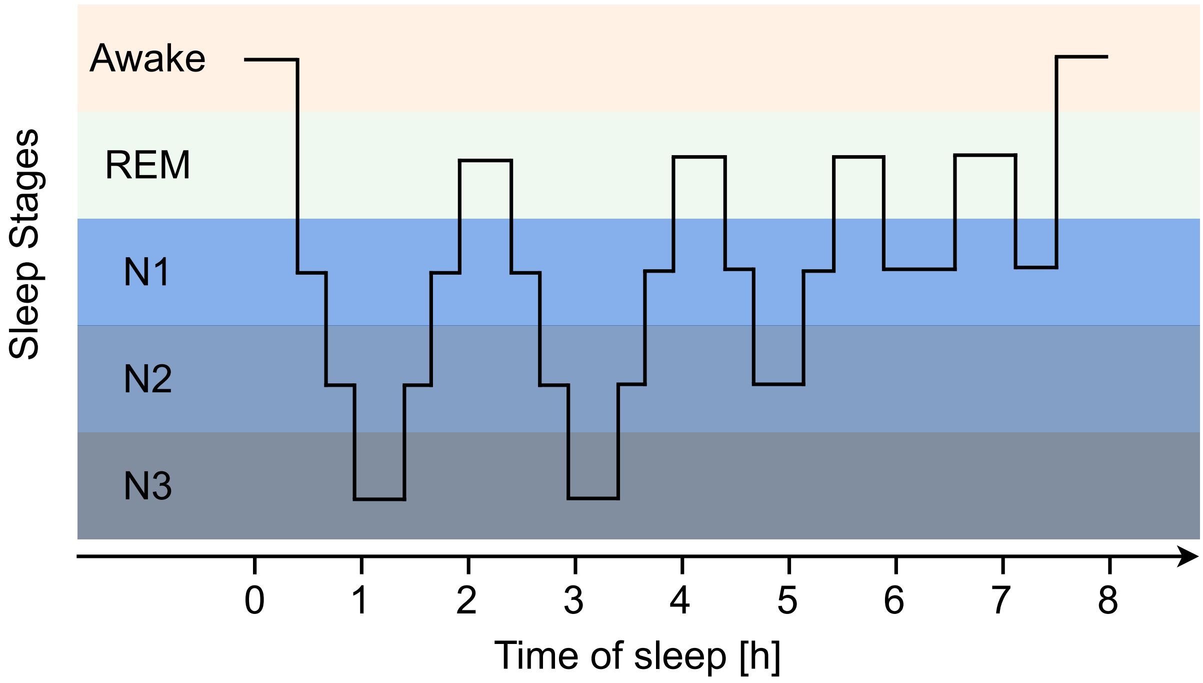

Figure 3.1: Example hypnogram of a healthy adult.

to track eye-movement activity as well as ***ECG*** to examine changes in cardiac activity. Further parameters are pulse oximetry asessed by a finger or ear clip as well as respiration measures acquired by a stretch belt [Bar02, Sta05]. Typically, ***PSG*** recordings are staged in 30s epochs according to the ***AASM*** criteria [Sta05, Mos09, Car16]. The sleep process over night is usually visualized in a hypnogram, which is a graph that represents sleep stages as a function of time. Figure 3.1 depicts a typical hypnogram of a heatly adult.

In a normal night, the sleep process consists of repetitive, but slightly changing sleep cycles, and lasts about 8 hours. Typically, there is a deep N3 sleep phase of about one hour shortly after falling asleep, followed by alternating phases of ***NREM*** and ***REM*** sleep at $60 - 90$ min intervals throughout the rest of the night [Car16] [Sta05]. Usually, most of the deep N3 phases occur in the first part of the night, whereas ***REM*** sleep predominates the second part of the night [Car16]. Overall, ***REM*** sleep accounts for $20 - 25\%$ of the night, while ***NREM*** sleep represents $75 - 80\%$ of the sleep phases [Sta05].

3.2 Sleep Disorders

It is well known that sufficient sleep is essential and sleep deprivation resulting either from lifestyle or from sleep disorders causes short- and long-term health consequences [Cho10]. Short-term effects of sleep deprivation are present in a reduced ability to focus or to solve complex tasks, lower productivity, and thus reduced quality of life. Long-term consequences include a higher morbidity and mortality as well as a higher risks for several diseases like obesity, diabetes, strokes, and coronary artery diseases [Ban07, Cho10, PH13]. While lifestyle-induced sleep deprivation can be easily remedied, sleep disorders are highly complex diseases. Several epidemiological studies showed that sleep problems are common in the population [Sin15]. According to a report of the National Center of Sleep Disorders Research, 35% of the population in the USA have problems falling asleep, maintaining sleep, and suffer from awakening too early in the morning [oSDRU93]. Thereby, the four most common sleep problems in society are excessive daytime somnolence, insomnia, abnormal movements or abnormal behaviors during sleep, and an inability to sleep at the desired time [Cho10, Sin15].

However, most of the sleep problems are not relevant from an acute medical perspective as they need to be treated by a physician. To assess patients with serve sleep disturbances, physical examinations are undertaken, followed by an evaluation of treatment history and laboratory tests [Cho10]. Furthermore, it is of particular importance to study the family disease history, because several sleep diseases have a genetic component [Seh11]. Symptoms that occur when the patient is lying in the bed may suggest a diagnosis of ***Restless legs syndrome (RLS)***, a widespread movement-induced sleep disturbance with a strong genetic component [Seh11]. ***RLS*** is associated with sensory symptoms like unpleasant crawling, burning, aching, or itching sensations that mostly occur between the knees and ankles. As the symptoms are typically noted when patients are sitting and lying, these conditions have a major impact on sleep initiation although sleep interruptions can be a problem, too [Cho10, Sin15].

The most common sleep disorder in medical treatment is ***Insomnia (INS)***. Individuals suffering from ***INS*** mostly experience a lack of sleep time, triggered by difficulties to initiate and maintain sleep as well as early-morning awakenings. Acute ***INS*** may be associated with acute stress, but commonly, insomnia cases are chronic [Cho10].

To assess sleep diseases and their severities, laboratory tests need to be conducted. The most important laboratory tests are an overnight ***PSG***, as described in Section 3.1, and ***Multiple Sleep Latency Test (MSLT)***. ***MSLT*** defines a sequence of sleep latency tests to provide an objective measure of daytime sleepiness [Cho10, Sin15]. For instance, the presence of two ***REM*** phases shortly after falling asleep or a sleep onset latency of less than 8 min suggests a diagnosis of

narcolepsy, that is associated with excessive sleepiness at daytime [Cho10, Sin15].
The ***Apnea–hypopnea Index (AHI)*** is a commonly used score to assess the severity and frequency of apneas and hypneas. The scores usually depend on the airflow and oxygen desaturation, however, there are no unique criteria to define hypopneas in the ***AHI*** [Man01]. The definition used in this work is provided[1].
Additionally to this techniques, it may be useful to use standardized questionnaires about usual sleep behaviour to assess sleep quality as well as sleep disturbances longitudinally [Sin15]. Commonly used questionnaires for that are the ***Pittsburgh Sleep Quality Index (PSQI)*** for adults and the ***Pediatric Daytime Sleepiness Scale (PDSC)*** that assesses daytime sleepiness in children [Buy89, Dra03, Sin15]. In this work, the ***PSQI*** was used, providing values from 0-21, with ***PSQI*** values above five indicating poor sleep quality. The ***MESA*** dataset, used in this work employed the ***Women's Health Initiative Insomnia Rating Scale (WHIIRS)*** questionnaire, which is a five-item scale evaluating ***INS*** symptoms. This score ranges from 0 to 20, with scores above nine indicating a ***INS*** diagnosis [Lev03].

[1] https://sleepdata.org/datasets/mesa/variables/ahi_a0h4a [Resc]

Chapter 4

Fundamentals

4.1 Actigraphy

Within the last decades, the technical characteristics of batteries, sensors, and digital data storage developed, and thereby also the usage of wearable devices became increasingly popular since data quality, battery runtime and on-device storage drastically increased [LT19]. This opened up the possibility to transfer sleep medicine partially into the patients' homes [Mar11]. A major advantage of this development is that wearable devices are more reliable and have less subjective information compared to sleep questionnaires.

Figure 4.1 shows a movement recording via wrist-worn actigraphy, acquired with concurrent ***PSG***. As depicted, high activity counts correlate with wake phases which opens the possibility to estimate sleep/wake states from movement data.

The general working principle of actigraphy is to aggregate physical movement that is sampled several times per second in epochs of 30 s or 1 min, whereas sampling- and epoch rates can be set by the investigator. As illustrated in Figure 4.2, the three commonly used methods to extract activity counts are ***Time Above Threshold (TAT)***, ***Zero Crossing Mode (ZCM)***, and ***Digital Integration Mode (DIM)***. The ***TAT*** method assesses the time when the movement is above a certain threshold, whereas the ***ZCM*** method accounts the number of times the signal crosses zero. However, both methods neither take amplitude nor acceleration into account which potentially cause high-frequency artifacts to be counted as movement. In contrast, ***DIM*** integrates over the highly sampled input signal and thereby considers both the amplitude and acceleration, but not the duration or frequency of the input signal. For that reason, some actigraphs use more than one method to cover the weakness of one single method [Sto17].

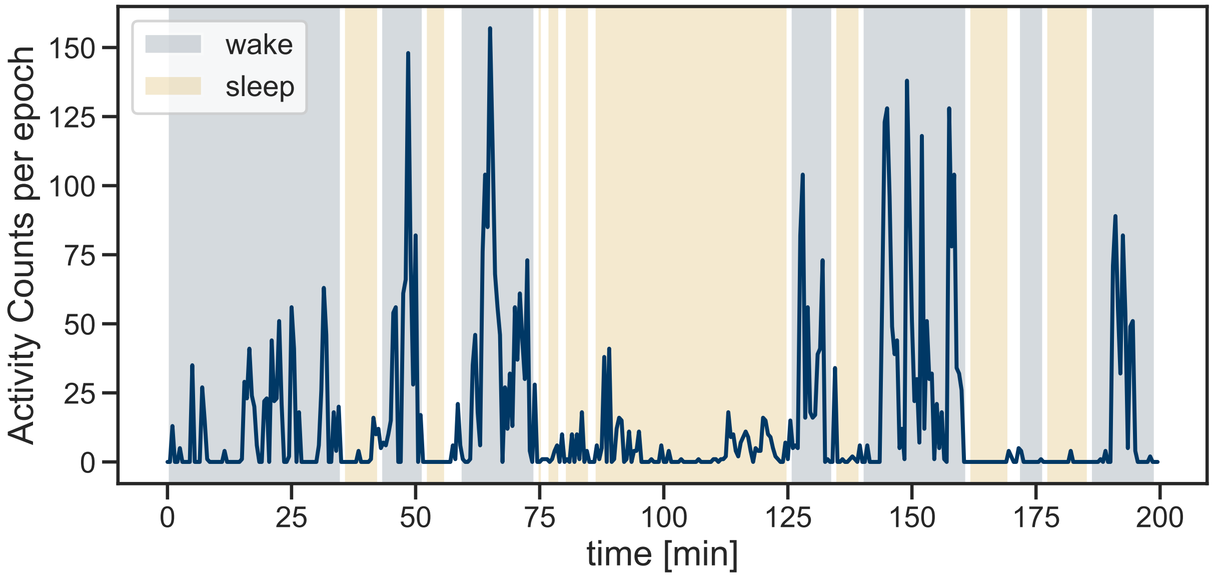

Figure 4.1: Activity counts with concurrent ***PSG*** acquisition.

To record movements during sleep actigraph devices can be worn on wrist, chest, or other body positions in an unobtrusive way. This modality is used frequently in sleep medicine because it provides information about the sleep habits of individuals in their natural environment [Mar11]. In 2007, a report of the ***AASM*** supported the usage of actigraphy for clinical applications and particularly for the diagnosis and evaluation of insomnia, hypersomnia, and obstructive sleep apnea [Mor07]. However, actigraphy has some limitations in sleep medicine. When assessing sleep, actigraphy tends to overestimate sleep due to a lack of movement in short periods of wakefulness. Moreover, different sleep phases cannot be distinguished by movement alone [Mar11]. Therefore, multimodal approaches including ***ECG*** or ***EEG*** may offer a better performance [Imt21]. Another major drawback is the limited comparability due to many manufacturers and closed-book algorithms [Sad11].

To allow researchers to compare the results of accelerometer-based studies with ActiGraph-based studies, Brønd et al. [Brø17] developed an algorithm that converts raw accelerometer signals into activity counts for the most widely used device ActiGraph. To ensure compatibility with the Nyquist-Shannon sampling theorem, the authors applied an aliasing filter to the raw signal. In a second step, a frequency band-pass filter was applied. The coefficients of the filter were extracted from the frequency response, measured from the original ActiGraph. To filter high and low frequencies, the data was truncated at 2.13 g and convoluted with a dead-band filter of 0.068 g. After conversion into 8-bit resolution, the activity counts were accumulated in all window sizes,

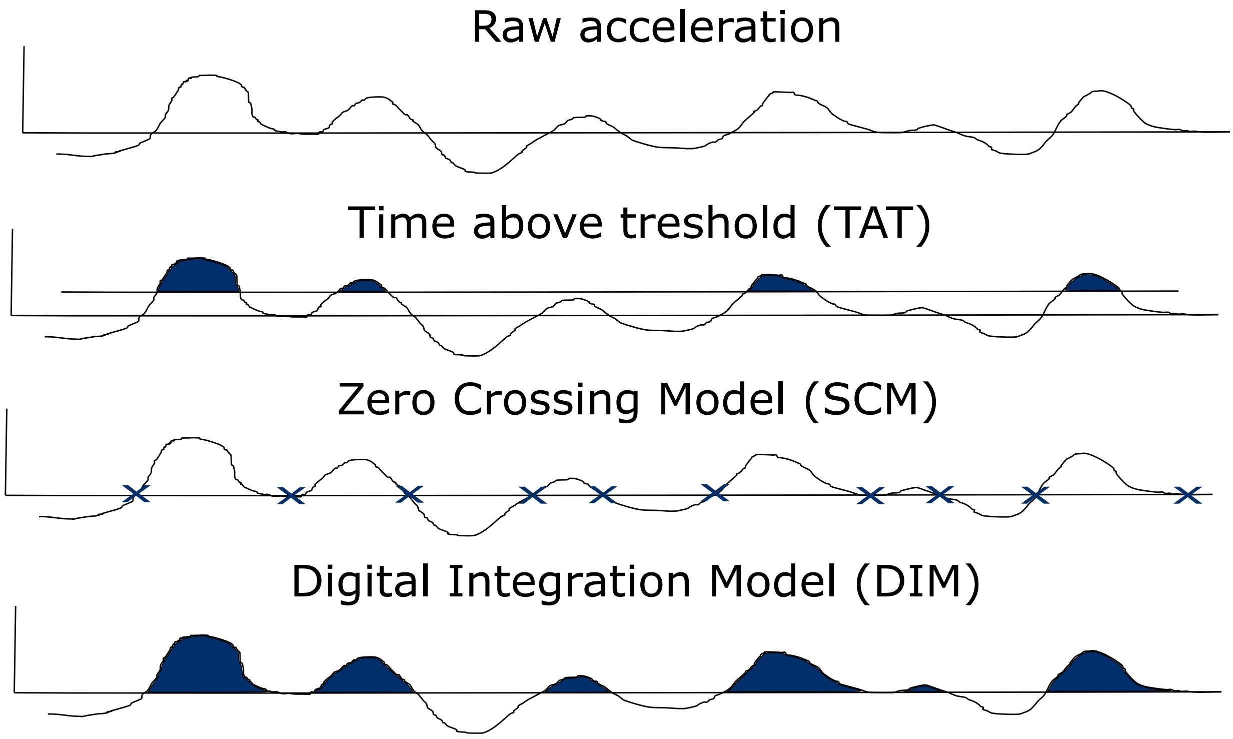
Raw acceleration
Time above treshold (TAT)
Zero Crossing Model (SCM)
Digital Integration Model (DIM)

Figure 4.2: Methods to process movement data to actigraphy, modified from [Sto17]

whereas the authors present activity counts sampled at 1 Hz. To validate their algorithm, the authors conducted a 24 h free-living study with accelerometer and ActiGraph worn on the right hip in an elastic belt. They found that the algorithm enables a conversion with a Cohen's κ of 0.945, indicating an almost perfect agreement [Brø17]. However, the algorithm was validated with acceleration sensors sampled at 30 Hz. The same research group published a paper which indicates that the usage of sampling rates different than 30 Hz leads to considerable errors that researchers have to be aware of [Brø16].

4.2 Sleep Parameters

In order to obtain relevant information about the sleep habits of individuals, various sleep statistics can be applied, which are frequently used in the literature [Pal19, Li20, Hag21]. In this work, the usage of sleep statistics enables to find more information about classification performance for different sleep patterns like sleep onset or the time awake during the night. The following section lists the sleep statistics calculated and evaluated in this work:

- ***Total Sleep Duration (TSD)*** is the total duration spent sleeping, i.e., the duration between the beginning of he first sleep interval and the end of the last sleep interval in minutes.

- ***Net Sleep Duration (NSD)*** denotes the total time actually sleeping in minutes.

- ***Sleep Efficiency (SE)*** is defined as the ratio of Net Sleep Duration and Total Sleep Duration in percent.

- ***Sleep Onset Latency (SOL)*** is the time difference between going to bed and the beginning of the first sleep interval in minutes.

- ***Wake after Sleep Onset (WASO)*** is the summation of all epochs being awake between the first sleep interval and the last sleep interval.

4.3 Sleep/Wake Detection Algorithms

With the development of wearable sensors it became increasingly popular to extend sleep medicine with a home monitoring assessment. The acquired data can then be processed using different algorithms for sleep/wake prediction. Generally sleep/wake detection algorithms can be grouped into three concepts: Heuristic algorithms, machine learning algorithms and algorithms based on deep learning. These algorithms differ in the concept of classification as well as the input modalities that can be used. Figure 4.3 gives an overview about the algorithms implemented for this work and the following sections will present the algorithms in detail.

4.3.1 Heuristic Algorithms

In the initial phase of actigraphy-based sleep assessment, researchers developed various algorithms to predict sleep and wakefulness. One of the first actigraphy-based algorithms was published by Webster et al. [Web82] in the early 1980s. They recorded an actigraphy dataset which was sampled with one activity count per minute, and developed an heuristic Formula (see Equation 4.1) to estimate whether individuals were asleep or not. It consists of weights W for the current as well as for the four preceding and two subsequent epochs as well as a scaling factor S. Thereby, i denotes the epoch that needs to be scored.

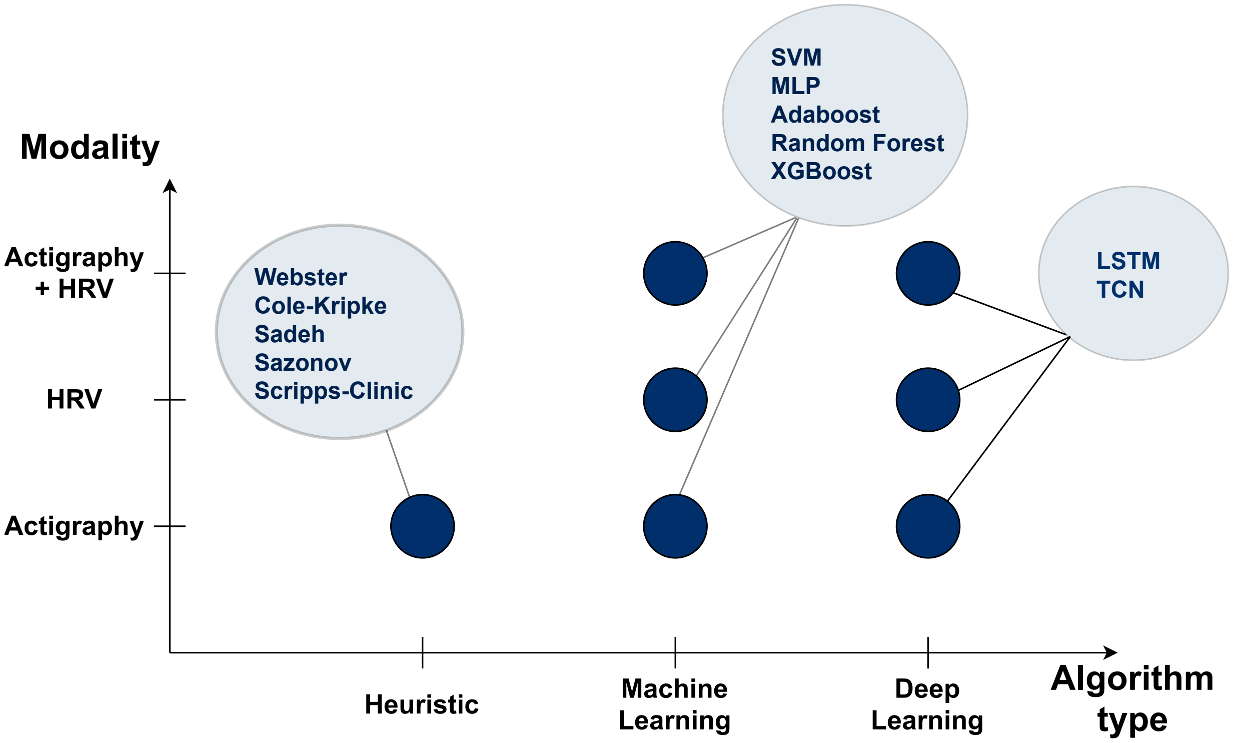

Figure 4.3: Overview of algorithms implemented for this work.

$$D(i) = S \cdot (W_1 \cdot T_{i-4} + W_2 \cdot T_{i-3} + W_3 \cdot T_{i-2} + W_4 \cdot T_{i-1} + W_5 \cdot T_i + W_6 \cdot T_{i+1} + W_7 \cdot T_{i+2}) \quad (4.1)$$

$$D = 0.25 \cdot (0.15 \cdot T_{i-4} + 0.15 \cdot T_{i-3} + 0.15 \cdot T_{i-2} + 0.08 \cdot T_{i-1} + 0.21 \cdot T_i + 0.12 \cdot T_{i+1} + 0.13 \cdot T_{i+2}) \quad (4.2)$$

According to the scoring rules, the current epoch *i* is scored as wake if $D \geq 1$. To find the optimal weights and scaling factor, they performed a grid search, and compared the sleep/wake estimation with a sleep scoring, obtained by ***EEG***. Applying Equation 4.1 on a small test dataset of three subjects, they found over 90% agreement. The weights and scaling factors resulting from the grid search yield the final scoring formula (Equation 4.2) [Web82].

One decade later, Cole et al. [Col92] published an updated version of the Webster algorithm, which was validated in a study with 42 participants, using actigraphy with concurrent ***PSG*** recording. They present the Webster Formula (Equation 4.1) optimized for different epoch lengths. Since the weights in Equation 4.2 were optimized based on the dataset from Webster et al., Cole et al. recalculated the weights and scaling factor according to their dataset. Furthermore, they extended the evaluation with relevant sleep parameters, like ***WASO***, ***SOL***, and ***SE***. Several sleep parameters calculated and evaluated in this work are listed in Section 4.2. The optimized formula (known as the *Cole-Kripke* algorithm) for an epoch length of 30 s is shown in Equation 4.3:

$$D = 0.0001 \cdot (50 \cdot T_{i-4} + 30 \cdot T_{i-3} + 14 \cdot T_{i-2} + 28 \cdot T_{i-1} + 121 \cdot T_i + 8 \cdot T_{i+1} + 50 \cdot T_{i+2}) \quad (4.3)$$

Applying the optimized equations for different epoch lengths to the corresponding dataset, they achieved agreements between 86 and 89 $\%$.

However, these approaches are rather basic and do not take signal metrics like variability into account. Sadeh et al. [Sad94] developed a more complex algorithm, including standard deviation, thresholding as well as logarithmic characteristics of the actigraphy signal. They performed discriminant analysis to adjust the weights of the scoring algorithm, which is presented in Equation 4.4.

$$PS = 7.601 - 0.065 \cdot \mu_{11min} - 1.08 \cdot \text{NAT}_{11min} - 0.056 \cdot sd_{6min} - 0.703 \cdot LOG_{Act} \quad (4.4)$$

Thereby, μ_{11min} denotes the average number of activity counts during the scored epoch as well as five epochs preceding and following it. sd_{6min} is the standard deviation of the current, as well as the five preceding epochs, whereas NAT_{11min} describes the number of epochs with activity levels equal or higher than 50, but lower than 100 activity counts in a window of 11 minutes, including the current as well as the 5 preceding and 5 following epochs. LOG_{Act} is the natural logarithm of the number of activity counts during the scored epoch + 1. If *PS* (probability of sleep) is zero or greater, the epoch is scored as sleep, if PS is smaller than zero, the epoch is scored as wake. To evaluate this algorithm, the authors conducted a study including 20 adults and 16 children wearing actigraphs at the dominant and non-dominant wrist. The overall agreement rates ranged between 91% and 93%, whereas the results of the non-dominant or dominant wrist were within the same range [Sad94].

However, if the data are not normally distributed but have a high degree of skewness, discriminant analysis as performed by Sadeh et al. [Sad94] is not validated. For this reason, Sazonova et al. [Saz02, Saz04] developed an approach based on a neural network and logistic regression. As sensing modality, they used accelerometers instead of actigraphs and sampled them to an epoch length of 30 s. The first part of their analysis consisted of applying logistic regression, in the form presented in Equation 4.5, where n denotes the number of previous 30 s epochs and ACC is the the movement measure, captured in the corresponding period.

$$\eta = log(\frac{p}{1-p}) = \beta_0 + \beta_1 \cdot ACC_0 + \beta_2 \cdot ACC_{-1} + ... + \beta_i \\ \cdot ACC_{i-1} + ... + \beta_{n+1} \cdot ACC_{-n} \quad (4.5)$$

The model classifies the epoch as sleep if the probability of sleep η is greater than 0.5 and wakefulness otherwise. The final model, presented in Equation 4.6 was then built by pooling four subjects into a training set and validating it against the remaining four subjects.

$$\eta = log(\frac{p}{1-p}) = 1.727 - 0.256 \cdot ACC_0 + 0.154 \cdot ACC_{-1} - 0.136 \cdot ACC_{-2} \\ - 0.140 \cdot ACC_{-3} - 0.176 \cdot ACC_{-4} \quad (4.6)$$

Using that approach the authors yielded an agreement rate of 76%, but reported a strong subject-specific effect that results in a huge variation in agreement between subjects ranging from 64.4% to 89.4%.

Because the algorithms only got tested in small studies, it is difficult to gain generalizability. Kripke et al. [Kri10] compared their new sleep/wake detection algorithm with the manufacturer's one and validated their results in a study with 116 subjects. Using am Excel Visual Basic macro, they developed a program to iteratively optimize their sleep/wake estimation equation:

$$D = S \cdot \sum_{i=-10}^{10} b_i \cdot x_i \quad (4.7)$$

Thereby, D was the scaled polynomial sum of activity counts for 21 epochs, weighted by a scaling factor b, each. x_0 represents the activity count for the epoch currently being evaluated, while x_{-10} and x_{10} denote the actigraph signal of the 10th preceding and 10th subsequent epoch respectively. When $D \geq 1$, the epoch was scored wake, while values smaller than one were classified as sleep. They found that the optimal parameters from b_3 to b_{10} are zero, so these parameters were excluded in Table 4.1, which depicts the parameters of the optimized equation. The optimal scaler value was found for $S = 0.300$, which resulted in 87% agreement by excluding one outlier [Kri10].

Epoch	Parameter	Epoch	Parameter
x_{-10}	0.0064	x_{-3}	0.0188
x_{-9}	0.0074	x_{-2}	0.0280
x_{-8}	0.0112	x_{-1}	0.0664
x_{-7}	0.0112	x_0	0.0300
x_{-6}	0.0118	x_1	0.0112
x_{-5}	0.0118	x_2	0.0100
x_{-4}	0.0128	x_3	0.0000

Table 4.1: Optimal scoring parameters for the Scripps Clinic Algorithm.

Although these heuristic actigraphy-based algorithms were a huge step in unobtrusive sleep medicine, all of them suffer from massive overprediction of sleep caused by little movement. In the work of Cole et al., a misclassification of wake as sleep occured 3.5 times as vice versa. For this reason, Webster et al. [Web82] developed a rescoring algorithm that adjusts the primary scoring according to the following rules:

- After at least 4 min scored wake, the first period of 1 min scored sleep is rescored wake.
- After at least 10 min scored wake, the first 3 min scored sleep are rescored wake.
- After at least 15 min scored wake, the first 4 min scored sleep are rescored wake.
- 6 min or less sleep surrounded by at least 10 min (before and after) scored wake are rescored wake.
- 10 min or less scored sleep surrounded by at least 20 min (before and after) scored wake are rescored wake.

Using that rescoring algorithm, Webster et al. were able to reduce the overestimation of sleep relative to the whole night from 1.89% to 0.81%, whereas Cole et al. gained a rise of overall agreement from 87.91% to 88.25%.

4.3.2 Machine Learning Algorithms

Although these heuristic algorithms are well established and already widely used, they have some disadvantages. Due to the simple, linear weights, this type of algorithm can hardly detect and classify complex sleep patterns. Furthermore the usage of datasets based on other sensing modalities is not possible. To enable a more complex sleep-wake classification, including high-level features of sleep parameters collected with different modalities, different state-of-the-art machine learning algorithms can be used.

Usually, machine learning algorithms provide a sample by sample estimation, which means that one input sample leads to one class sample [Bon17]. However, the probability of sleeping in one epoch strongly depends on the information of the previous and subsequent epochs. Hence, to achieve time-dependency, the extracted features can be calculated from the input data over sliding windows which will be explained in further detail in Section 5.1.3.

The performance of machine learning algorithms depends on the choice of hyperparameters. To find the optimal combination, a grid search can be performed over all parameters in a given search space. However, if the model is too computationally intensive, or the search space is too large, a randomized search can be used, i.e. random hyperparameter combinations are selected from the parameter search space for a certain number of trials [Kot07].

To evaluate the performance of machine learning models on unknown sleep data, it is important to divide the dataset into training- and test data to avoid overfitting [Kot07, Bon17]. Different methods with increasing computational complexity can be used for this purpose: The simplest method is to split the dataset into a training set containing between 60 and 80% of the samples and a test set containing the remaining samples. However, the choice of the train-test split can influence the results. To overcome this issue, k-fold cross-validation can be used. Thereby, the dataset is divided into k subsets, whereas each subset serves as test set in one of the k folds, while the other k-1 subsets serve as train set. When evaluating the results of the different folds, a small standard deviation indicates a stable algorithm and, thus, good generalizability [Kot07].

4.3.3 Deep Learning Algorithms

One of the most recent innovations in the field of ***Artificial Intelligence (AI)*** in the last decade is Deep Learning to train ***ANN***. These networks consists of multiple processing layers to discover patterns and structures in large datasets. Each layer learns a concept, on which the subsequent layer is based. The higher the level, the more abstract concepts are learned. Thereby, deep learning does not require prior data processing and extracts features automatically [Rus16].

A special class of neural networks that is suited to processing time-series or other sequential data are ***Recurrent Neural Network (RNN)***s. The most popular ***RNN*** is ***LSTM***, which achieves its time dependency by using input gates, memory cells, and forget cells, with weights adjusted for each cell [DiP20]. As ***RNN***s are feedforward networks, the training usually consists of gradient-based optimization, where gradients are obtained using backpropagation.

Although Deep Learning models can calculate features independently, they need to be optimized. Popular tunable hyperparameters are sequence length, learning rate, number of layers, hidden size and batch size. Usually these parameters are optimized using randomized search in a defined parameter space [DiP20].

Chapter 5

Benchmarking of Sleep/Wake Detection Algorithms

5.1 Dataset

5.1.1 Dataset Description

The ***MESA*** dataset is, to the authors knowledge, one of the largest open-access dataset that combines gold standard measurements of ***PSG*** with actigraphy and ***ECG***. The ***MESA*** study is a longitudinal investigation of factors associated with sub-clinical and clinical cardiovascular disease. The study, conducted at six centers in the United States, included 6,814 black, white, Hispanic, and Chinese-American men and women. Of these, 2,237 participants were enrolled in a Sleep Exam (***MESA*** Sleep) which included a sleep questionnaire, one week of actigraphy, and one night of concurrent ***PSG*** [Che15, Zha18].
The ***MESA*** Sleep study was conducted using wrist-worn actigraphy devices (Actiwatch Spectrum, Philips Respironics, Cambridge, USA) that records movements and aggregate them as activity counts in 30 s epochs. To record an in-home ***PSG***, the Compumedics Somte System (Compumedics Ltd., Abbotsford, Australia) was used. The ***PSG*** incorporates cortical ***EEG***, bilateral ***EOG***, chin ***EMG***, thoracic and abdominal respiratory inductance plethysmography, airflow, ***ECG***, leg movements and finger pulse oximetry [Che15, Zha18]. Details of sampling rates, scoring rules, and data collection protocols are available [Resa, Resb, Resc]. Nocturnal recordings were transmitted to the centralized reading center at Brigham and Women's hospital (Boston, USA) and scored by trained technicians following current guidelines. The QRS-complexes (R-peaks) obtained by the ***ECG*** were detected using the Compumedics (Abbotsford, VIC, Australia) Somte

software Version 2.10 (Builds 99 to 101) and reviewed by a trained technician who corrected misscored annotations during the sleep period.

5.1.2 Preprocessing and Cleaning

In this work, 1,743 of the initially 2,237 participants were included. 494 participants were excluded because of at least one of the following reasons:

- no concurrent ***PSG***, ***ECG*** and actigraphy recording available.
- no overlap of ***PSG*** and actigraphy specified.
- less than 1h of total sleep time.

Dataset	Total	Age	Female/Male	*TSD*	Ethnicity
Full dataset	$1,743$	69.1 ± 9.0	$951/792$	921.8 ± 170.2	36.7% White (639) 27.9% Black (487) 11.0% Chinese (192) 24.4% Hispanic (425)
Train set	$1,394$	68.9 ± 9.0	$759/635$	921.0 ± 169.9	36.5% White (509) 27.4% Black (382) 10.9% Chinese (152) 25.2% Hispanic (351)
Test set	349	69.9 ± 9.2	$192/157$	926.0 ± 172.0	37.2% White (130) 30.1% Black(105) 11.5% Chinese (40) 21.2% Hispanic (74)

Table 5.1: Statistics of Mesa Sleep Dataset, age $(U = 2.29 \cdot 10^5, p = 0.101, g = 0.4724)$ and ***TSD*** $(U = 3.12 \cdot 10^5, p = 0.443, g = 0.0296)$ are given as mean ± SD.

The participants were randomly divided into a training dataset of 1,394 and a test dataset of 349 participants using an 80/20 split. Further details regarding age, gender, and ethnicity of the participants in train-, test- and full dataset are provided in Table 5.1.2.

The preprocessing consists of several steps (Figure 5.1). In this process, the RR-intervals extracted from the ***ECG*** data were further cleaned and filtered to be reliable. Using the python package *hrv-analyis*[1], outliers of RR-intervals shorter than 300 ms and longer than 2000 ms were removed using a method described by Tanaka et al. [Tan01]. Removed beats were then imputed by linear

[1] https://github.com/Aura-healthcare/hrv-analysis [Rob21]

interpolation. The next step was to detect ecotropic beats and remove them according to the work of Malik [Mal96]. Then, the RR-intervals were grouped into epochs of 30 s to match the time interval of the data obtained by ***PSG*** and actigraphy. To ensure that only overlapping epochs were considered, actigraphy, ***PSG*** and RR-intervals were aligned using metadata information from the ***MESA*** dataset indicating the overlap of different recordings. Thereafter, all non-overlapping epochs were dropped out. Participants were further excluded if the remaining data after alignment contained less than 1 h of total sleep.

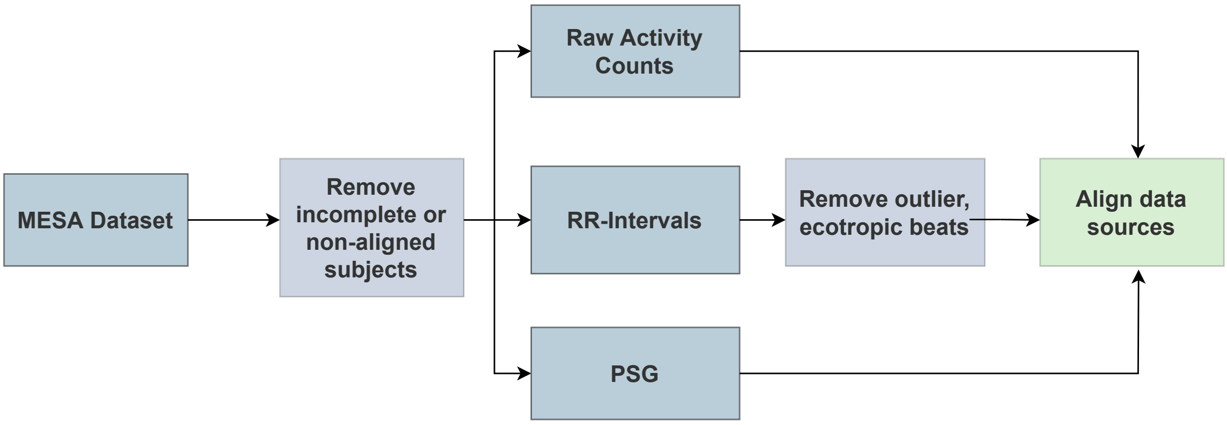
Raw Activity Counts
MESA Dataset
Remove incomplete or non-aligned subjects
RR-Intervals
Remove outlier, ecotropic beats
Align data sources
PSG

Figure 5.1: Preprocessing and cleaning of ***MESA*** Dataset.

5.1.3 Feature Extraction

As described in Section 4.3.2, machine learning models need appropriate features for good performance. This section describes feature extraction from actigraphy and cardiac data.
To evaluate sleep/wake detection using an motion-based approach, 370 handcrafted time-series features were extracted for an actigraphy-based machine learning approach. These features, described in Table 5.2, are basic statistical time-series features that are calculated over centered- and non-centered sliding windows of size $n = 1...19$. Moreover, these features were also used in recent literature [Pal19, Zha20].

As described in Section 4.1, using actigraphy alone has disadvantages such as massive over-prediction of sleep. Thus, it might be advantageous to add additional biosignals to the sleep/wake estimation. According to Section 3.1, the human body reduces the activity of core body functions during sleep, such as heart rate or blood pressure. For that reason, using cardiac information is a promising approach for sleep/wake detection. ***HRV*** features describe the variability of beat-to-beat intervals, extracted from the time interval between several R-peaks. Hence, 30 different ***HRV***

Feature Name	Description	Window Size
ACC	Raw actigraphy count	1
LOG	Natural logarithm of activity count	1
Mean	Mean value of activity counts	$1 \leq n < 20$
Median	Median value of activity counts	$1 \leq n < 20$
SD	Standard deviation of activity counts	$1 \leq n < 20$
Maximum	Maximum of activity counts	$1 \leq n < 20$
Minimum	Minimum of activity counts	$1 \leq n < 20$
Variance of	Variance of activity counts	$1 \leq n < 20$
NAT	Number of epochs with activity counts larger than 50, but lower than 100	$1 \leq n < 20$
ANY	Number of epochs that contain any activity count larger than 0	$1 \leq n < 20$
Skewness	Skewness of actigraph signal	$4 \leq n < 20$
Kurtosis	Kurtosis of actigraph signal	$4 \leq n < 20$

Table 5.2: Feature set extracted from raw actigraphy.

features were extracted for each 30 s interval according to previous work [Zha20]. The extracted features can be divided into four categories: time-domain, frequency-domain, non-linear, and geometrical features [Mal96]. Table 5.3 provides an overview of all ***HRV*** features used in this work.

5.2 Models and Settings

This section describes the application of heuristic, machine learning-based, and deep learning-based algorithms to the ***MESA***-Sleep dataset presented in Section 5.1. A multimodal approach including motion and cardiac data was compared to the two monomodal approaches using only actigraphy or ***HRV***.

5.2.1 Heuristic Algorithms

As already described, heuristic algorithms based on actigraphy were the first steps in unobtrusive sleep medicine. However these algorithms are static and only allow the evaluation of the movement-based approach using actigraphy. This work compares the following algorithms which were introduced in Section 4.3.1:

Feature	**Description**
	Time domain features
Mean NN	Mean over ***Normal-to-Normal (NN)***-Intervals
SDNN	Standard deviation of ***NN***-Intervals
SDSD	Standard deviation of ***NN*** differences
NN50	Number of ***NN***- Intervals greater than 50 ms
pNN50	Ratio between NN50 and number of ***NN***-Intervals
NN20	Number of ***NN***- Intervals greater than 20 ms
pNN20	Ratio between NN20 and number of ***NN***-Intervals
RMSSD	Root mean square of successive differences between ***NN***-Intervals
Median NN	Median of ***NN***-Intervals
Range NN	Range between smallest and largest ***NN***-Interval
CVSD	RMSSD divided by Mean NN (variation of successive differences)
CV NNI	The ratio of SDNN divided by Mean NN (variation of ***NN***-Intervals)
Mean HR	Mean Heart Rate
Max HR	Maximum Heart Rate
Min HR	Minimum Heart Rate
Std HR	Standard deviation of Heart Rate
	Geometrical domain features
Triangular Index	Integral of density distribution of ***NN***-Intervals (number of all ***NN***-Intervals) divided by the maximum of the density distribution
	Frequency domain features
LF	Variance (power) in low frequency (0.04 to 0.15 Hz)
HF	Variance (power) in high frequency (0.15 to 0.4 Hz)
VLF	Variance (power) in very low frequency (0.003 to 0.04 Hz)
LH/HF ratio	Ratio of Low frequency to high frequency
LF norm	Normalized LF power
HF norm	Normalized HF power
Total Power	Total power
	Non linear domain features
CSI	Cardiac Sympathetic Index [Jep14]
CVI	Cardiac Vagal Index [Jep14]
Modified CSI	Alternative measure of Cardiac Sympathetic Index [Jep14]
SD1	The standard deviation of the projection of the Poincaré plot [Beh13]
SD2	SD2 is defined as the standard deviation of the projection of the Poincaré plot on the line of identity [Beh13]
SD1/SD2 ratio	Ratio between SD2 and SD1 [Beh13]

Table 5.3: Feature set extracted from RR-intervals.

- Webster [Web82]
- Cole-Kripke [Col92]
- Sadeh [Sad94]
- Sazonov [Saz02, Saz04]
- Scripps-Clinic [Kri10]

These algorithms are easy to use because they do not require to compute higher-level features, since they only use the raw activity counts as input. To find the optimal scaling value (only for Webster, Cole-Kripke, Scripps-Clinic), and to assess whether applying Webster's rescoring algorithm (see Section 4.3.1), a grid search was performed over a pre-defined search space. Detailed information about the search space is provided in Appendix A in Table A.1.
As a part of this work, the heuristic algorithms listed above were implemented and included in the open-source python library *biopsykit*[2] [Ric21]. The code can be found online at:
https://github.com/mad-lab-fau/BioPsyKit.

5.2.2 Machine Learning Algorithms

Due to the weaknesses of actigraphy, explained in Section 4.1, it may be beneficial to make use of different modalities. This work compares the performance of actigraphy- and cardiac-based monomodal approaches with a multimodal examination combining both. The features used for this are presented in 5.1.3 and summarized in Table 5.4.

Modality	**Features**	**Number of features**
Actigraphy	Features derived of raw activity counts	370
HRV	***HRV*** features derived of RR-intervals	30
Multimodal	Combination of ***HRV*** and Actigraph-based features	400

Table 5.4: Modalities and input features for machine learning models.

Because of many publications using different studies as well as only a few algorithms, it is difficult to make a reliable and generalized statement about the performances of different well-established machine learning algorithms in sleep/wake detection. To gain comparability, the

[2] https://biopsykit.readthedocs.io/en/latest/api/biopsykit.sleep.sleep_wake_detection.algorithms.html

mono- and multimodal approaches were also applied to several state-of-the-art machine learning models:

- ***Adaptive Boosting (AdaBoost)***
- ***Support Vector Machine (SVM)***
- ***Multi Layer Perceptron (MLP)***
- Random Forest
- ***XGB***

As hyperparameters are govern to influence the training and therefore also the final model, an appropriate combination of hyperparameters has a major impact on the performance of machine-learning models. To find the best set of hyperparameters for ***AdaBoost***, ***MLP***, and ***SVM***, a grid search with an embedded 5-fold cross-validation was performed over a defined parameter search space. As ***SVM*** and ***MLP*** calculate distances between datapoints, the features were normalized by removing mean and scaling to unit variance. Due to a large number of hyperparameters for Random Forest and ***XGB*** and the associated high computational cost, a 1,000 trial-Random Search including an embedded 5-fold cross-validation was performed for each algorithm and modality. In this context, Random Search means that the hyperparameters are selected from a defined parameter search space in a random manner for a given number of trials or time. In this work, the python package *optuna*[3] was used, which is an open-source framework to automate hyperparameter search [Aki19]. To determine the hyperparameters for each trial, *optuna* fits one ***Gaussian Mixture Model (GMM)*** G_1 to the set of parameter values associated with the best objective values, and another ***GMM*** G_2 to the remaining parameter values. The next hyperparameters are then chosen by the maximation of the ratio G_1/G_2. Thereby, the algorithm performance of all machine-learning algorithms was optimized towards accuracy. More information, including the hyperparameter search space, can be found in the Appendix A in Table A.2. The pipeline used for the machine learning algorithms is illustrated in Figure 5.2.

5.2.3 Deep Learning Algorithms

As explained in Section 4.3.3, ***RNN***s are powerful deep learning-based methods for the classification of time series. For that reason, this work compares the performance of deep learning models with the various heuristic and machine learning algorithms presented above. Thereby, two

[3] https://optuna.org/

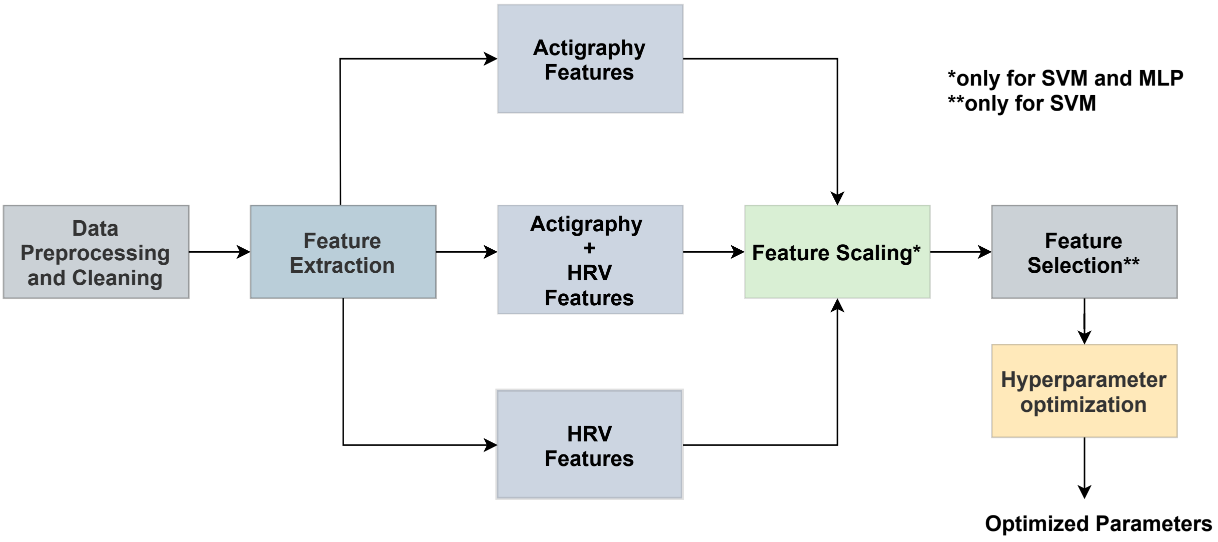

Figure 5.2: Pipeline for the machine learning based approach.

monomodal approaches using actigraphy and ***HRV*** were compared with a multimodal approach combining both inputs. The deep learning algorithms that were evaluated in this work are:

- ***LSTM***
- ***Temporal Convolutional Network (TCN)***

This work uses an LSTM provided by the python package *torch*[4] [Col11] and a ***TCN*** published by Bai et al. [Bai18]. In time-series classification, the choice of the optimal sequence length is crucial. For that reason, both networks expect special input shapes. Thus, the first step was to prepare the input data for further processing. The ***LSTM*** expects 3-dimensional input data with shape `(batch_size, sequence_length, number_of_features)`. To prevent sequences that exist of more than one participant, participant-wise sequencing was performed. The shape of the input data `(number_of_samples, sequence_length, number_of_features)` was created using a sliding window function from the python package *biopsykit*[5] [Ric21]. Sequence length and sliding window overlap were considered as optimizable parameters. The ***TCN*** network expects the data with an input shape `(batch_size, number_of_features, sequence_length)`. For that, the same data shaping as for the ***LSTM*** was applied, followed by flipping the second and third dimensions. Since deep learning models require feature scaling,

[4] https://pytorch.org/docs/stable/torch.html
[5] https://biopsykit.readthedocs.io/en/latest/api/biopsykit.utils.array_handling.html

Modality	**Input data**
Actigraphy	Raw activity counts
HRV	Mean NN SDNN SDSD VLF LF HF LF/HF ratio Total Power
Multimodal	Concatenation of Actigraphy and ***HRV*** features

Table 5.5: Modalities and input data for deep learning models.

all input data were standardized by means of z-normalization, corresponding to a distribution with $\mu = 0$ and $\sigma = 1$, using the python package *sklearn*[6] [Ped11]. For an epoch-wise training of the Deep Learning networks, ***Adaptive Moment Estimation (Adam)*** was applied, a gradient descent-based optimization method that computes adaptive learning rates for each parameter. Equation 5.1 depicts the ***Adam*** update rule [Kin17], where $\hat{m_t}$ and $\hat{v_t}$ are the bias-corrected first and second moment estimates, η the learning rate and ϵ a constant with a default value of 10^{-8}. σ_t and σ_{t+1} denote the weights of the current epoch *t* and the following epoch *t + 1*.

$$\sigma_{t+1} = \sigma_t - \frac{\eta}{\sqrt{\hat{v_t}} + \epsilon}\hat{m_t} \tag{5.1}$$

The loss was calculated via binary cross-entropy loss, which is depicted in Equation 5.2. Here, y is the label for being asleep or not, while $p(y)$ is the predicted probability of the epoch being classified as sleep for all N epochs.

$$H_p(q) = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{5.2}$$

As described in Section 4.3.3, Deep Learning models automatically extract high-level features. Therefore, raw activity counts served as input for the movement-based approach, while 8 basic ***HRV*** features were considered as input for the cardiac-based classification (see Section 5.1.3). To compare actigraphy-based and cardiac-based classification, the deep learning algorithms were applied first trained with monomodal data followed by a multimodal approach. Table 5.5 gives

[6] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

an overview of the different approaches along with the input data. Since the performance of Deep Learning models is highly dependent on the choice of hyperparameters such as learning rate, number of layers, or hidden layer size, an *optuna*[7]-based randomized search was performed. Additional information about the hyperparameter search space is provided in Appendix A in Table A.3. The pipeline used for the deep learning algorithms is illustrated in Figure 5.3.

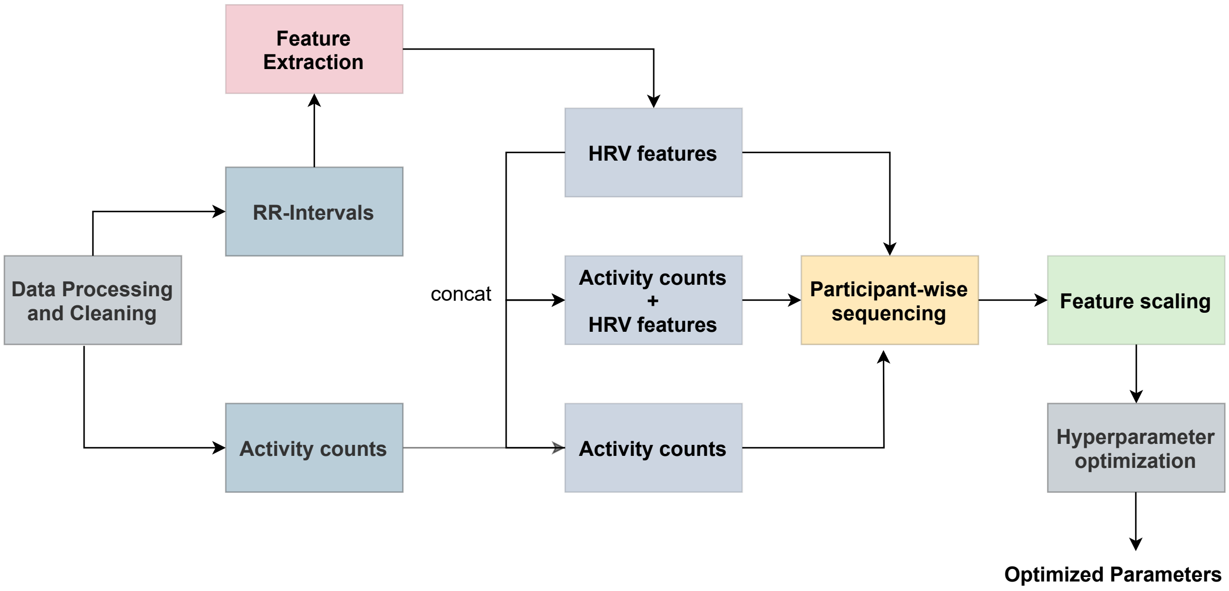

Figure 5.3: Pipeline for the deep learning based approach.

[7] https://optuna.org/

Chapter 6

Sleep/Wake Classification with Real-world Data

6.1 Real-World Study

6.1.1 Dataset and Study Protocol

Since the acquisition of the ***MESA*** dataset was performed under controlled laboratory environment conditions, one purpose of this thesis was to evaluate the developed algorithms using real-world data.

The study conducted as part of this work involved 22 participants recording concurrent ***IMU*** and ***ECG*** data during sleep for 3 nights. During the study, the participants were able to sleep in their natural environment without any supervision or interferences.

Thus, a new dataset including ***ECG*** and ***IMU*** data was collected using two wearable sensors (Portabiles NilsPod, Porabiles GmbH, Erlangen, Germany) worn at the chest and on the wrist of the non-dominant hand. The chest sensor recorded 1-channel ***ECG*** and 6-d ***IMU*** data consisting of 3-d acceleration (range ±16g) and 3-d angular rate (range $\pm 2000°/s$) whereas the wrist sensor only recorded 6-d ***IMU*** data. Both sensors recorded with a sampling rate of 256 Hz and were synchronized wirelessly [Rot18]. Thus, data from both sensors can be processed on a common time axis.

All data are logged onto the internal sensor storage and downloaded afterwards using the application *PortabilesHomeMonitoring*[1] for Android-based smartphones. The application automatically

[1] PortabilesHomeMonitoring(1.1.12)https://play.google.com/store/apps/details?id=de.portabiles.homemonitoring

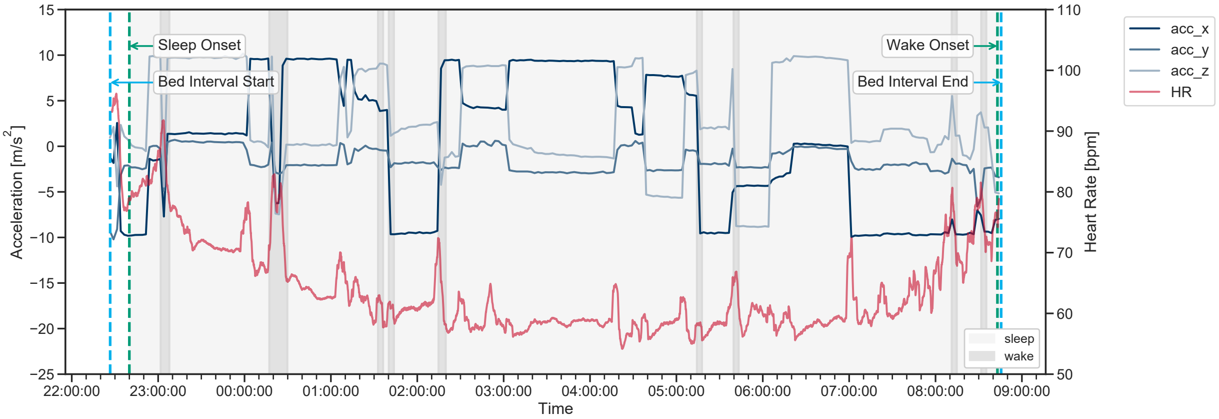

Figure 6.1: Full overnight recording of ***IMU*** signal with concurrent heart rate monitoring. acc_x, acc_y, and acc_z denote the accelerometer axes whereas HR denotes the heart rate.

downloads the sessions from the sensors via ***Bluetooth Low Energy (BLE)*** when smartphone and sensors both are charging, in Bluetooth range, and the sensors have new sessions available. Upon download, the sessions are deleted from the sensors to allow the recording of further sessions. Figure 6.1 shows an overnight recording of 3-axis accelerometer with concurrent heart rate monitoring and Figure 6.2 depicts the placement of both sensors.

To validate the sleep/wake detection, this work used a clinically validated sleep mat (Withings Sleep Analyzer, Withings France SA, Issy-les-Moulineaux, France) as ground truth. The sleep mat was placed below the mattress on the slatted frame. With the help of highly sensitive sensors movement, heart rate and respiration can be measured and used for sleep staging [Edo21].

The study was structured as follows:

Task 1: Sensor configuration: The ***ECG*** and ***IMU*** sensors were configured and registered in the *PortabilesHomeMonitoring* app.

Task 2: Participant briefing: The participants were briefed in the facilities of the Machine Learning and Data Analytics Lab of Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) in Erlangen, Germany. Thereby, the sensors and sleep mat were handed over, and a user profile was created in the *Withings HealthMate*[2] app.

[2] WithingsHealthMate(5.7.1)https://play.google.com/store/apps/details?id=com.withings.wiscale2

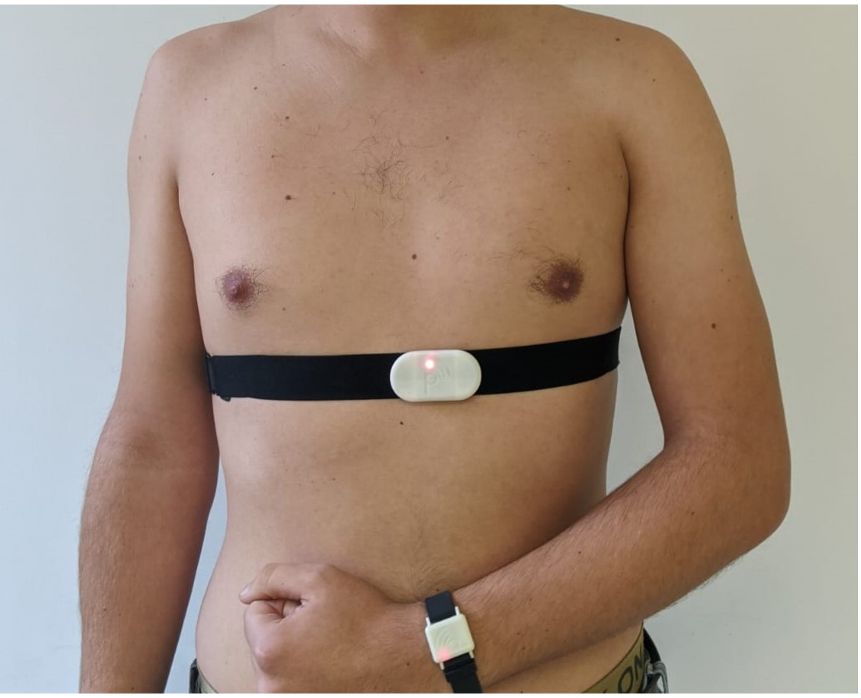

Figure 6.2: Placement of ***IMU*** and ***ECG*** sensors.

Task 3: Installation of sleep mat and sensor charging: At their home, participants were instructed to install the sleep mat themselves according to an explanation in the study protocol and charge the sensors in their charging cradles. For potential technical issues, contact with the study supervisor was provided.

Task 4: Recording start (evening): The sensors were configured to automatically start recording data five seconds after removal from the charging cradle. Thus, before going to bed, participants were asked to remove the sensors out of the charging cradles, attach them as depicted in Figure 6.2, and check whether they are recording synchronously.

Task 5: Recording stop (morning): After waking up the next morning, participants were instructed to stop data collection as soon as they got out of bed by putting the sensors back into the charging cradles.

Task 6: Data transfer: To ensure correct data transfer, participants were instructed to check in the *PortabilesHomeMonitoring* app whether the session download had been started correctly. If not, they were asked to restart the session download manually. The sleep mat was connected to WiFi and transferred data from the preceding night to the Withings cloud.

Task 7: Sleep diary questionnaire: Afterwards, participants were asked to fill out a sleep diary questionnaire about the last night. The questionnaire includes questions about subjective sleep quality, bed time, self-estimated sleep onset, wake onset, etc. Furthermore, it was assessed whether participants had consumed more than two alcoholic beverages the evening before since it is known that alcohol affects sleep [Ebr13] and, thus, it was hypothesized that alcohol consumption might also influence the classification performance of sleep/wake detection algorithms. All questions can be found in Appendix C.

Task 8: Night 2 and 3: On the second and third night, the participants had to repeat tasks 4-7.

Task 9: PSQI: Upon recording three nights, participants were instructed to fill out the ***PSQI***, which assesses sleep quality of the last four weeks.

To obtain a larger dataset, a second study conducted at the Machine Learning and Data Analytics Lab of Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) was merged with the study conducted for this thesis. This study included 22 individuals as well and was conducted as part of the ***Digital Psychology Lab (DiPsyLab)*** seminar. Thereby, it dealt with the assessment of nightly heart rate patterns and their relation to stress and stress coping strategies. Because this study is based on the same ground truth acquired by the same sleep mat as well as the same wearable sensors, it was possible to combine both studies to obtain a larger dataset. However, this study included only two nights per participant and a slightly deviating questionnaire.

6.1.2 Data Cleaning

The combined studies initially included 107 nights of 42 participants. In order to use only high-quality data, strict data cleaning was performed, excluding 22 nights for at least one of the following reasons:

- Missing ground truth data (more than 40% of the night with missing ground truth labels).
- Ground truth not available (Technical problems of the sleep mat).
- Corrupted sensor data.
- Invalid sensor output.

The remaining dataset contained 50 nights of 20 different participants from the study conducted for this thesis as well as 35 nights of 22 participants from the ***DiPsyLab*** study. Table 6.1 provides additional information and statistics about the participants who participated in the studies.

Dataset	Participants	Nights	Age [yr]	Female/Male	Height [cm]	Weight [kg]
Thesis	20	50	25.5 ± 4.9	11 / 11	175.2 ± 9.7	68.3 ± 12.7
DiPsyLab	22	35	22.9 ± 2.3	11 / 8	178.3 ± 8.5	70.8 ± 9.0
Combined	42	85	24.3 ± 4.0	22 / 19	176.7 ± 9.2	69.5 ± 11.1

Table 6.1: Information about the examined participants.

Since the ***DiPsyLab*** study was conducted without the sleep diary questionnaire described above, only the participants of the other study could be evaluated according to the characteristics described in Table 6.2. Moreover, demographic information of three individuals participated in the ***DiPsyLab*** study were missing.

Total nights	Alcohol	No alcohol	Alarm	No alarm
50	13	37	39	11

Table 6.2: Information about the examined nights collected within this work.

6.1.3 Data Processing

The ground truth data from the sleep mat were converted into time-series data and exported as csv file using *BioPsyKit*[3] [Ric21]. Since the sleep mat measures different sleep phases the sleep stages were converted into binary sleep/wake classes.
Before the data from the wearable sensors were used for sleep/wake classification they first had to be preprocessed. For that the binary files were imported into python using *BioPsyKit*[4] [Ric21]. As technical problems occurred with some sensors, further error handling had to be performed for some nights, such as truncating the start or end of a session due to invalid sample counters or incorrect start timestamps. Unfortunately, a total of 9 sessions of the study conducted for this work as well as 5 nights of the ***DiPsyLab*** study had to be excluded due to corrupted sensor data. After data loading, the data of both sensors were aligned using the synchronized counter. The synchronized session was then aligned with the sleep labels extracted from the sleep mat. Thereby, all non-overlapping epochs were removed. As next step, the sensor data were calibrated using the python library *imucal*[5] [Kü21] which implements the IMU-infield calibration based on Ferraris et al. [Fer94]. The Ferraris calibration is a simple and affordable calibration method for 6 DOF IMUs as it can be performed by placing the sensor on each side and rotating it around each axis. To use the ***ECG*** data for further processing and feature extraction, RR-intervals were extracted. This involved cleaning and filtering of the raw ***ECG*** signal followed by R-peak detection and outlier correction with *BioPsyKit*[6] [Ric21].

6.1.4 Feature Extraction

One goal of this thesis was to evaluate the newly developed algorithms on real-world data by collecting a new dataset. However, the data collected within this thesis was ***IMU*** data, while the ***MESA*** benchmark dataset only contained actigraphy data. To compare the results of both datasets, a conversion from acceleration data to actigraphy according to Brønd et al. [Brø17] (Section 4.1) was performed using an implementation from *BioPsyKit*[7] [Ric21]. Using the activity counts resulting from this conversion, the same actigraphy-based features were calculated as for the ***MESA*** dataset (see Table 5.2). For the ***HRV***-based sleep/wake detection, the extracted

[3] https://biopsykit.readthedocs.io/en/latest/api/biopsykit.io.sleep_analyzer.html
[4] https://biopsykit.readthedocs.io/en/latest/api/biopsykit.io.nilspod.html
[5] https://github.com/mad-lab-fau/imucal
[6] https://biopsykit.readthedocs.io/en/latest/api/biopsykit.signals.ecg.html
[7] https://biopsykit.readthedocs.io/en/latest/api/biopsykit.signals.imu.activity_counts.html

Feature Name	Description	Window Size (in s)
Mean$_{acc\_gyr}$	Mean of $\{$Acc, Gyr$\}$ norm	$30 \leq n < 600$
Median$_{acc\_gyr}$	Median of $\{$Acc, Gyr$\}$ norm	$30 \leq n < 600$
SD$_{acc\_gyr}$	Standard Deviation of $\{$Acc, Gyr$\}$ norm	$30 \leq n < 600$
Variance$_{acc\_gyr}$	Variance of $\{$Acc, Gyr$\}$ norm	$30 \leq n < 600$
RMS$_{acc\_gyr}$	Root Mean Square of $\{$Acc, Gyr$\}$ norm	$30 \leq n < 600$
Minimum$_{acc\_gyr}$	Minimum of $\{$Acc, Gyr$\}$ norm	$30 \leq n < 600$
Maximum$_{acc\_gyr}$	Maximum of $\{$Acc, Gyr$\}$ norm	$30 \leq n < 600$
Absolute Energy$_{acc\_gyr}$	Absolute energy of $\{$Acc, Gyr$\}$ norm	$30 \leq n < 600$

Table 6.3: Features extracted from accelerometer- and gyroscope norm, respectively.

RR-intervals were processed similarly as described in Section 5.1.3 using the python package *hrv-analysis*[8] [Rob21] resulting in 30 high level ***HRV*** features. An overview of these features is provided in Table 5.3.

Since actigraphy is usually sampled in 30 s or 1 min epochs, important movement information may get lost. For this reason, this work aimed to investigate whether sleep/wake detection performance can be further improved by using other modalities, such as raw ***IMU*** data, instead of aggregated actigraphy data. Thus, several high-level features based on accelerometer and gyroscope data were extracted. Using the python package *tsfresh*[9], eight basic features were calculated from the norm vectors of gyroscope and accelerometer data, resulting in 16 different features (Table 6.3). Because the features were calculated in 20 different sized sliding windows, starting from 30 s and rising in 30 s steps up to 10 min, the feature extraction resulted in a set of 304 motion-based features. As the ***IMU*** data was sampled at 256 Hz, downsampling to 32 Hz was performed in advance. According to the work of [Kha16], this enables faster processing while retaining important information.

6.2 Sleep/Wake Classification

To evaluate the newly developed sleep/wake detection algorithms based on the real-world dataset using different algorithms and modalities, the same algorithms as presented in Chapter 5 were used. Moreover, the search spaces of the hyperparameter optimization were similar and are listed in Appendix A.

While the heuristic algorithms can only be investigated in a motion-based approach, the machine- and deep-learning-based approaches were evaluated in monomodal approaches of ***HRV*** and

[8] https://github.com/Aura-healthcare/hrv-analysis

[9] https://tsfresh.readthedocs.io/en/latest/

motion data as well as in a multimodal approach combining both input modalities.

Since the real-world data set used in this work is highly imbalanced and about $90\ \%$ of the samples are labeled as sleep, optimization towards accuracy is not feasible. For that reason, algorithms trained on real-world data were optimized on Cohen's κ, which is known to be more robust to imbalanced datasets compared to classical measures such as accuracy [Can13, Sui19]. More information about the evaluation metrics used in this work is provided in Chapter 7.

The sleep/wake estimation based on real-world data examined in this work was separated into two parts. To compare the actigraph-based approach with the ***IMU***-based approach, the raw accelerometer samples were converted into activity counts which served as base for sleep/wake detection. The second part consists of an ***IMU***-based approach which is based on accelerometer and gyroscope data.

6.2.1 Actigraphy-based Sleep/Wake Classification

To compare the performance of the heuristic motion-based algorithms, the converted activity counts were applied to the heuristic algorithms described in Section 4.3.1. Since the heuristic algorithms only accept activity counts as input, no multimodal or ***HRV***-based approach was examined.

In contrast, the machine learning approach was evaluated using two monomodal approaches, including actigraphy-based and ***HRV***-based features, as well as a multimodal approach combining both modalities. To evaluate the machine learning algorithms for real-world data, the actigraphy-based features described in Table 5.2 were computed as described above. The sleep/wake classification was performed in two different approaches: First, the actigraphy and ***HRV*** features were applied to the machine learning algorithms. The models were optimized using a grid search with embedded cross-validation for ***SVM***, ***MLP***, and ***AdaBoost***, while the Random Forest and the ***XGB*** models were optimized using a randomized search. Here, the hyperparameter search space was identical to the hyperparameter optimization used for the benchmark dataset which is provided in Appendix A.

The actigraphy-based sleep/wake detection for the real-world dataset was performed in two ways. First, the algorithms were both trained and tested on the ***HRV*** features and the activity counts extracted from the ***IMU*** signal (see Section 4.1). Additionally, the machine and deep learning models trained on the benchmark dataset were used to be evaluated on the real-world dataset since for both datasets, actigraphy and ***HRV*** data were present. Thus, the same features were extracted for both datasets.

6.2.2 IMU-based Sleep/Wake Classification

Due to low sampling rates and built-in pre-processing of actigraphy devices, activity counts as output are easy to handle and easy to understand. However, information can be lost due to aggregation into epochs of 1 min or 30 s. For this reason, this work examines the performance of ***IMU***-based sleep/wake detection and compares it with the actigraph-based approaches.

Since the heuristic algorithms presented in this work are only developed for actigraphy as input modality, no examination of heuristic algorithms using ***IMU***- or cardiac data was conducted.

For the machine learning-based sleep/wake classification, the ***IMU***-based features (see Table 6.3), were used. The eight basic features presented in Section 5.3 were used as cardiac features.

The deep learning algorithms were applied using raw accelerometer and gyroscope data as movement-based input modality. Thereby, the norm of the raw accelerometer and gyroscope data was used. To save computational cost, the signal was downsampled to 8 Hz. The correct input was generated via participant-wise sequencing, as described in Section 5.2.3. The ***HRV***-features presented in Section 5.5 were used as cardiac-based input modality.

Chapter 7

Evaluation

7.1 Evaluation Metrics

To evaluate and compare the classification performance of the different algorithms, several evaluation metrics were applied which are based on ***True Positive (TP)***, ***False Positive (FP)***, ***True Negative (TN)***, and ***False Negative (FN)*** obtained from the confusion matrix.

1. **Accuracy** accounts for the number of correctly classified epochs, divided by the total amount of epochs:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7.1}$$

2. **Precision** is the number of correctly classified sleep epochs, divided by the total number of epochs classified as sleep:

$$Precision = \frac{TP}{TP + FP} \tag{7.2}$$

3. **Recall (also known as sensitivity)** denotes the number of correctly classified epochs being asleep divided by the total number of epochs labeled asleep:

$$Recall = \frac{TP}{TP + FN} \tag{7.3}$$

4. **F1-score** is the harmonic mean of precision and recall:

$$F1 = \frac{2}{recall^{-1} \cdot precision^{-1}} = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{7.4}$$

5. **Cohen's κ** is a statistic to measure inter-rater agreement comparing observed accuracy with expected accuracy, where p_0 denotes the observed, and p_e the expected accuracy. Thereby, the expected accuracy denotes the accuracy occurring by chance. In sleep/wake detection, Cohen's κ is considered to be more robust against class imbalance [Can13, Sui19].

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{7.5}$$

7.2 Evaluation of Benchmark Dataset

To make reliable statements about the performances of the different algorithms with different modalities, respectively, several techniques to prevent overfitting were used. Thereby, the first step was to randomly separate 20% of the dataset to use as test set which is only used for the final evaluation with the final classifier models.

Since the heuristic algorithms are static and hence do not need to be trained, no separation into train/test set would be required. However, to compare the heuristic algorithms with the machine- and deep learning-based algorithms, the evaluation was performed on the same test set. Because no intrinsic function exists to find the optimal scaling parameters, different parameter combinations were tested using a grid search applied on the training set. This grid search was then repeated in a 5-fold cross-validation train/test splits. The upper part of Figure 7.1 illustrates a 5-fold cross-validation. The final scaling parameter was calculated as the mean over all cross-validation folds.

As the machine learning models introduced in Chapter 5 have tunable hyperparameters, they were optimized and evaluated using a parameter search with an embedded 5-fold cross-validation. The performance of the classifiers for different hyperparameter combinations were evaluated within five-fold cross-validation. These different splits of train and validation set are supposed to prevent overfitting during hyperparameter optimization. The best-performing classifier is yielded by the hyperparameter combination which achieved the best average performance over the five folds. The classifier is then retrained with this hyperparameter set on the complete training data. Finally, the retrained classification model is evaluated on the test set and the performance metrics are reported. Due to the computational cost of deep learning algorithms, no cross-validation was

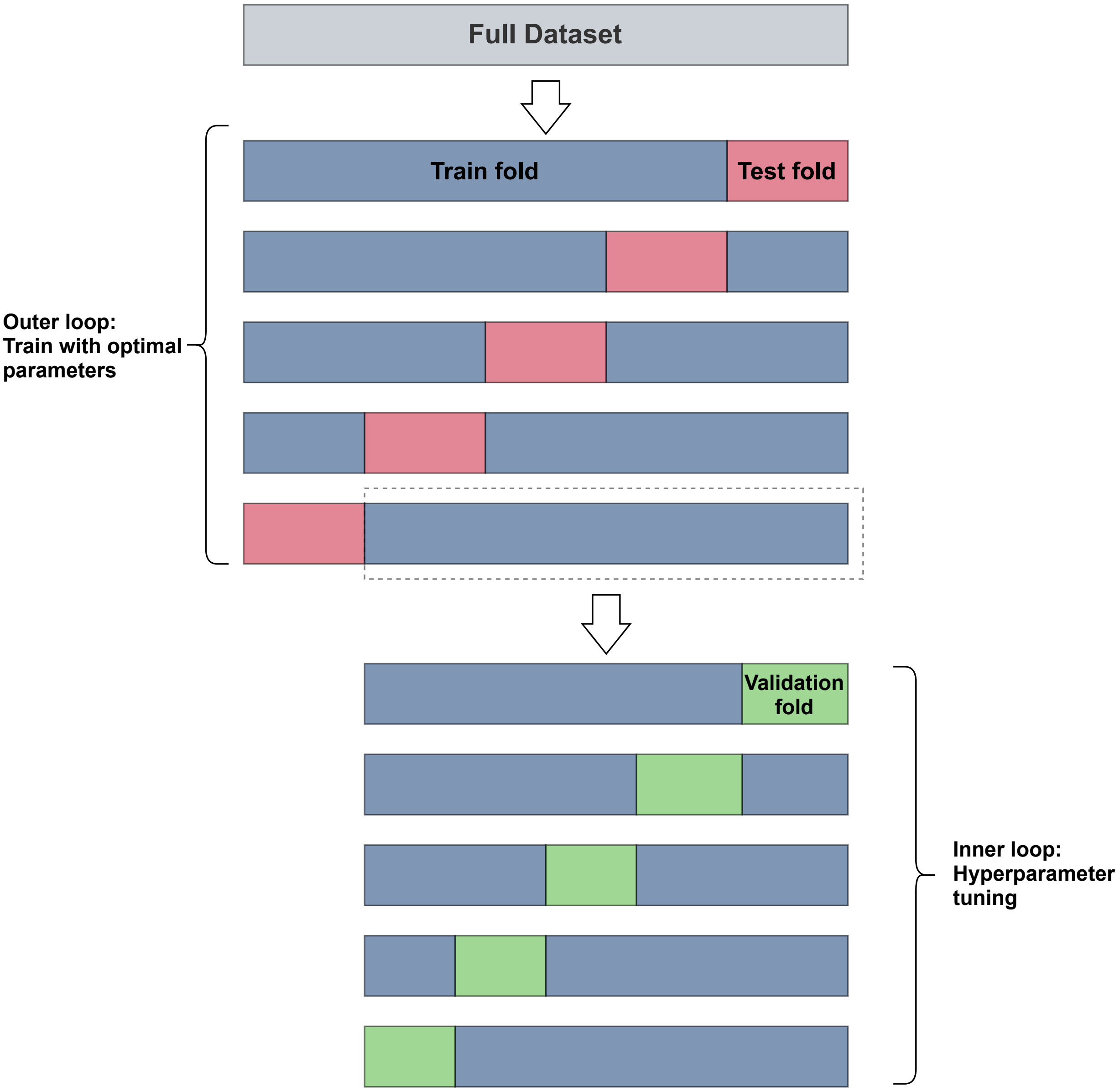

Figure 7.1: Nested cross-validation.

performed. Instead, the epoch-wise trained algorithms were evaluated using a single validation set split off from the training set. Thereby, after each epoch, the validation set was applied to the current model and the corresponding loss was calculated. Each time a new lowest validation loss was found, the model was saved. After a defined number of epochs, the optimization aborted and the model was evaluated on the test set.

7.3 Evaluation of Real-world Study

The real-world study contains sleep data from 85 nights, including special characteristics that occur infrequently, i.e. waking up with an alarm clock is only reported for 11 nights while consuming alcohol before bed time occurred 13 times. Using the same approach as for the benchmark dataset with a hold-out test set with an 80/20 split would result in only 2 or 3 nights with these characteristics being included in the test set, which would limit generalizability due to strong subject-dependent influences.

For this reason, the machine learning models were evaluated using nested cross-validation. Thereby, the hyperparameters of the algorithms were optimized like described for the benchmark approach using 5-fold cross-validation with different train and validation splits. However, instead of testing the final optimized model with a hold-out test set, this optimization is done using another 5-fold cross-validation including the whole dataset. This means that all data are used as test set once on a trained model, optimized with the rest of the data. Figure 7.1 gives an overview of the nested cross-validation applied in this work.

7.4 Statistical Analyses

In addition to the overall performance of sleep-wake recognition, several other investigations were conducted to determine the influence of participant- and study-specific characteristics. As the ***MESA*** dataset contains a large amount of additional demographic and clinical information about each participant, the following properties that would suggest the greatest influence were extracted and evaluated:

- **Signal Quality of *PSG* and Actigraphy:** The signal quality for both ***PSG*** and Actigraphy was rated on a range of two (poor) to seven (perfect). To compare the influence of signal quality, the scale was converted into a dichotomous scale which values less than four being labeled as bad quality and all values larger or equal then four being labeled as good quality.

- **Gender:** The gender was provided as a dichotomous variable indicating male or female gender of participants.

- **Race:** The dataset contains 36.7% White/Caucasian, 11% Chinese American, 27.9% Black/African-American, and 24.4% Hispanic participants.

- **Sleep Quality:** The ***WHIIRS*** questionnaire was filled out, which was designed to assess ***INS*** symptoms. The resulting scale was converted into a dichotomous variable to distinguish between good and bad sleep. A value of nine was chosen as cut-off value as ***WHIIRS*** scores above nine indicate ***INS*** (Section 3.2).

- **Extra Workload:** One value obtained from the data set is the additional workload per week. This value was classified as high workload if individuals worked more than 5 h of overtime per week.

- **Sleep Diseases:** The ***MESA*** dataset contained clinical information about the diagnosis of ***RLS***, ***INS***, and sleep apnea. It was evaluated if one of the diseases or being sick in general influences the classification performance.

- **Age:** As an important demographic information, age was provided for each participant. All participants were between 54 and 93 years old (Mean ± SD: 69.9±9.2 years).

- ***AHI*****:** The ***AHI*** seves as indicator for frequency and severity of apnea and hypopnea (Section 3.2).

For the real-world study conducted in this work, a sleep diary questionnaire was filled out. Furthermore, the following demographic and sleep quality scores were evaluated as potential confounders for sleep/wake classification performance:

- **Alcohol consumption:** It was evaluated whether the consumption of more than two alcoholic beverages influenced the classification performance.

- **Subjective sleep quality per night:** The subjective sleep quality score of each night night was rated from one (bad sleep) to seven (perfect sleep) (Questionnaire: C). To evaluate the classification performance according to the subjective sleep quality, all quality scores larger than five were labeled as good sleep while scores lower or equal than five were labeled as bad sleep quality.

- ***PSQI*:** As ***PSQI*** scores larger than five indicate poor sleep quality (see Section 3.2), sleep quality was rated as good for scores from zero to five and poor for scores from six to 21. (see Section 3.2).
- **Mode of awakening:** It was assessed whether the mode of awakening (alarm vs. no alarm) influenced sleep/wake classification.
- **Profession:** It was evaluated whether the current profession (student vs. employee) has an influence on classification performance.
- ***Body Mass Index (BMI)*:** With weight and height collected from the ***PSQI*** questionnaire, the ***BMI*** was collected. It was evaluated if high or low ***BMI*** influenced the classification performance.

As the resulting classification metrics violated the assumptions of normal distribution, non-parametric statistical analysis was performed. For features with only two conditions, such as gender (male vs. female), the *Mann-Whitney-U-Test* [McK10b], the non-parameteric version of the t-test for independent samples was applied to detect possible differences. Since several algorithms from different modalities got tested, Bonferroni correction for multiple-comparison correction was applied.

For characteristics with more than two groups, such as the comparison of different algorithms or different races, the *Kruskal-Wallis-Test* [McK10a] was applied to determine group differences. As post-hoc tests, *Mann-Whitney-U-Tests* with Bonferroni correction for multiple-comparison correction were applied.

The significance level was set to $\alpha = 0.05$. In all Figures and Tables, the following notation is used to indicate statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Chapter 8

Results

8.1 Classification of Benchmarking Dataset

One major research goal of this work was to systematically compare different algorithms with different modalities on the same benchmarking dataset. For this, twelve different algorithms were applied on the ***MESA*** dataset. The resulting performance measures are presented in Table 8.1. Additionally, Figure 8.1 visualizes the best algorithm in terms of Cohen's κ of every group separated between different input modalities.

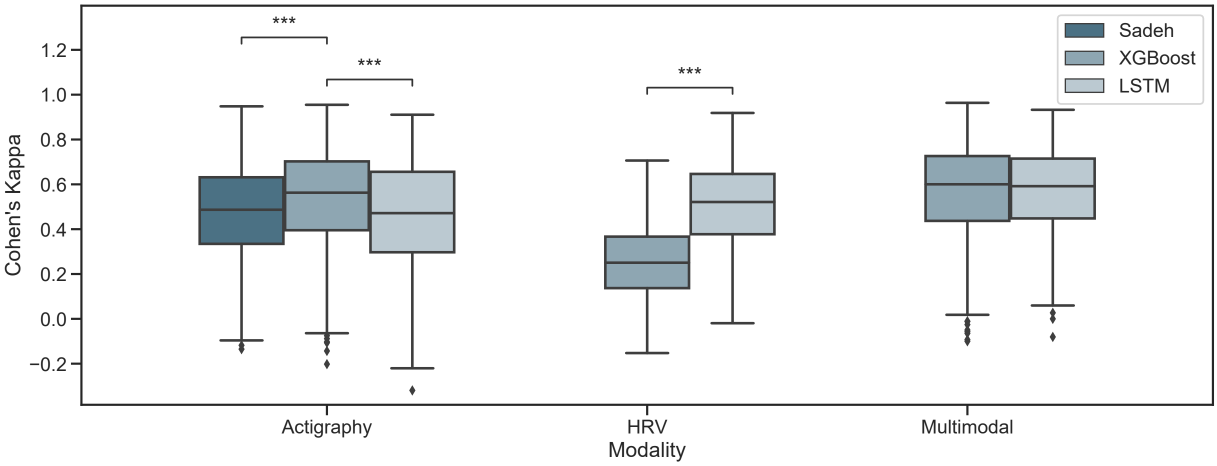

Figure 8.1: Best-performing algorithm of every category for mono- and multimodal approaches; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

For the actigraphy-based, monomodal approach, ***XGB*** achieved the highest accuracy of 81.0% correctly classified epochs, while all deep learning and machine learning approaches performed within the same range of $80 - 81\%$. In comparison, the heuristic algorithms performed slightly worse in a range from $77 - 79\%$. Among all heuristic algorithms, *Sazonov* poses an outlier performance of only 73.7% accuracy. Thereby, the comparably low recall and high precision of *Sazonov* show under-prediction of sleep. Furthermore, in terms of κ, accuracy, and F1 score, the algorithm of *Sazonov* performed worst of all algorithms. In terms of κ values, it is visible that the machine learning-based approaches perform better than both deep learning and heuristics for the actigraphy-based approach (Figure 8.1).

For the ***HRV***-based approach the deep learning models were able to achieve the highest classification performance (Table 8.1) with the ***LSTM*** as the best-performing model (accuracy: $80.2 \pm 10.5\%$). The ***TCN*** achieved an accuracy of 76.8%, while the machine learning models only achieved average accuracies of $69 - 71\%$.

Overall, the multimodal approach combining both modalities worked better than both monomodal approaches. Only the ***SVM*** could not profit from the ***HRV*** data resulting in an equal performance between the actigraphy-based and the multimodal approach. The best overall performance was obtained with ***LSTM***, achieving $83.7 \pm 9.6\%$ accuracy.

Table 8.2 presents sleep statistics in absolute values and ***Mean Absolute Error (MAE)***. As visible, most algorithms tend to overestimate sleep, expressed in a high ***SE*** and low ***WASO*** compared to the ground truth. Thereby, the machine learning algorithms in the ***HRV***-based approach led to the highest overestimation of sleep.

As visible in high ***SE***, and low ***WASO*** most of the algorithms underestimate wake phases occurring between sleep onset and wake onset. Participants had an average ***SE*** of 75.5%, while most algorithms predicted efficiencies of more than 80%. As visible from the other sleep metrics, the *Sazonov* algorithm overestimates wake phases during sleep with 228.5 ± 118.6 min instead of 152.4 ± 112.2 min.

Further sleep statistics are provided in Appendix B.1. It was observed that most of the algorithms underestimate the time to fall asleep, expressed in the ***SOL***. All ***HRV***-based approaches show ***MAE*** larger than $50\ min$, while the machine learning-based approaches in the actigraphy and multimodal approach show error rates of only $33 - 35$ min. In contrast, the ***NSD*** observed in the multimodal approach is lower than for the actigraphy-based approach for all algorithms except ***SVM***. That is caused by a slightly higher ***WASO***, which results in less overprediction of sleep and therefore the boost in performance.

Algorithm	**Accuracy [%]**	**Precision [%]**	**Recall [%]**	**F1-score [%]**	**Cohen's κ**
Always wake	33.2 ± 13.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.00 ± 0.00
Always sleep	66.8 ± 12.3	66.8 ± 12.3	100 ± 0.0	79.3 ± 10.2	0.00 ± 0.00
Ground truth	100 ± 0.0	100 ± 0.0	100 ± 0.0	100 ± 0.0	1.00 ± 0.00
Actigraphy					
Cole-Kripke	78.0 ± 10.1	78.4 ± 12.9	91.6 ± 7.4	83.7 ± 9.3	0.45 ± 0.20
Sadeh	78.6 ± 10.5	77.5 ± 13.0	**94.8 ± 5.8**	84.6 ± 9.2	0.45 ± 0.21
Sazonov	73.7 ± 10.0	81.4 ± 12.2	76.9 ± 13.3	78.1 ± 11.1	0.41 ± 0.20
Scripps-Clinic	77.8 ± 9.9	78.9 ± 12.9	90.1 ± 8.0	83.4 ± 9.3	0.46 ± 0.20
Webster	78.2 ± 10.1	78.6 ± 12.9	91.5 ± 7.8	83.8 ± 9.3	0.46 ± 0.21
AdaBoost	80.8 ± 10.7	81.3 ± 12.5	91.6 ± 11.1	85.3 ± 10.2	0.52 ± 0.23
MLP	80.9 ± 10.8	81.4 ± 12.5	91.5 ± 11.1	85.3 ± 10.3	0.53 ± 0.23
Random Forest	80.8 ± 10.8	81.2 ± 12.5	91.6 ± 10.8	85.3 ± 10.2	0.52 ± 0.23
SVM	80.4 ± 10.7	80.3 ± 12.7	93.0 ± 9.9	85.3 ± 9.8	0.51 ± 0.23
XGB	**81.0 ± 10.7**	81.6 ± 12.4	91.4 ± 11.0	85.4 ± 10.2	**0.53 ± 0.22**
LSTM	80.5 ± 11.1	**83.2 ± 12.1**	89.8 ± 11.8	85.6 ± 10.6	0.46 ± 0.23
TCN	80.4 ± 10.9	81.9 ± 12.3	92.2 ± 10.1	**86.0 ± 9.8**	0.44 ± 0.23
HRV					
AdaBoost	70.5 ± 11.9	71.5 ± 13.6	92.3 ± 12.9	79.2 ± 11.6	0.24 ± 0.16
MLP	71.1 ± 12.0	71.9 ± 13.6	92.3 ± 13.0	79.5 ± 11.8	0.26 ± 0.16
Random Forest	70.8 ± 12.1	71.9 ± 13.6	91.5 ± 14.3	79.1 ± 12.5	0.25 ± 0.16
SVM	69.2 ± 12.8	69.1 ± 13.6	**96.7 ± 10.0**	79.4 ± 11.3	0.16 ± 0.14
XGB	70.9 ± 12.0	72.2 ± 13.6	90.9 ± 14.3	79.0 ± 12.4	0.26 ± 0.16
LSTM	**80.2 ± 10.5**	**85.4 ± 12.7**	87.1 ± 12.3	**85.0 ± 10.3**	**0.50 ± 0.19**
TCN	76.7 ± 11.5	81.3 ± 13.9	87.1 ± 11.9	82.9 ± 11.0	0.41 ± 0.18
Multimodal					
AdaBoost	81.5 ± 10.4	82.4 ± 12.3	91.2 ± 11.4	85.6 ± 9.9	0.55 ± 0.22
MLP	81.9 ± 10.3	82.7 ± 12.3	91.4 ± 10.6	86.0 ± 9.7	0.56 ± 0.22
Random Forest	81.7 ± 10.4	82.3 ± 12.4	91.7 ± 10.4	85.9 ± 9.7	0.55 ± 0.22
SVM	80.4 ± 10.7	80.2 ± 12.7	**93.0 ± 9.8**	85.3 ± 9.8	0.51 ± 0.23
XGB	82.1 ± 10.2	83.0 ± 12.2	91.3 ± 10.6	86.1 ± 9.6	**0.57 ± 0.21**
LSTM	**83.7 ± 9.6**	**86.5 ± 12.0**	91.0 ± 9.9	**87.8 ± 9.1**	**0.57 ± 0.19**
TCN	80.9 ± 9.3	84.9 ± 12.0	88.2 ± 9.2	85.8 ± 8.8	0.50 ± 0.18

Table 8.1: Performance measures for the algorithm benchmarking on the ***MESA*** dataset sorted by input modality. The types of algorithms are separated by a line. The best-performing performance scores per algorithm group and modality, respectively, are written in bold.

Algorithm	***SE* [%]**	***MAE SE* [%]**	***WASO* [min]**	***MAE WASO* [min]**
Ground truth	75.8 ± 12.9	0.0 ± 0.0	152.4 ± 112.2	0.0 ± 0.0
Actigraphy				
Cole-Kripke	80.2 ± 12.0	9.8 ± 8.9	125.4 ± 92.8	68.5 ± 71.1
Sadeh	85.6 ± 10.0	12.2 ± 10.2	86.2 ± 70.8	81.4 ± 80.6
Sazonov	68.0 ± 13.4	11.4 ± 9.3	228.5 ± 118.6	102.5 ± 85.7
Scripps-Clinic	**81.1 ± 10.1**	**9.6 ± 8.9**	**129.2 ± 82.7**	**67.0 ± 71.2**
Webster	80.9 ± 11.6	10.0 ± 9.1	121.3 ± 90.1	69.5 ± 71.3
AdaBoost	81.8 ± 13.6	10.8 ± 9.5	112.0 ± 104.0	75.9 ± 78.8
MLP	81.6 ± 13.6	10.6 ± 9.5	113.6 ± 104.9	75.7 ± 78.3
Random Forest	81.8 ± 13.3	10.6 ± 9.4	111.9 ± 102.4	75.4 ± 78.0
SVM	83.5 ± 12.7	11.7 ± 9.8	100.0 ± 95.9	79.3 ± 81.3
XGB	81.6 ± 13.3	10.5 ± 9.4	111.8 ± 102.9	75.5 ± 77.2
LSTM	76.6 ± 14.8	10.1 ± 9.2	147.4 ± 122.7	91.0 ± 90.5
TCN	79.8 ± 13.5	10.8 ± 9.3	128.1 ± 111.3	92.1 ± 87.4
HRV				
AdaBoost	87.9 ± 13.8	18.1 ± 13.3	86.0 ± 102.7	114.3 ± 103.9
MLP	87.4 ± 13.7	17.4 ± 13.2	**89.8 ± 102.7**	**108.7 ± 103.1**
Random Forest	86.8 ± 15.1	17.9 ± 13.7	93.0 ± 113.2	114.2 ± 106.9
SVM	93.9 ± 10.5	21.5 ± 13.2	42.5 ± 76.7	128.3 ± 107.6
XGB	85.8 ± 14.9	17.1 ± 13.5	100.8 ± 112.4	109.1 ± 104.4
LSTM	**72.4 ± 14.3**	**11.8 ± 10.6**	211.4 ± 131.8	114.9 ± 108.8
TCN	75.9 ± 12.4	12.3 ± 10.9	211.5 ± 120.7	122.8 ± 102.1
Multimodal				
AdaBoost	80.5 ± 14.0	10.5 ± 9.6	122.2 ± 108.8	74.8 ± 78.5
MLP	80.3 ± 13.4	10.0 ± 9.3	**124.0 ± 105.0**	**74.2 ± 76.8**
Random Forest	81.0 ± 13.2	10.1 ± 9.3	117.1 ± 101.0	72.4 ± 76.5
SVM	83.6 ± 12.6	11.7 ± 9.8	99.4 ± 95.7	79.6 ± 81.3
XGB	80.3 ± 13.3	9.9 ± 9.1	122.2 ± 103.6	72.7 ± 75.7
LSTM	74.6 ± 13.6	10.0 ± 8.9	167.4 ± 119.9	91.0 ± 90.0
TCN	**73.6 ± 12.2**	**9.7 ± 8.3**	211.4 ± 115.2	106.2 ± 93.0

Table 8.2: ***SE*** and ***WASO*** with corresponding ***MAE*** measured on the benchmarking dataset.

The following results show the assessment of the influence of different study- and participant-specific influences on classification performance. Statistical tests were applied for this purpose and can be found in Appendix B.2. Thereby, only the best-performing algorithm of each group (heuristic, machine learning, and deep learning) and each modality, according to accuracy, were analyzed.

Signal Quality

It was examined, whether the signal quality of actigraphy and ***PSG*** recording influenced the performance. Figure 8.2 shows the performance of the algorithms in all three modalities distinguished between good and bad ***PSG***-signal quality. As it is visible, all algorithms achieved higher classification accuracy for high-quality ***PSG*** data (Figure 8.2). Concurrently, higher actigraph signal quality led to a significantly ($p < 0.05$) better classification accuracy. The statistical tests are provided in Appendix (Table B.5).

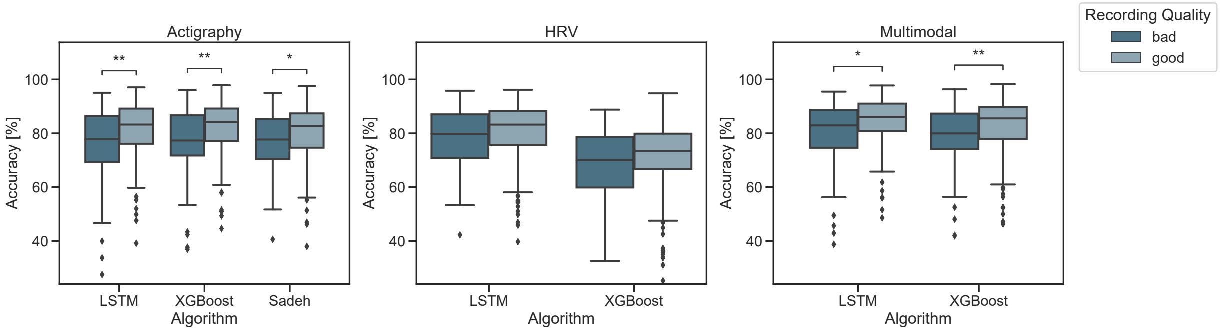

Figure 8.2: Classification performance dependent on ***PSG*** Recording Quality; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Gender

Figure 8.3 depicts differences in classification performance between male and female participants. As visible, both the actigraph and the multimodal approaches achieved significantly higher classification performance for female participants (see Appendix, Table B.7).

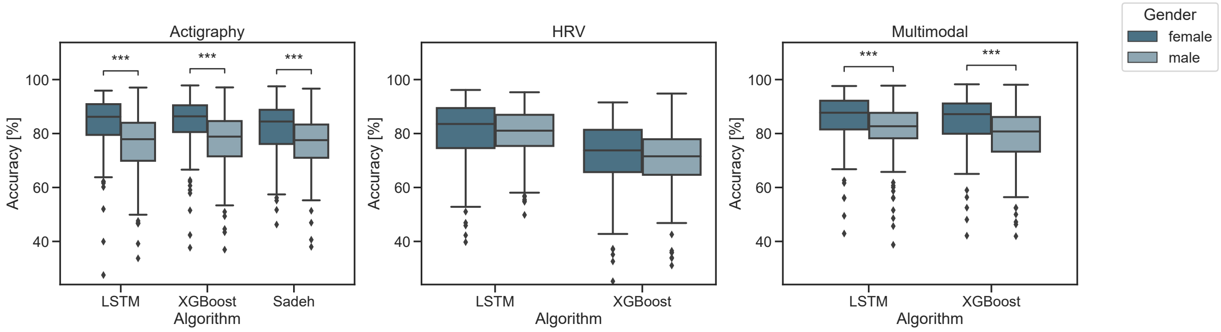

Figure 8.3: Performance dependent on gender of the participants; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Race

Furthermore, the sleep/wake estimation was evaluated towards different races. Thereby, no significant differences were found (see Appendix, Table B.11).

Extra Workload

Another examination was performed to determine the influence of extra work (more than 5 h per week). Results show that individuals with more than 5 h of extra work showed better classification performance (see Figure 8.4 and Appendix, Table B.10).

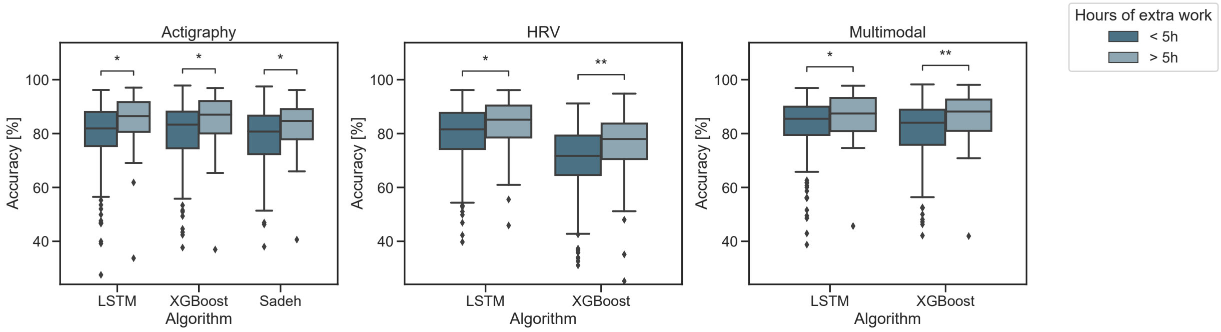

Figure 8.4: Influence of more than 5 h of extra work per week; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Sleep Quality

The sleep quality score obtained from the ***WHIIRS*** questionnaire led to no significant differences between individuals with good or bad sleep (see Appendix, Table B.9).

Sleep Diseases

In addition, it was investigated whether sleep-related diseases influence classification performance. The ***MESA*** dataset included information about the diagnosis of ***RLS***, ***INS***, and sleep apnea. However, no significant effects were found (see Appendix, Table B.8).

Age

Figure 8.5 shows the relationship between age and accuracy for ***XGB***, all trained on different modalities. It is visible that the classification performance does not depend on age for the actigraphy- and only little for the multimodal approach but increasing age tends to decrease the accuracy for the ***HRV***-based approach.

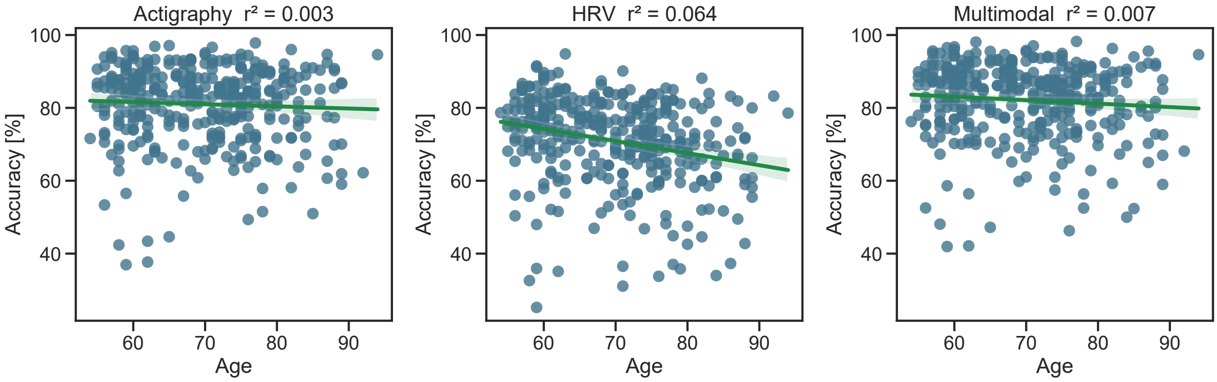

Figure 8.5: Sleep/wake detection accuracy as a function of age. *Green line*: Linear regression slope with 95% confidence interval; predictor: age; dependent variable: accuracy.

AHI

Figure 8.6 presents the relationship between ***AHI***, which serves as an indicator for apnea and hypopnea frequency and severity, (Section 3.2) and classification performance with ***XGB***. The correlation indicates worse classification for individuals with higher ***AHI***, with a similar effect for all three modalities.

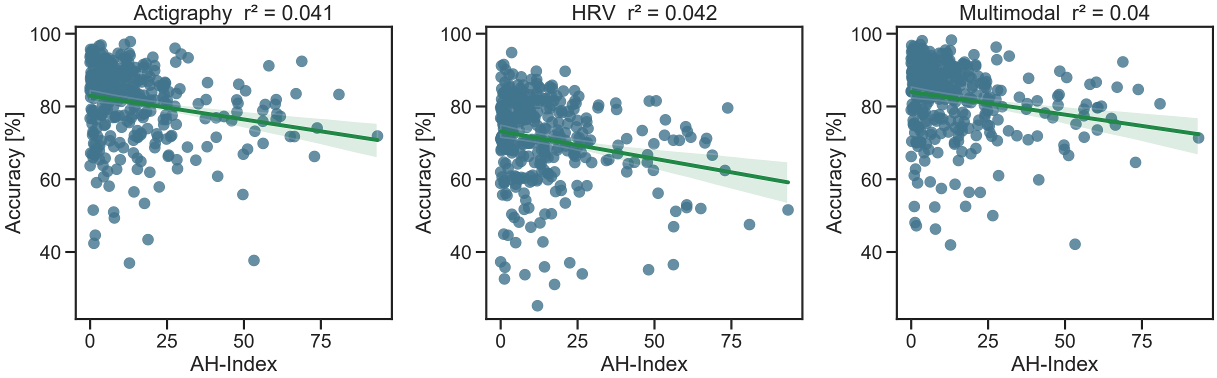

Figure 8.6: Sleep/wake detection accuracy as a function of ***AHI***. *Green line*: Linear regression slope with 95% confidence interval; predictor: ***AHI***; dependent variable: accuracy.

8.2 Classification on Real-World Data

The second major part of this thesis was to evaluate the algorithms used for the benchmarking approach on a real-world data set. For that, different analysis approaches were conducted that will be presented in this section: First, results of applying machine learning models, pre-trained on the benchmark dataset, on the real-world dataset, are outlined. Second, all algorithms were re-trained on the real-world dataset and evaluated. For both evaluations, raw ***IMU*** data were converted to activity counts to allow comparison with the benchmark dataset. In the third evaluation, it was assessed whether classification performance on the real-world dataset can be further improved by using features extracted from raw ***IMU*** data instead of activity counts.

Since the dataset acquired for this work is highly imbalanced, mostly containing sleep, ranking the results by accuracy is not very meaningful. Therefore, the performance evaluation was mainly focused on Cohen's κ.

8.2.1 Classification using Pre-trained Pipelines

Table 8.3 provides an overview of the performance, sorted according to input modality. Machine- and deep learning algorithms are separated by a line. The best performance of every metric and every modality is written in bold. Because the heuristic algorithms are not trainable, only the machine-learning- and deep learning algorithms presented in Chapter 5 were used.

Comparing the different modalities, the deep learning algorithms perform worse than machine learning algorithms in all three approaches. While the ***LSTM*** in the motion-based approach reached a κ of 0.12 ± 0.15, all other approaches resulted in κ values very close to zero or negative, implying a classification performance worse than random guessing [Can13, Sui19]. Regarding κ, the ***XGB*** performed best in all modalities, with the best performance obtained for the actigraphy-based and multimodal approach. The ***HRV***-based approach performed worse for all machine learning algorithms. Furthermore, they were not able to profit from ***HRV*** data and reached their best κ values in the actigraphy-based approach.

Due to the imbalance of sleep and wake samples, all machine learning algorithms reached accuracies higher than 85%. In contrast, the deep learning algorithms predicted more wake samples, resulting in considerably worse classification performance.

Algorithm	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]	Cohen's κ
Always wake	9.2 ± 6.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.00 ± 0.00
Always sleep	90.8 ± 6.9	90.8 ± 6.9	100 ± 0.0	95.0 ± 4.1	0.00 ± 0.00
Ground truth	100 ± 0.0	100 ± 0.0	100 ± 0.0	100 ± 0.0	1.00 ± 0.00
Actigraphy					
AdaBoost	91.1 ± 7.7	93.0 ± 6.4	97.4 ± 5.5	95.0 ± 5.0	0.29 ± 0.24
MLP	91.0 ± 7.5	93.0 ± 6.4	97.3 ± 5.1	94.9 ± 4.8	0.30 ± 0.23
Random Forest	90.9 ± 7.4	92.9 ± 6.4	97.5 ± 4.8	94.9 ± 4.7	0.28 ± 0.23
SVM	90.1 ± 7.6	93.3 ± 6.3	96.0 ± 5.9	94.4 ± 5.0	0.30 ± 0.24
XGBoost	**91.2 ± 7.5**	93.1 ± 6.6	**97.7 ± 4.8**	**95.1 ± 4.7**	**0.32 ± 0.23**
LSTM	86.1 ± 7.8	**98.4 ± 3.3**	87.1 ± 8.3	92.1 ± 4.9	0.12 ± 0.15
TCN	3.2 ± 4.0	40.4 ± 49.9	0.3 ± 0.4	0.5 ± 0.9	0.00 ± 0.00
HRV					
AdaBoost	87.3 ± 9.6	91.4 ± 6.1	94.3 ± 10.0	92.5 ± 7.2	0.06 ± 0.08
MLP	86.4 ± 9.9	91.5 ± 6.1	93.1 ± 10.6	92.0 ± 7.4	0.06 ± 0.08
Random Forest	86.5 ± 10.4	91.5 ± 6.2	93.1 ± 11.3	91.9 ± 8.0	0.06 ± 0.08
SVM	**89.1 ± 7.3**	91.2 ± 6.3	**97.2 ± 5.0**	**93.9 ± 4.7**	0.04 ± 0.08
XGBoost	85.9 ± 10.7	91.6 ± 6.1	92.4 ± 11.9	91.5 ± 8.3	**0.07 ± 0.08**
LSTM	73.8 ± 17.9	**97.9 ± 3.8**	74.6 ± 18.9	83.2 ± 14.4	0.03 ± 0.12
TCN	66.7 ± 20.7	97.7 ± 3.6	67.4 ± 21.5	77.4 ± 18.7	0.01 ± 0.05
Multimodal					
AdaBoost	90.3 ± 7.6	92.9 ± 6.3	96.7 ± 5.4	94.6 ± 4.9	0.26 ± 0.22
MLP	89.5 ± 7.2	92.7 ± 6.1	95.8 ± 5.8	94.0 ± 4.8	0.22 ± 0.19
Random Forest	90.8 ± 7.5	92.8 ± 6.2	**97.1 ± 5.1**	**94.8 ± 4.9**	0.26 ± 0.22
SVM	90.2 ± 7.6	93.3 ± 6.3	96.1 ± 5.9	94.5 ± 5.0	0.30 ± 0.24
XGBoost	**90.6 ± 7.2**	93.2 ± 5.7	96.4 ± 5.9	94.6 ± 4.8	**0.31 ± 0.19**
LSTM	12.2 ± 11.8	86.3 ± 33.2	10.1 ± 12.4	16.2 ± 17.9	0.01 ± 0.03
TCN	18.5 ± 20.8	**94.5 ± 12.5**	16.5 ± 21.6	23.4 ± 27.1	0.00 ± 0.04

Table 8.3: Performance measures for the sleep/wake classification of the real-world dataset using pre-trained pipelines from the benchmarking approach on the ***MESA*** dataset.

Algorithm	***SE*** **[%]**	***MAE SE*** **[%]**	***WASO*** **[min]**	***MAE WASO*** **[min]**
Ground truth	90.8 ± 6.9	0.0 ± 0.0	24.4 ± 40.1	0.0 ± 0.0
Actigraphy				
AdaBoost	95.1 ± 6.0	6.1 ± 7.3	30.1 ± 32.3	30.2 ± 35.5
MLP	94.9 ± 5.6	5.9 ± 7.0	34.0 ± 35.5	33.3 ± 39.1
Random Forest	95.2 ± 5.3	5.9 ± 7.0	**28.6 ± 28.7**	**28.9 ± 33.2**
SVM	**93.4 ± 6.4**	**5.6 ± 7.0**	42.4 ± 39.1	38.1 ± 39.0
XGBoost	95.3 ± 5.3	6.0 ± 7.1	25.5 ± 26.5	29.2 ± 32.8
LSTM	85.9 ± 8.8	8.4 ± 7.0	100.9 ± 56.5	94.9 ± 59.9
TCN	0.3 ± 0.4	92.7 ± 3.9	79.1 ± 184.5	87.7 ± 182.3
HRV				
AdaBoost	93.7 ± 10.8	7.9 ± 6.0	50.7 ± 74.0	50.5 ± 63.1
MLP	92.4 ± 11.3	8.1 ± 6.4	62.0 ± 78.9	58.9 ± 67.6
Random Forest	92.4 ± 12.0	8.0 ± 6.9	61.6 ± 84.9	57.5 ± 72.7
SVM	**96.8 ± 6.0**	**7.4 ± 5.6**	25.7 ± 40.5	34.3 ± 43.3
XGBoost	91.7 ± 12.5	8.3 ± 7.3	68.2 ± 88.9	62.5 ± 76.3
LSTM	74.1 ± 18.8	20.0 ± 16.3	**17.5 ± 11.9**	**20.4 ± 19.7**
TCN	67.0 ± 21.7	26.4 ± 20.7	275.8 ± 182.3	267.2 ± 187.1
Multimodal				
AdaBoost	94.5 ± 5.9	6.0 ± 6.6	37.7 ± 36.8	34.7 ± 37.3
MLP	93.8 ± 6.3	6.0 ± 5.7	45.9 ± 42.9	40.9 ± 43.7
Random Forest	94.9 ± 5.7	5.7 ± 5.8	**31.0 ± 34.3**	**28.1 ± 35.8**
SVM	93.5 ± 6.4	5.6 ± 7.0	41.1 ± 38.5	37.3 ± 38.8
XGBoost	**93.9 ± 7.1**	**5.5 ± 5.7**	40.0 ± 43.6	36.1 ± 43.4
LSTM	10.3 ± 12.7	82.7 ± 12.6	384.1 ± 238.3	377.4 ± 239.1
TCN	16.6 ± 21.3	76.4 ± 21.4	521.1 ± 234.0	512.5 ± 231.5

Table 8.4: ***SE*** and ***WASO*** with corresponding ***MAE*** measured on the real-world dataset with classification models trained on the benchmark dataset.

The sleep statistics (Table 8.4) show a strong overprediction of wake phases for the ***TCN*** in all modalities and the ***LSTM*** in the multimodal approach expressed in a low ***SE*** and high ***WASO***. In contrast, the machine learning algorithms systematically overestimated sleep, leading to overall lower ***SOL*** and higher ***NSD*** (see Appendix, Table B.2). Moreover, ***WASO*** was systematically overestimated by all machine learning algorithms.

8.2.2 Actigraphy-based Sleep/Wake Classification

The second real-world evaluation was to train and test the algorithms presented in Chapter 5 using the activity counts extracted from raw ***IMU*** data. Thereby, the machine learning algorithms reached higher κ values compared to the heuristic algorithms, while the deep learning algorithms performed worst (Table 8.2.2). Comparing the mono- and multimodal approaches, the motion-

Algorithm	**Accuracy [%]**	**Precision [%]**	**Recall [%]**	**F1-score [%]**	**Cohen's κ [%]**
Always wake	9.2 ± 6.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.00 ± 0.00
Always sleep	90.8 ± 6.9	90.8 ± 6.9	100 ± 0.0	95.0 ± 4.1	0.00 ± 0.00
Ground truth	100 ± 0.0	100 ± 0.0	100 ± 0.0	100 ± 0.0	1.00 ± 0.00
Actigraphy					
Cole-Kripke	87.6 ± 7.8	93.3 ± 6.1	93.0 ± 7.1	92.9 ± 5.3	0.25 ± 0.21
Sadeh	86.9 ± 8.3	94.8 ± 6.0	90.5 ± 7.9	92.3 ± 5.9	0.35 ± 0.19
Sazonov	79.6 ± 7.0	94.1 ± 5.7	82.8 ± 6.7	87.8 ± 5.1	0.18 ± 0.14
Scripps-Clinic	87.0 ± 7.5	93.4 ± 6.0	92.2 ± 6.6	92.5 ± 5.1	0.24 ± 0.19
Webster	85.3 ± 8.2	93.7 ± 5.8	89.7 ± 8.1	91.4 ± 5.7	0.23 ± 0.20
AdaBoost	91.2 ± 7.7	92.8 ± 6.6	98.0 ± 5.3	95.1 ± 4.9	0.28 ± 0.25
MLP	91.3 ± 7.5	92.7 ± 6.6	98.2 ± 4.5	95.2 ± 4.6	0.30 ± 0.24
Random Forest	91.5 ± 7.1	92.8 ± 6.7	98.3 ± 3.4	95.3 ± 4.3	0.30 ± 0.25
SVM	90.9 ± 8.0	93.3 ± 6.5	96.9 ± 6.1	94.8 ± 5.2	0.30 ± 0.28
XGBoost	91.8 ± 7.3	93.2 ± 6.7	98.1 ± 4.2	95.4 ± 4.6	**0.38 ± 0.26**
LSTM	**96.2 ± 4.5**	**97.5 ± 4.0**	98.7 ± 2.5	**98.0 ± 2.4**	0.15 ± 0.22
TCN	**96.2 ± 4.1**	97.3 ± 4.0	**98.8 ± 1.6**	**98.0 ± 2.2**	0.12 ± 0.19
HRV					
AdaBoost	90.5 ± 6.8	91.0 ± 6.8	99.4 ± 1.7	94.8 ± 4.1	0.03 ± 0.05
MLP	90.4 ± 7.4	91.0 ± 6.9	99.2 ± 3.2	94.8 ± 4.5	0.04 ± 0.10
Random Forest	90.3 ± 7.4	91.0 ± 6.8	99.0 ± 3.5	94.7 ± 4.5	**0.05 ± 0.09**
SVM	85.2 ± 13.5	90.7 ± 7.2	92.7 ± 13.8	91.0 ± 10.9	**0.05 ± 0.11**
XGBoost	90.6 ± 7.0	91.0 ± 6.8	99.4 ± 2.2	94.9 ± 4.2	0.04 ± 0.08
LSTM	96.5 ± 4.0	97.2 ± 3.8	99.2 ± 1.9	98.2 ± 2.1	0.03 ± 0.11
TCN	**96.8 ± 3.8**	**97.2 ± 3.8**	**99.6 ± 1.0**	**98.3 ± 2.0**	**0.05 ± 0.09**
Multimodal					
AdaBoost	91.0 ± 7.4	92.9 ± 6.6	97.5 ± 4.7	94.9 ± 4.7	0.28 ± 0.24
MLP	91.1 ± 7.3	92.7 ± 6.6	97.8 ± 4.5	95.0 ± 4.5	0.28 ± 0.24
Random Forest	91.6 ± 7.0	92.8 ± 6.7	98.4 ± 3.4	95.4 ± 4.3	0.30 ± 0.25
SVM	90.9 ± 8.0	93.3 ± 6.5	96.9 ± 6.1	94.8 ± 5.2	0.30 ± 0.28
XGBoost	91.6 ± 7.5	93.3 ± 6.4	97.7 ± 5.3	95.2 ± 4.9	**0.38 ± 0.24**
LSTM	**96.1 ± 4.6**	**97.5 ± 3.9**	**98.5 ± 2.6**	**98.0 ± 2.5**	0.19 ± 0.23
TCN	95.3 ± 4.0	97.2 ± 3.9	98.0 ± 2.0	97.6 ± 2.1	0.02 ± 0.05

Table 8.5: Performance measures of sleep/wake classification using activity counts extracted from ***IMU*** data.

Algorithm	***SE* [%]**	***MAE SE* [%]**	***WASO* [min]**	***MAE WASO* [min]**
Ground truth	90.8 ± 6.9	0.0 ± 0.0	24.4 ± 40.1	0.0 ± 0.0
Actigraphy				
Cole-Kripke	90.4 ± 7.5	6.3 ± 6.6	72.0 ± 51.6	59.6 ± 48.1
Sadeh	86.5 ± 8.4	7.1 ± 7.4	81.9 ± 60.3	66.5 ± 57.6
Sazonov	79.9 ± 7.1	11.9 ± 7.5	173.4 ± 59.7	149.0 ± 59.9
Scripps-Clinic	89.6 ± 7.0	5.9 ± 6.3	79.6 ± 50.4	64.1 ± 48.4
Webster	86.9 ± 8.7	7.2 ± 7.6	102.8 ± 64.0	83.9 ± 62.0
AdaBoost	95.8 ± 5.7	6.6 ± 7.5	26.2 ± 28.2	31.6 ± 35.6
MLP	96.1 ± 4.8	6.6 ± 7.0	27.5 ± 26.9	29.2 ± 33.1
Random Forest	96.2 ± 4.0	6.4 ± 6.7	24.7 ± 27.2	28.2 ± 34.3
SVM	94.3 ± 6.6	6.2 ± 7.4	37.3 ± 36.3	35.6 ± 37.7
XGB	95.5 ± 4.8	6.1 ± 6.5	27.1 ± 31.9	31.5 ± 37.4
LSTM	98.3 ± 3.0	$\mathbf{5.6 \pm 3.8}$	14.4 ± 23.9	19.3 ± 29.8
TCN	98.6 ± 1.8	$\mathbf{5.6 \pm 4.1}$	11.6 ± 14.7	$\mathbf{16.9 \pm 24.6}$
HRV				
AdaBoost	99.2 ± 1.9	8.5 ± 6.6	6.9 ± 16.3	24.2 ± 38.4
MLP	99.0 ± 3.2	8.6 ± 6.6	8.4 ± 28.7	25.2 ± 39.8
Random Forest	98.7 ± 3.5	8.4 ± 6.7	11.3 ± 31.3	26.9 ± 40.6
SVM	92.6 ± 13.3	9.4 ± 10.8	66.7 ± 125.3	69.2 ± 116.5
XGB	99.2 ± 2.3	8.6 ± 6.7	7.1 ± 21.0	24.2 ± 38.6
LSTM	99.1 ± 2.0	$\mathbf{6.1 \pm 4.4}$	8.1 ± 17.1	12.9 ± 21.6
TCN	99.4 ± 1.2	6.4 ± 4.1	5.2 ± 10.5	$\mathbf{10.9 \pm 20.5}$
Multimodal				
AdaBoost	95.3 ± 5.2	6.2 ± 6.6	33.2 ± 35.1	32.8 ± 37.5
MLP	95.8 ± 4.9	6.4 ± 6.9	32.4 ± 33.0	31.5 ± 35.4
Random Forest	96.3 ± 4.1	6.5 ± 6.6	24.4 ± 27.3	29.8 ± 34.0
SVM	94.3 ± 6.6	6.2 ± 7.4	37.3 ± 36.3	35.6 ± 37.7
XGB	95.1 ± 5.8	5.8 ± 6.3	30.2 ± 34.8	32.7 ± 38.1
LSTM	98.1 ± 3.1	5.4 ± 3.7	15.7 ± 24.1	$\mathbf{19.0 \pm 28.5}$
TCN	97.9 ± 2.1	$\mathbf{5.0 \pm 4.5}$	17.1 ± 17.7	21.7 ± 21.6

Table 8.6: ***SE*** and ***WASO*** with corresponding ***MAE*** measured on the real-world dataset using the extracted activity counts from ***IMU*** data.

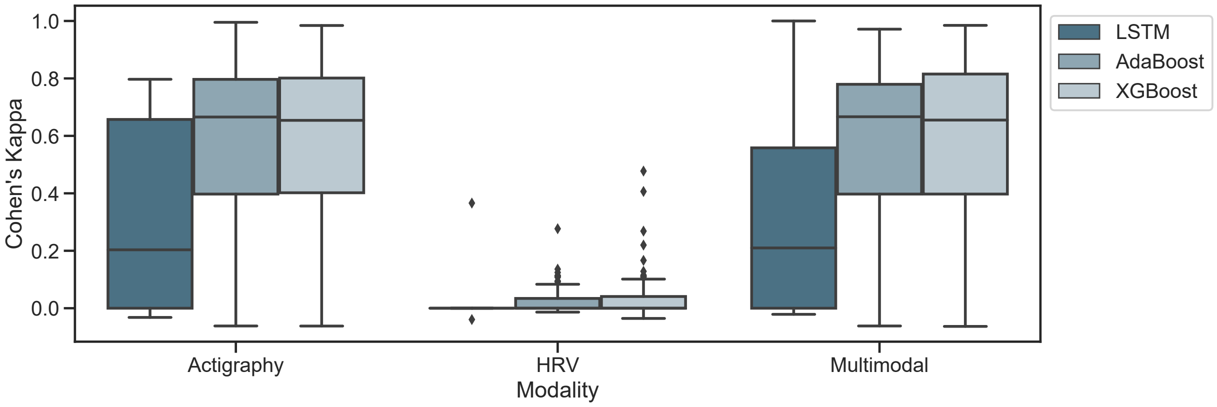

Figure 8.7: Best performing algorithm of machine learning and deep learning algorithms for mono- and multimodal approaches.

based approach worked similarly compared to the multimodal approach, while the cardiac-based approach performed worst. The highest performance was observed for the ***XGB*** algorithm with $\kappa = 0.38$ for both the motion-based and the multimodal approach.
Except *Sazonov*, the heuristic algorithms predicted a similar ***SE*** as measured by the sleep mat, but a higher ***WASO***. In contrast, the motion-based and multimodal machine learning algorithms had a good prediction of ***WASO***, but an overprediction of ***SE***. The cardiac-based approach, which performed worst, suffered from a strong systematic overprediction of sleep (see Table 8.6). Moreover, all algorithms in all modalities systematically underestimated ***SOL*** (see Appendix, Table B.3).

8.2.3 IMU-based Sleep/Wake Classification

The third evaluation of real-world data was to assess whether sleep/wake classification using raw ***IMU*** data instead of aggregated activity counts can yield better performance than using aggregated activity counts. Table 8.7 presents the results of this assessment using the algorithms presented in Chapter 5. As already observed for the real-world ***IMU*** data converted to activity counts (Section 8.2.2), the deep learning algorithms performed poorly compared to the machine learning algorithms. The ***IMU*** and multimodal approaches performed similarly with κ ranging between 0.54 and 0.59 for the machine learning algorithms as well as 0.26 and 0.31 for the deep learning algorithms. In comparison, the ***HRV***-based approach performed considerably worse with κ ranging between 0.00 and 0.10 for both types of algorithms. The best performing algorithms were ***XGB*** and ***AdaBoost***, reaching κ values of 0.59 ± 0.28 and 0.59 ± 0.27, in the ***IMU***- and the

Algorithm	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]	Cohen's κ
Always wake	9.2 ± 6.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.00 ± 0.00
Always sleep	90.8 ± 6.9	90.8 ± 6.9	100 ± 0.0	95.0 ± 4.1	0.00 ± 0.00
Ground truth	100 ± 0.0	100 ± 0.0	100 ± 0.0	100 ± 0.0	1.00 ± 0.00
Actigraphy					
AdaBoost	93.4 ± 7.0	95.2 ± 6.7	97.7 ± 3.7	96.3 ± 4.2	**0.59 ± 0.27**
MLP	93.4 ± 6.8	94.9 ± 6.5	98.0 ± 3.7	96.3 ± 4.1	0.56 ± 0.26
Random Forest	92.5 ± 12.0	93.7 ± 12.3	97.2 ± 11.5	95.2 ± 11.4	0.58 ± 0.28
SVM	93.6 ± 7.0	94.9 ± 6.7	98.2 ± 3.2	96.4 ± 4.2	0.58 ± 0.28
XGBoost	92.7 ± 11.7	93.7 ± 12.3	97.4 ± 11.2	95.3 ± 11.3	**0.59 ± 0.28**
LSTM	**95.9 ± 4.7**	96.4 ± 4.2	**99.2 ± 1.2**	**97.7 ± 2.6**	0.30 ± 0.34
TCN	95.3 ± 4.4	**96.9 ± 3.5**	98.0 ± 2.6	97.4 ± 2.5	0.28 ± 0.32
HRV					
AdaBoost	90.5 ± 6.8	91.0 ± 6.8	99.4 ± 1.7	94.8 ± 4.1	0.03 ± 0.05
MLP	90.6 ± 7.2	91.0 ± 6.8	99.3 ± 2.4	94.8 ± 4.3	0.04 ± 0.10
Random Forest	90.2 ± 7.5	91.0 ± 6.8	98.9 ± 3.8	94.6 ± 4.5	0.05 ± 0.09
SVM	81.9 ± 22.6	88.9 ± 15.8	88.1 ± 27.8	85.9 ± 24.7	0.02 ± 0.08
XGBoost	90.6 ± 6.9	91.0 ± 6.8	99.5 ± 2.0	94.9 ± 4.1	0.04 ± 0.08
LSTM	96.5 ± 4.0	97.2 ± 3.8	99.2 ± 1.9	98.2 ± 2.1	0.03 ± 0.11
TCN	**96.8 ± 3.8**	**97.3 ± 3.7**	**99.5 ± 1.4**	**98.3 ± 2.0**	**0.10 ± 0.23**
Multimodal					
AdaBoost	93.5 ± 6.9	95.1 ± 6.6	98.0 ± 3.5	96.3 ± 4.1	**0.59 ± 0.27**
MLP	92.7 ± 7.0	94.9 ± 6.4	97.2 ± 4.2	95.9 ± 4.3	0.54 ± 0.24
Random Forest	92.6 ± 11.8	93.8 ± 12.3	97.2 ± 11.4	95.3 ± 11.3	0.58 ± 0.28
SVM	93.6 ± 7.0	94.9 ± 6.7	98.2 ± 3.2	96.4 ± 4.2	0.58 ± 0.28
XGBoost	92.6 ± 11.9	93.7 ± 12.2	97.3 ± 11.4	95.3 ± 11.3	**0.59 ± 0.28**
LSTM	**94.5 ± 5.9**	**97.1 ± 3.6**	96.8 ± 4.2	97.0 ± 3.5	0.31 ± 0.36
TCN	95.4 ± 4.4	96.8 ± 3.8	**98.3 ± 1.9**	**97.5 ± 2.5**	0.26 ± 0.34

Table 8.7: Performace metrics of the ***IMU***-based sleep/wake detection.

multimodal approach. Figure 8.7 shows the best performing algorithm or each category (***XGB***, ***AdaBoost***, and ***LSTM***) for all three modalities.

Results of sleep statistics (Table 8.2.3) show that all algorithms of all modalities tend to underestimate ***SOL*** and overestimate ***SE***, while the ***HRV***-based approach as well as the deep learning approaches for all modalities show the most deviation. In general, the machine learning algorithms show comparably small deviations in ***SOL***, ***SE***, ***WASO***, and ***NSD***. Due to the best overall agreement rates, the more detailed analysis of sleep metrics is only presented for the ***IMU***-based classification. Moreover, only 50 out of 80 nights were considered for this analysis since the *DiPsyLab* study did not include the same set of questionnaires in the study protocol, thus had to be excluded for the more detailed analysis. Results of statistical test for the best-performing algorithms, ***XGB*** and ***AdaBoost***, are provided in Appendix B.2.2.

Algorithm	***SE* [%]**	***MAE SE* [%]**	***WASO* [min]**	***MAE WASO* [min]**
Ground truth	90.8 ± 6.9	0.0 ± 0.0	24.4 ± 40.1	0.0 ± 0.0
Actigraphy				
AdaBoost	93.2 ± 4.6	4.7 ± 6.4	21.5 ± 32.2	25.0 ± 32.6
MLP	**93.7 ± 4.7**	**4.6 ± 6.6**	25.7 ± 34.2	27.2 ± 34.2
Random Forest	93.1 ± 11.3	6.1 ± 12.0	15.4 ± 24.8	24.6 ± 34.1
SVM	94.0 ± 4.4	4.8 ± 6.5	15.5 ± 29.3	23.4 ± 32.9
XGBoost	93.2 ± 11.0	5.9 ± 11.8	16.9 ± 27.4	24.6 ± 33.1
LSTM	97.4 ± 4.6	5.8 ± 3.4	**23.0** ± **45**.5	**21.8** ± **33.7**
TCN	**95.8** ± **5.9**	**4.6** ± **3.2**	27.5 ± 39.5	23.2 ± 32.8
HRV				
AdaBoost	99.2 ± 1.9	8.5 ± 6.6	6.9 ± 16.3	24.2 ± 38.4
MLP	99.0 ± 2.6	8.3 ± 6.7	8.4 ± 23.4	23.9 ± 37.9
Random Forest	98.6 ± 3.8	8.4 ± 6.7	11.8 ± 33.6	27.9 ± 41.4
SVM	88.0 ± 27.8	16.5 ± 22.2	73.8 ± 187.8	84.5 ± 172.7
XGBoost	99.3 ± 2.1	8.7 ± 6.7	6.2 ± 19.2	24.4 ± 38.4
LSTM	**99.1** ± **2.0**	**6.1** ± **4.4**	8.1 ± 17.1	12.9 ± 21.6
TCN	99.4 ± 1.2	6.4 ± 4.1	5.2 ± 10.5	**10.9** ± **20.5**
Multimodal				
AdaBoost	93.5 ± 4.4	4.7 ± 6.5	17.9 ± 28.5	25.1 ± 32.7
MLP	92.9 ± 5.0	4.3 ± 6.2	36.3 ± 36.7	30.3 ± 28.2
Random Forest	93.0 ± 11.3	6.0 ± 12.0	16.7 ± 26.5	24.9 ± 32.9
SVM	94.0 ± 4.4	4.8 ± 6.5	15.5 ± 29.3	23.4 ± 32.9
XGBoost	93.0 ± 11.2	5.9 ± 11.9	17.4 ± 27.8	24.6 ± 33.1
LSTM	**94.5** ± **6.5**	**3.4** ± **2.9**	**39.0** ± **57**.0	**21.4** ± **35.6**
TCN	96.2 ± 5.7	4.7 ± 3.4	25.5 ± 43.5	25.1 ± 41.2

Table 8.8: ***SE*** and ***WASO*** with corresponding ***MAE*** measured on the real-world dataset using ***IMU*** data.

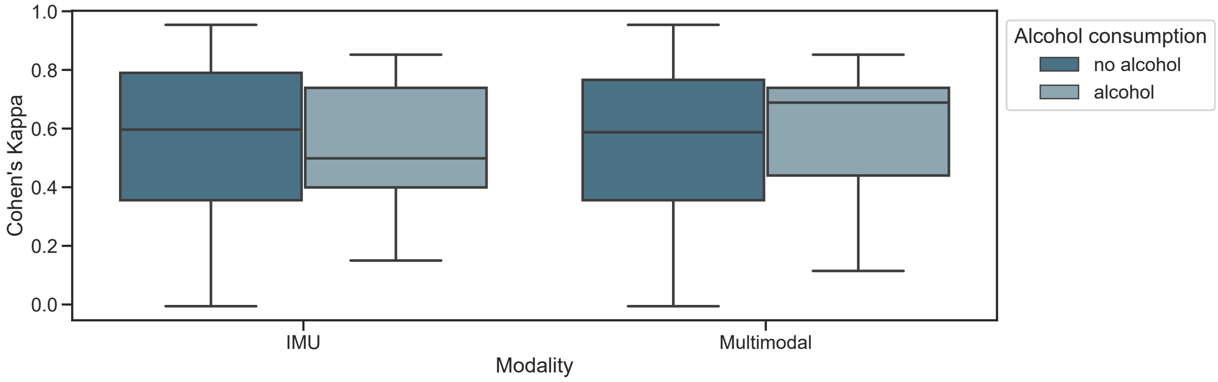

Figure 8.8: Evaluation of ***IMU***-based sleep/wake classification by ***AdaBoost*** with or without the influence of alcohol.

One question the participants were asked was whether they had consumed more than two alcoholic beverages before going to bed. It was found that the consumption of alcohol did not significantly influence the classification performance (Figure 8.8).

Figure 8.9 depicts the classification performance of ***AdaBoost*** categorized by subjective sleep quality. The classification performance significantly increased with better subjective sleep quality of the last night for the actigraphy-based approach. Additionally, the performance was evaluated using the ***PSQI*** score (Section 3.2). However, no significant results were found (Appendix, Tables B.14 and B.15, respectively)

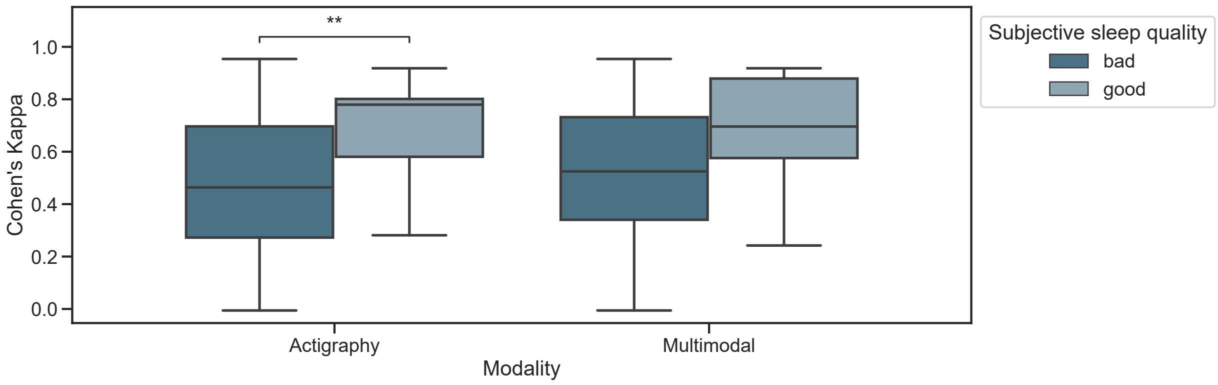

Figure 8.9: Evaluation of ***IMU***-based sleep/wake classification by ***AdaBoost*** depending on subjective sleep quality; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

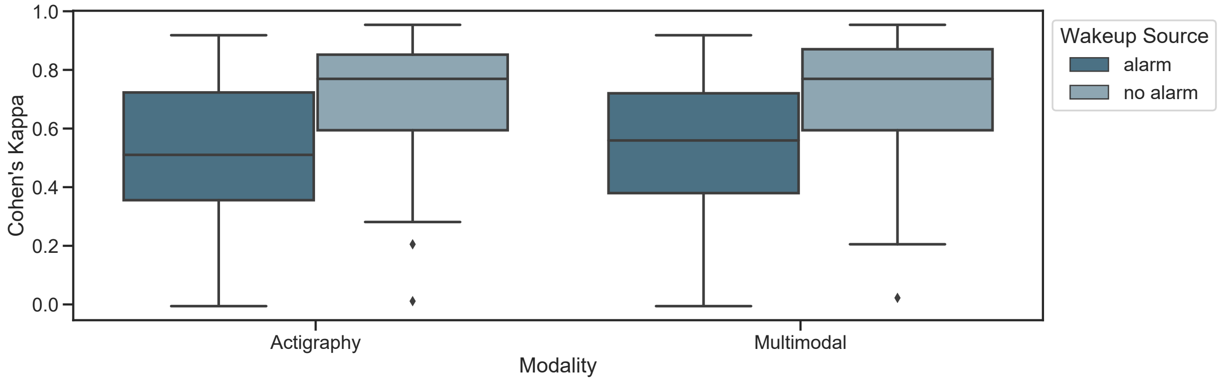

Figure 8.10: Evaluation of ***IMU***-based sleep/wake classification by ***AdaBoost*** with or without alarm clock in the morning.

Additionally, it was assessed whether the mode of awakening (alarm vs. no alarm) influenced sleep/wake classification (see Figure 8.10). The results indicate better classification on nights without awakening by an alarm clock the next morning, but the results are not significant (Appendix, Table B.16). Similarly, differences in current profession (student vs. employee) did not influence classification performance (Appendix, Table B.17). In contrast, ***BMI*** seemed to influence classification performance with higher ***BMI*** leading to poorer results (Figure 8.11).

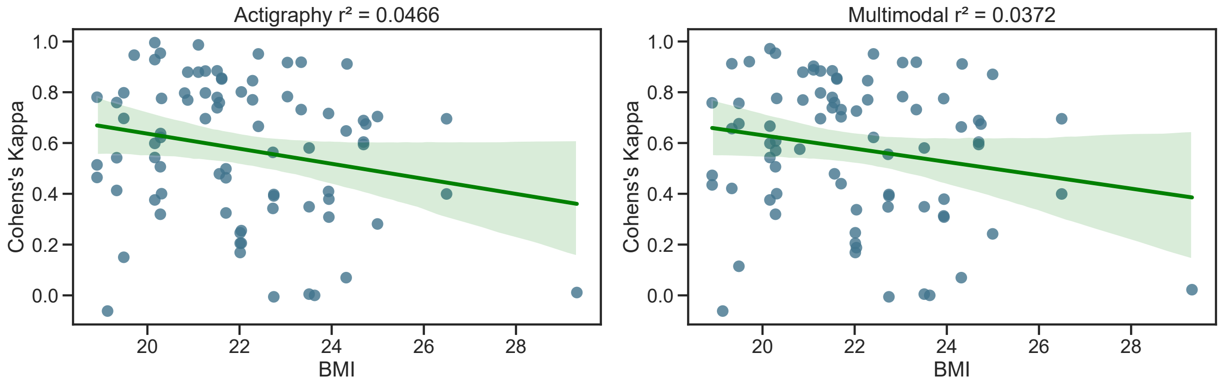

Figure 8.11: κ of sleep/wake detection as a function of ***BMI***. *Green line*: Linear regression slope with 95% confidence interval; predictor: ***BMI***; dependent variable: Cohen's κ

8.2.4 Comparison of Actigraphy- and IMU-based Sleep/Wake Classification

This section compares the different sleep/wake classification approaches presented before. As visible in Figure 8.12, the deep learning algorithms performed best using the benchmarking approach. All assessments using deep learning and data acquired in the real-world study performed poorly. In contrast, the machine learning-based classification worked best using raw ***IMU*** data, while classification on activity counts extracted from raw ***IMU*** data achieved only poor performance. For the cardiac-based approach, only the benchmark dataset reached considerable performance. Additionally, the multimodal approach was only able to improve classification performance for the benchmark dataset. In contrast, classification models trained on the real-world dataset were not able to benefit from additional cardiac information.

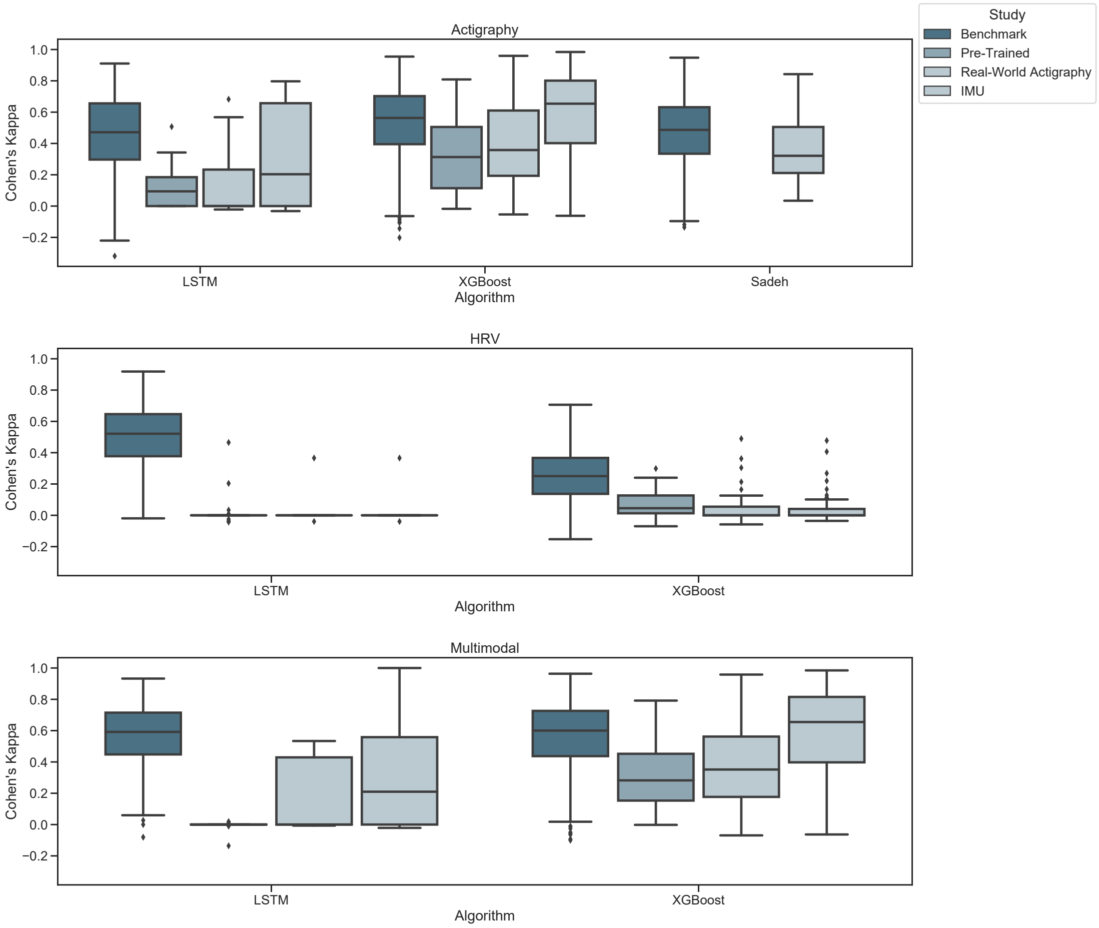

Figure 8.12: Classification of sleep/wake states using different modalities and approaches.

Similar to results from the benchmark dataset classification performance was higher for female participants (Figure 8.13).

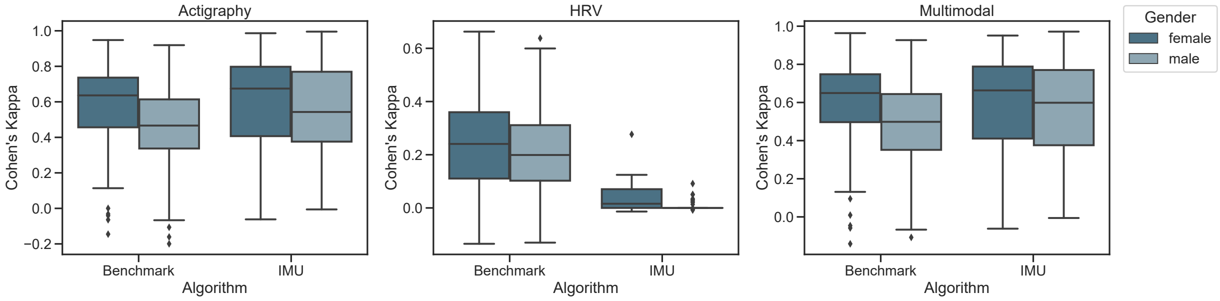

Figure 8.13: Performance dependent on gender of the participants for ***IMU***-based classification using ***AdaBoost***. The x-axis denotes the approach used for sleep/wake classification i.e., ***IMU*** denotes the usage of the real-world dataset.

Chapter 9

Discussion

9.1 Benchmark of Algorithms

One main objective of this thesis was to systematically benchmark sleep/wake detection using different types of algorithms with different input modalities on a large, diverse dataset. The results obtained from actigraphy-based, ***HRV***-based and multimodal approaches confirm previous findings that the ***LSTM*** performed best in sleep/wake detection accuracy. Palotti et al. [Pal19] already performed sleep/wake detection on the ***MESA*** dataset in an actigraphy-based approach, and Zhai et al. [Zha20] compared a motion-based approach with a cardiac-based and a multimodal approach. As found similarly by Palotti et al., the ***LSTM*** works best with an overall accuracy of about 83%, followed by various machine learning algorithms performing in the range of $80 - 82\%$ accuracy. Furthermore, this work additionally included ***XGB*** as classifier, which turned out to be the best-performing machine learning algorithm in all three modalities. For the deep learning algorithms, ***TCN*** performed slightly worse than the ***LSTM***. In the work of Palotti et al. [Pal19] as well as this thesis, the heuristic algorithms performed slightly worse with agreement rates mostly below 80%. The data used by Palotti et al. were distributed slightly different with about 60% of the epochs labeled as sleep and 40% as wake. In contrast, the cleaned and preprocessed ***MESA*** dataset for this thesis contained about 66% sleep and 34% wake epochs. The overprediction of sleep, which is a common problem in sleep/wake detection, might have led to slightly better agreement rates observed for the machine learning approaches assessed by Palotti et al.

Zhai et al. [Zha20] extended the classification by introducing a cardiac-based and a multimodal approach. This thesis confirms and extends the findings, that machine learning algorithms detected sleep/wake states slightly worse compared to deep learning algorithms, while the ***HRV***-based approach only shows satisfying results using deep learning. The authors stated that the addition of cardiac-based data did not improve the performance of motion-based algorithms. However, in this thesis a clear improvement for machine- and deep learning algorithms was found.

In total, the best overall performance was achieved using the ***LSTM*** in the multimodal approach with an accuracy of $83.7 \pm 9.6\%$. This can be explained with the ***HRV*** data that are combined with the actigraphy. Since only the ***LSTM*** reached a remarkable performance in the cardiac-based approach the ***LSTM*** was able to benefit most from additional cardiac data. This finding is strengthened because the ***LSTM*** did not perform better than the machine learning algorithms in the motion-based classification.

Focusing on the motion-based approach, the heuristic algorithms performed worse compared to both machine and deep learning algorithms. This can be explained by static weights that cannot be adjusted according to different actigraph devices or different individuals participating in the study. To account for this issue the internal weights of the heuristic algorithms would need to be optimized for the benchmarking dataset to gain optimal performance, e.g., via grid search. This was, for instance, performed by Cole et al. [Col92].

The results show that the choice of algorithm and modality can highly influence the interpretation and diagnostic of individuals' sleep behavior. An algorithm that systematically overestimates sleep might result in undiagnosed sleep disorders, whereas algorithms that overpredict wake phases might lead to unnecessary clinical evaluations and treatments. It can be stated that all motion-based algorithms, except *Sazonov*, overpredict sleep, which is a well-known behavior in literature [Til09, Pal19]. However, *Sazonov* heavily overpredicts wake phases resulting in the worst accuracy, recall, and F1-score. One possible reason for of overprediction of sleep among heuristic algorithms is that, shortly before falling asleep or in wake phases during sleep, only little movement occurs. This might lead to more epochs falsely classified as sleep, which results in lower ***SOL*** and ***WASO***. In contrast, algorithms using ***HRV*** information tend to less overprediction of sleep, which might be caused by a differing heart rate between the phase shortly before falling asleep and the sleep phase itself. This can be observed in both ***HRV***-based and multimodal machine and deep learning models, respectively. Additionally, the actigraph-based deep learning approach also shows less sleep overprediction, however, only to a lesser extend. The cardiac-based machine learning algorithms only performed poorly in overall accuracy what might be the reason why these algorithms also suffered a strong overprediction of sleep.

For the deep learning algorithms, hyperparameter optimization results indicate that the sequence length influences classification performance with higher sequence lengths leading to higher performance. This demonstrates the dependency of sleep/wake epochs on the subsequent and the following epochs. Larger search spaces may yield even better performance and should be assessed in future work.

In addition to the overall classification performance, participant- and study-specific influences were assessed. One limitation of this thesis is that no strict in-depth quality control of actigraphy and ***PSG*** was enforced. For instance, as reported by Zhai et al. [Zha20], 30 individuals (2%) of the total dataset contained no ***REM*** epochs at all. Absence of ***REM*** sleep phases is considered as abnormal sleep patterns and, thus, probably poses invalid data [Bug21]. Future work using the ***MESA*** dataset should stricter data cleaning paradigms which also take physiological signal quality into account. However, this thesis assessed the signal quality issue retrospectively. It was examined whether bad ***PSG*** or actigraphy signal quality impairs classification accuracy. Results showed that for both the actigraph-based as well as the multimodal approaches, lower signal quality led to statistically significant decreases ($p < 0.05$) in classification performance. That is obvious because a bad actigraphy signal might lead to incorrect detection of movement patterns, whereas a bad ***PSG*** signal makes the whole recording less reliable. As visible in Table B.5, the ***HRV***-based approach with ***XGB*** was also significantly ($p < 0.05$) influenced by actigraph quality. This is probably caused by a confounder because 54 of the 69 participants with bad ***PSG*** signal quality were also labeled with bad actigraphy signal.

Another assessment was performed to find differences between gender. It was found that classifying nights of female participants led to significantly ($p < 0.05$) higher performance measures compared to males. Recent research about gender-specific sleep habits in large-scale, objective sleep data found that females have a smaller sleep duration and less ***SE*** compared to men. This might indicate that they do not fulfill their sleep needs and therefore have deeper sleep [Sou17, Li21], which might be easier to classify [Sil08]. However, these assumptions would need to be further confirmed in future studies.

No significant differences were found between healthy individuals and individuals that suffer from sleep disorders. This was unexpected because sleep disorders influence sleep quality [Cho10, Sin15]. Thus, it was hypothesized that the sleep/wake classification performance is also influenced. A possible reason for that is the small number of sick participants included in this study (17.6%). Future research should therefore examine the influence of sleep disorders on sleep/wake classification on large-scale datasets including more individuals with pathological sleep.

Furthermore, it was assessed if sleep/wake classification performance differs for individuals working more than 5 extra hours per week. Results showed that participants with a heavy amount of extra work experienced a significantly ($p < 0.05$) better classification, which might be explained by higher ***SE*** (88.4% vs. 85.1%), and lower ***WASO*** (59.6 min vs. 90.2 min). As short wake periods during night are difficult to predict, classifying individuals with fewer wake patterns during night probably leads to higher performance measures.

When analyzing the effect of age on classification performance by means of linear regression the ***HRV***-based as well as the multimodal approach showed a negative relationship between age and classification performance while the effect turned out to be stronger for ***HRV*** than for the multimodal approach. In contrast, the actigraphy-based approach was nearly uninfluenced. That can be explained by decreasing ***HRV*** with increasing age [Yer97], which makes classification of sleep patterns harder.

The last examination performed on the ***MESA*** dataset was to assess the influence of apneas and hypopneas for sleep/wake classification. In a regression analysis, it was found, that high ***AHI*** which indicates more frequent and more severe apneas and hypopneas, correlated with worse classification performance in all modalities. As suspected by Vallat et al. [Val21], this might be driven by an increased number of sleep stage- and sleep/wake transitions.

The ***MESA*** dataset is a huge, diverse study that includes individuals from different ethnicity. However, one limitation of the ***MESA*** dataset is that it only includes adults aged above 54. An ideal dataset would represent the distribution of the population including toddlers, children, and adolescents. Furthermore, more individuals suffering from various diseases should be included.

Another limitation of the ***MESA*** dataset is the controlled laboratory environment in which the dataset was collected, whilst the possible applications of sleep/wake detection via wearable sensors are mostly in real-world environments. This environment led to high data quality, but sleep habits and sleep quality might have been affected, which is why some conclusions are not completely applicable to real-world scenarios [Ibe04].

9.2 Real-World Study

To overcome this limitation, a real-world study including 22 participants was conducted for this work. This study was merged with the ***DiPsyLab*** study using the same sensors as well as the same ground truth, resulting in a dataset of 85 nights of 42 participants.

To compare the performance of actigraph-based and ***IMU***-based classification, the ***IMU*** data was converted into activity counts using the algorithm of Brønd et al. [Brø17]. These activity counts were then used to train and test the algorithms presented in Chapter 5. To compare the performance between benchmark and real-world dataset machine and deep learning models trained on the benchmark dataset were used to predict sleep and wake epochs from activity counts extracted from the real-world dataset.

One problem of research in the field of sleep/wake detection is that many publications compare different small-sized studies with each other without considering differences in study procedure. For instance, it is harder to reach high-performance measures by classifying night-only datasets than to perform sleep/wake classification on datasets that also contain daytime. For instance, Palotti et al. [Pal19] compared a night-only with a full-day approach resulting in increased performance from 83.1 to 88.2% for the best-performing algorithm.

This thesis compares the performance obtained by the benchmarking approach with the performance measured by the application of the other three approaches. Because the ***MESA*** dataset not only contains the time in bed but also time before and after waking up, the dataset collected for this thesis might be slightly harder to classify.

However, comparing the benchmarking approach with the actigraphy-based and the ***IMU***-based approach, the ***IMU***-based algorithms reached the best results in terms of Cohen's κ. Especially the ***IMU***-based machine learning algorithms outperform the results of the benchmark dataset, as well as both actigraphy-based approaches on the real-world dataset. This might indicate, that the use of highly sampled accelerometer and gyroscope data provides valuable additional information for accurate sleep/wake prediction which gets lost when aggregating acceleration data to activity counts. This results in higher classification performance for ***IMU***-based sleep/wake classification compared to actigraphy-based approaches. In contrast, the actigraphy-based approaches both performed worse compared to the benchmarking approach. One reason for that might be the more imbalanced dataset containing more than 90% sleep. As datasets which contain less wake time are harder to classify, the ***MESA*** dataset might leads to better performance.

However, another possible reason for the worse performance is the non-ideal conversion from raw accelerometer data to activity counts. Brønd et al. [Brø17] only validated the algorithm for accelerometer data sampled at 30 Hz and stated in another publication that sampling rates different than 30 Hz might influence the conversion [Brø16].

This also leads to a general drawback of actigraphy in sleep/wake detection researchers need to be aware of. As explained in Section 4.1, many different manufacturers produce actigraph devices, each with different built-in closed-book algorithms that may change from version to version. This can lead to strong device-specific differences and, thus, to limited generalizability. For this reason, future research should remove barriers such as closed book algorithms to open up the possibility of pooling and comparing the results of different studies.

However, comparing the performance of the two actigraphy-based approaches applied on the real-world dataset, yields that the performance of training and testing on the real-world dataset is only slightly better than using the machine learning models trained on the benchmark dataset. The best-performing algorithm of both approaches (***XGB***) reached a κ value of 0.32 compared to 0.38 for the approach where data were both trained and tested on the real-world dataset. This indicates that the application of machine learning models, that are pre-trained on larger datasets, is a promising approach for future clinical practice.

Interestingly, both real-world actigraphy-based approaches showed higher ***WASO*** compared to the ground truth and, concurrently, overpredicted sleep. As the ***SOL*** of these approaches was very low, it is conceivable that the participants were classified asleep directly after going to bed even though they were still awake. Subsequent correctly classified wake epochs then resulted in a higher ***WASO***.

Although the machine learning algorithms performed well in the ***IMU***-based approach, ***TCN*** and ***LSTM*** performed poorly in all assessments using the real-world dataset. A possible reason for that is the comparably little amount of data from the real-world dataset, while another probable issue are large sequence lengths. Because sleep data is highly dependent on sleep states before and after the epoch being observed, an appropriate choice of sequence length is crucial. However, the ***IMU*** data were highly sampled, resulting in long sequences, which might cause deep learning algorithms facing the vanishing gradient problem [Hoc98].

Another finding was that ***HRV*** data, other than in the benchmarking approach, did not improve classification performance, in contrast to the benchmarking approach. Comparing the ***IMU***-based with the multimodal approach shows that both performance metrics and sleep statistics are a similar range for both machine- and deep learning algorithms. Furthermore, the monomodal ***HRV***-based approach reached only κ values close to zero which implies a classification performance that is as good as random guessing. This might be caused by the poorly performing deep learning algorithms for the real-world dataset that profited the most from cardiac data in the algorithm benchmarking.

Because alcohol negatively affects sleep quality, it was hypothesized that the consumption of alcohol also influences the classification performance of sleep and wake epochs. However, some algorithms in some modalities showed better performance on nights with alcohol, while others showed opposite behavior. One possible reason for that could be the small sample size of 13 nights with alcohol included compared to 37 nights without alcohol. Furthermore, the definition of a night influenced by alcohol was to drink two or more beverages of alcohol in the evening before. However, no distinction between moderate and excessive alcohol consumption was made.

Moreover, the duration from alcohol consumption to bed time might also change the influence of alcohol on sleep.

Another evaluation was performed to find the influence of subjective sleep quality immediately assessed in the morning after. It was found that higher subjective sleep quality leads to a significantly ($p < 0.05$) higher κ score in classification performance. This can be explained with shorter time to fall asleep (***SOL***) and less time awake during the night (***WASO***) that is typically difficult to classify, due to commonly observed overprediction of sleep. In contrast, the ***PSQI*** showed no influence on classification performance. One possible reason might be the cohort the dataset was collected from. All subjects were young healthy adults with no diagnosed sleep diseases. Additionally, the ***PSQI*** assesses sleep quality of the last four weeks. For that reason, the sleep quality of one single night might a higher influencing factor on sleep/wake classification performance than sleep quality of the last four weeks.

Waking up without alarm clock led to a higher classification performance than vice versa. Analyzing the total sleep duration as well as the bed time and wake-up time of participants, it is visible that participants waking up without alarm clock tended to sleep longer and go bed as well as wake up later (Figure B.1). As participants waking up with alarm clock might have has fixed appointments in the morning, they were getting up faster, and thus, the recording did not include a longer wake-up time including active patterns in the bed which are rather easy to detect. However, only 11 of 50 nights were recorded without alarm clock in the morning. For that reason, this topic needs further research and a larger, more balanced dataset in order to draw reliable conclusions.

Because employees are known to have a more regular lifestyle than students, it was hypothesized that the sleep habits and, thus, sleep/wake classification performance, might differ. However it turned out that no significant differences were observed. A probable reason for this is that the sleep habits of students and employees did not differ considerably in the cohort assessed in this thesis (Figure B.2).

In contrast, a negative correlation between ***BMI*** and sleep/wake classification performance was found (Figure 8.11). However, a recent publication of Vallat et al. [Val21] found no correlation using in a large-scale dataset. Blackwell et al. [Bla08] even found a slightly better classification performance for individuals with higher ***BMI***, and only a worsened performance for individuals classified as obese (***BMI*** $>$ 30 according to Krebs et al. [Kre07]). Thus, no obese participants were included in the study conducted for this work.

Chapter 10

Conclusion and Outlook

The first research goal of this thesis was to systematically benchmark a variety of heuristic, machine learning, and deep learning algorithms in monomodal approaches using actigraphy and ***HRV*** data, respectively, and multimodal approaches that combine both data modalities using the ***MESA*** dataset. This work mainly confirmed the findings from previous works [Pal19, Zha20], such as the ***LSTM*** as the best performing algorithm. Furthermore, only the deep learning approaches were able to reach remarkable performance in ***HRV***-based classification. However, in contrast to Zhai et al. [Zha20], this thesis found a clear improvement in classification performance when combining actigraph data with ***HRV*** data. The best overall performance was achieved by an ***LSTM*** with multimodal input data with an accuracy of $83.7 \pm 9.6\%$.

The performance boost achieved by adding ***HRV*** data to the actigraph-based approach shows, that the introduction of biosignals other than movement data can boost classification performance. Furthermore, only the deep learning algorithms were able to extract valuable information from the ***HRV***-data. For that reason, future research using cardiac information should focus on the application and improvement of deep learning algorithms. Possible approaches could be the application of ensemble deep learning algorithms presented by Zhai et al. [Zha20] or algorithms using attention mechanisms presented by Chen et al. [Che20a].

However, the ***MESA*** dataset was solely collected in a large controlled laboratory environment whilst possible applications of sleep/wake detection using wearable sensors primarily lie in real-world scenarios. For that reason, it is of particular importance for future research to collect large datasets in real-world settings which are standardized and diverse regarding demographics, and health status (especially sleep-related diseases).

To examine sleep/wake detection in a real-world setting, a new real-world dataset, including ***IMU*** and ***ECG*** data during sleep, was collected and assessed as second major part of this thesis. To compare the performance with the ***MESA*** dataset, the accelerometer signal was converted and aggregated into activity counts and applied to the same algorithms as the benchmark dataset. In a second examination, these activity counts were applied to pre-trained models extracted from the benchmarking approach.

Comparing both real-world actigraphy-based approaches, the classification via pre-trained models performed only slightly worse than to train and test the algorithms with the activity counts. Consequently, applying real-world data to pre-trained models is a promising approach for future clinical practice.

Due to a low sampling rate and the aggregation of accelerometer data into activity counts, important information may get lost. The third main objective of this thesis was therefore to examine whether the sleep/wake detection performance can be further improved by introducing new modalities, e.g, using features computed from highly sampled acceleration and angular velocity instead of activity counts. This ***IMU***-based approach outperformed, the benchmarking as well as both actigraphy-based approaches using the real-world dataset. These results suggest using ***IMU*** data instead of activity counts, due to increased performance. Possible reasons for this were the addition of gyroscope data as well as the preservation of valuable movement information which gets lot when aggregating acceleration data to activity counts. That might led to more granular features. Another drawback of actigraphy data is the limited comparability due to manufacture specific actigraph devices each with different built-in closed-book algorithms.

As visible in the results, all algorithms in both datasets produced heavy outliers. Future research should investigate whether these outliers are caused from the same data across all algorithms. For datasets containing more than one night, it would be interesting if a subject specific influence on classification performance can be found. Analysis of these outlier datasets would provide further insight into the requirements for accurate classification performance.

However, further research needs to be done to confirm these results, such as the collection and evaluation of a large, high-quality real-world dataset containing movement information of highly sampled ***IMU*** data. Thereby, new high-level features might further improve sleep/wake classification. For that reason, it might be advantageous to develop new features from different modalities and perform a feature importance analysis.

The sensors for motion-based sleep/wake detection are usually wrist-worn [Imt21]. The dataset collected for this thesis contained ***IMU*** sensors at wrist and chest, however, the data collected from the chest-worn ***IMU*** sensor has not yet been used. Possible future research could assess sleep/wake classification using ***IMU*** recordings from the chest and compare the results with classification based on ***IMU*** data collected from wrist sensors. In addition, combining both ***IMU*** recordings could be beneficial and requires further research.

Furthermore, it may be advantageous to introduce more biosignals as input modalities. For instance, respiration is, similar to cardiac activity, drastically reduced during sleep [Dou82]. An unobtrusive way to measure respiration was presented by Schäfer et al. [Sch08]. In their work, average respiration was approximated from ***HRV*** data via respiratory sinus arrhythmia. Moreover, Lenis et al. [Len15] reported, that the electrical impedance of the torso changes depending on the volume of air inside the lungs, which leads to potential differences which can be measured by the ***ECG*** electrodes. The combination of these approaches is very promising, because ***ECG*** data, which turned out to be beneficial in the benching approach, is sufficient to approximate respiration. Therefore, more biosignals can be integrated into the classification without additional obtrusiveness. Another way to measure respiration was proposed by Cesareo et al. [Ces19]. They were able to extract respiration with a principal component analysis from ***IMU*** data. As the dataset collected for this thesis contains ***IMU*** and ***ECG*** data, a possible follow-up work could be the assessment of multimodal sleep/wake detection using ***IMU***, ***HRV***, and respiration data.

Appendix A

Hyperparameter Optimization

Algorithm	Parameter	Range
Webster	scale_value	[0.025, 0.055], *stepsize* = 0.005
	rescoring algorithm	[True, False]
Cole-Kripke	scale_value	$[1 \cdot 10^{-4}, 25 \cdot 10^{-3}]$, *stepsize* = $5 \cdot 10^{-4}$
	rescoring algorithm	[True, False]
Scripps Clinic	scale_value	[0.15, 0.65], *stepsize* = 0.05
	rescoring algorithm	[True, False]
Sadeh	rescoring algorithm	[True, False]
Sazonov	rescoring algorithm	[True, False]

Table A.1: Parameter search space for heuristic algorithms.

Algorithm	Parameter	Range
SVM	penalty	[l1, l2, elasticnet]
	alpha	$[1 \cdot 10^{-4}, 0.1]$, *log distribution*
	fit_intercept	[True, False]
	warm_start	[True, False]
	learning_rate	[optimal, constant, adaptive, invscaling]
	eta0	[0.1, 0.5], *stepsize* = 0.1
	power_t	[0.0, 1.0], *stepsize* = 0.1
Random Forest	n_estimators	[10, 450], *stepsize* = 10
	criterion	[gini, entropy]
	max_depth	[3, 30], *stepsize* = 1
	min_samples_split	[2, 50], *stepsize* = 5
	min_samples_leaf	[1, 50], *stepsize* = 5
	min_weight_fraction_leaf	[0.0, 0.5], *stepsize* = 0.1
	max_features	[auto, sqrt, log2]
	max_leaf_nodes	[20, 10000], *stepsize* = 100
	min_impurity_decrease	[0.0, 0.1], *stepsize* = 0.01
	bootstrap	[True, False]
	oob_score	[True, False]
	max_samples	[100, 2000, None], *stepsize* = 50
	ccp_alpha	[0.0, 0.15], *stepsize* = 0.05
AdaBoost	learning_rate	$[1 \cdot 10^{-4}, 1 \cdot 10^{-3}, 0.01, 0.1, 1.0]$
	n_estimators	[10, 50, 100, 500]
MLP	hidden_layer_sizes	[(50, 50, 50), (50, 100, 50), (100,)]
	activation	[tanh, relu]
	solver	[adam]
	alpha	$[1 \cdot 10^{-4}, 0.05]$
	learning_rate	[constant, adaptive]
XGB	n_estimators	[200, 400], *stepsize* = 1
	max_depth	[10, 20], *stepsize* = 1
	reg_alpha	[0, 20], *stepsize* = 1
	reg_lamda	[0, 15], *stepsize* = 1
	min_child_weight	[0, 15], *stepsize* = 1
	gamma	[5, 20], *stepsize* = 1
	learning_rate	[0.01, 0.1], *stepsize* = 0.01
	colsample_by_tree	[0.1, 1], *stepsize* = 0.1

Table A.2: Hyperparameter search space for machine learning models.

Algorithm	Parameter	Range
LSTM	sequence_length	[21, 51, 101]
	num_layers	[1, 5], *stepsize* = 1
	learning_rate	[$1 \cdot 10^{-4}$, $5 \cdot 10^{-4}$, $1 \cdot 10^{-3}$, $5 \cdot 10^{-3}$, 0.01, 1]
	batch_size	[64, 128]
	overlap (of sequences)	[0.8, 0.9]
	hidden_size	[8, 16, 32, 64, 128, 256, 512]
TCN	sequence_length	[21, 51, 101]
	num_chanels	[2, 5], *stepsize* = 1
	n_hid	[8, 16, 32, 64, 128, 256, 512]
	kernel_size	[2, 5], *stepsize* = 1
	dropout	[0.1, 0.5], *stepsize* = 0.1
	learning_rate	[$1 \cdot 10^{-4}$, $5 \cdot 10^{-4}$, $1 \cdot 10^{-3}$, $5 \cdot 10^{-3}$, 0.01, 1]
	batch_size	[64, 128]

Table A.3: Hyperparameter search space for the Deep Learning algorithms.

Appendix B

Additional Statistics and Figures

B.1 Algorithm Comparison

B.1.1 Benchmark of Algorithms

Algorithm	*NSD* [min]	*MAE NSD* [min]	*SOL* [min]	*MAE SOL* [min]
Ground truth	698.9 ± 166.7	0.0 ± 0.0	50.4 ± 73.5	0.0 ± 0.0
		Actigraphy		
Cole-Kripke	734.7 ± 176.4	9.8 ± 8.9	30.1 ± 50.8	37.1 ± 59.3
Sadeh	784.5 ± 173.4	12.2 ± 10.2	24.0 ± 42.0	41.6 ± 65.9
Sazonov	622.3 ± 169.1	11.4 ± 9.3	17.8 ± 32.3	40.8 ± 69.5
Scripps-Clinic	**743.0 ± 168.5**	**9.6 ± 8.9**	11.5 ± 25.6	44.4 ± 71.6
Webster	740.9 ± 174.7	10.0 ± 9.1	29.1 ± 51.5	38.0 ± 61.8
AdaBoost	748.5 ± 188.4	10.8 ± 9.5	28.4 ± 46.2	34.2 ± 59.1
MLP	746.8 ± 187.4	10.6 ± 9.5	28.1 ± 46.4	34.5 ± 59.3
Random Forest	748.7 ± 186.3	10.6 ± 9.4	28.7 ± 46.0	33.9 ± 58.0
SVM	764.4 ± 184.3	11.7 ± 9.8	26.4 ± 45.1	35.2 ± 60.0
XGB	747.4 ± 186.6	10.5 ± 9.4	**29.9 ± 47.4**	**33.7 ± 57.4**
LSTM	750.2 ± 181.0	10.1 ± 9.2	53.7 ± 74.6	51.8 ± 68.8
TCN	782.2 ± 176.2	10.8 ± 9.3	44.0 ± 66.8	51.8 ± 67.5
		HRV		
AdaBoost	806.5 ± 205.4	18.1 ± 13.3	3.1 ± 35.9	55.1 ± 85.4
MLP	801.5 ± 203.3	17.4 ± 13.2	2.6 ± 30.0	54.9 ± 83.2
Random Forest	796.8 ± 212.2	17.9 ± 13.7	2.4 ± 24.5	54.8 ± 81.5
SVM	860.4 ± 192.3	21.5 ± 13.2	3.0 ± 34.7	55.1 ± 84.7
XGB	787.6 ± 209.0	17.1 ± 13.5	2.4 ± 24.5	54.7 ± 81.6
LSTM	**709.2 ± 174.8**	**11.8 ± 10.6**	**27.7 ± 47.6**	**50.2 ± 69.7**
TCN	745.0 ± 171.4	12.3 ± 10.9	5.7 ± 13.1	52.6 ± 77.0
		Multimodal		
AdaBoost	737.1 ± 190.2	10.5 ± 9.6	26.8 ± 44.3	35.1 ± 60.3
MLP	734.8 ± 185.2	10.0 ± 9.3	26.0 ± 42.7	35.0 ± 62.2
Random Forest	741.2 ± 184.9	10.1 ± 9.3	**28.8 ± 47.6**	**34.1 ± 60.6**
SVM	765.3 ± 184.1	11.7 ± 9.8	26.1 ± 45.0	35.3 ± 60.4
XGB	734.8 ± 185.0	9.9 ± 9.1	27.9 ± 44.9	34.7 ± 61.6
LSTM	730.3 ± 169.9	10.0 ± 8.9	50.7 ± 73.5	51.8 ± 74.6
TCN	**720.6 ± 161.3**	**9.7 ± 8.3**	24.3 ± 48.6	49.0 ± 71.5

Table B.1: Mean absolute error compared to ground truth of sleep statistics measured on the benchmark dataset.

B.1.2 Classification on Pre-Trained Models

Algorithm	***NSD* [min]**	***MAE NSD* [min]**	***SOL* [min]**	***MAE SOL* [min]**
Ground truth	853.8 ± 163.4	0.0 ± 0.0	44.8 ± 29.5	0.0 ± 0.0
Actigraphy				
AdaBoost	893.9 ± 169.8	52.3 ± 51.0	6.3 ± 10.4	39.0 ± 29.7
MLP	892.8 ± 169.2	50.4 ± 50.7	5.6 ± 9.6	39.5 ± 29.8
Random Forest	895.4 ± 168.2	50.8 ± 50.5	5.6 ± 9.7	39.4 ± 30.1
SVM	**878.7 ± 170.5**	**47.6 ± 48.3**	7.6 ± 11.5	37.7 ± 29.4
XGB	896.2 ± 169.1	51.5 ± 51.7	**8.4 ± 9.4**	**36.8 ± 29.4**
LSTM	714.1 ± 135.8	151.8 ± 70.7	0.3 ± 1.2	39.8 ± 20.7
TCN	2.2 ± 3.8	863.6 ± 125.3	194.4 ± 288.5	197.9 ± 255.4
HRV				
AdaBoost	886.0 ± 193.6	72.4 ± 55.6	1.3 ± 3.0	43.3 ± 29.1
MLP	874.4 ± 196.3	74.0 ± 57.5	1.4 ± 2.5	43.2 ± 29.0
Random Forest	874.6 ± 199.1	72.6 ± 61.3	1.7 ± 3.8	42.9 ± 29.1
SVM	**911.5 ± 176.5**	**70.2 ± 57.7**	0.7 ± 2.0	43.9 ± 29.4
XGB	868.0 ± 201.3	75.3 ± 63.2	1.6 ± 3.0	43.0 ± 29.1
LSTM	55.9 ± 16.4	809.9 ± 117.7	0.6 ± 1.9	39.5 ± 20.4
TCN	551.9 ± 187.7	313.9 ± 178.3	**2.4 ± 8.2**	**38.8 ± 21.1**
Multimodal				
AdaBoost	889.3 ± 172.2	51.8 ± 48.0	6.0 ± 9.1	39.5 ± 29.4
MLP	884.1 ± 174.8	54.2 ± 46.6	4.5 ± 6.8	40.2 ± 29.4
Random Forest	**894.1 ± 172.5**	**50.0 ± 45.5**	5.6 ± 9.7	39.4 ± 29.7
SVM	880.1 ± 170.8	48.1 ± 48.7	7.5 ± 11.4	37.7 ± 29.4
XGB	886.6 ± 179.2	49.8 ± 46.1	**7.9 ± 8.4**	**37.2 ± 28.9**
LSTM	84.4 ± 107.2	781.5 ± 158.9	178.1 ± 223.4	157.3 ± 197.8
TCN	128.2 ± 163.0	737.6 ± 217.3	160.9 ± 207.7	158.5 ± 178.3

Table B.2: Mean absolute error compared to ground truth of ***NSD*** and ***SOL*** measured on the real-world dataset using machine and deep learning models trained on the benchmarking dataset.

B.1.3 Classification on Actigraphy from Real-World Data

Algorithm	***NSD* [min]**	***MAE NSD* [min]**	***SOL* [min]**	***MAE SOL* [min]**
Ground truth	853.8 ± 163.4	90.8 ± 6.9%	44.8 ± 29.5	24.4 ± 40.1
Actigraphy				
Cole-Kripke	851.4 ± 170.0	54.8 ± 44.2	5.3 ± 10.0	39.7 ± 29.2
Sadeh	815.9 ± 170.2	62.4 ± 51.3	**21.4 ± 17.6**	**28.5 ± 28.9**
Sazonov	751.4 ± 153.4	107.0 ± 58.2	4.7 ± 9.2	40.3 ± 30.0
Scripps-Clinic	**843.8 ± 168.4**	**51.0 ± 40.7**	4.6 ± 8.9	40.3 ± 29.2
Webster	818.4 ± 170.7	63.2 ± 54.0	6.5 ± 12.2	38.6 ± 29.0
AdaBoost	901.5 ± 171.5	57.1 ± 54.2	5.5 ± 9.8	39.6 ± 30.0
MLP	903.8 ± 169.7	57.6 ± 52.8	5.1 ± 7.8	40.3 ± 30.0
Random Forest	903.9 ± 166.0	56.6 ± 51.2	6.5 ± 9.0	39.1 ± 29.4
SVM	887.4 ± 173.3	52.7 ± 50.5	5.4 ± 11.6	40.0 ± 29.4
XGB	898.3 ± 168.8	53.4 ± 50.2	10.1 ± 9.9	35.8 ± 29.5
LSTM	816.0 ± 130.3	56.5 ± 28.8	0.0 ± 0.0	40.1 ± 20.4
TCN	818.7 ± 130.6	53.6 ± 29.3	0.1 ± 0.5	39.9 ± 20.5
HRV				
AdaBoost	931.0 ± 167.8	78.4 ± 59.8	0.1 ± 0.4	44.6 ± 29.5
MLP	929.5 ± 169.0	79.0 ± 60.5	0.2 ± 1.0	44.6 ± 29.5
Random Forest	926.4 ± 170.4	77.0 ± 60.3	0.4 ± 1.1	44.4 ± 29.4
SVM	869.6 ± 206.0	88.2 ± 104.6	1.4 ± 2.5	43.3 ± 29.2
XGB	930.6 ± 167.5	78.9 ± 59.7	0.3 ± 1.1	44.5 ± 29.5
LSTM	822.6 ± 129.7	53.6 ± 26.2	**0.0 ± 0.0**	**40.1 ± 20.4**
TCN	**825.4 ± 127.7**	**50.2 ± 25.5**	**0.0 ± 0.0**	**40.1 ± 20.4**
Multimodal				
AdaBoost	896.4 ± 170.9	53.9 ± 50.1	5.4 ± 8.9	39.7 ± 29.6
MLP	900.1 ± 167.8	55.6 ± 50.0	4.2 ± 6.8	41.1 ± 29.3
Random Forest	904.5 ± 165.7	56.6 ± 50.4	6.2 ± 9.2	39.2 ± 29.3
SVM	887.4 ± 173.3	52.7 ± 50.5	5.4 ± 11.6	40.0 ± 29.4
XGB	**895.6 ± 173.1**	**51.3 ± 49.6**	**9.5 ± 9.6**	**36.4 ± 28.6**
LSTM	814.6 ± 130.9	56.4 ± 31.4	0.0 ± 0.0	40.1 ± 20.4
TCN	813.5 ± 129.5	61.5 ± 28.5	0.1 ± 0.2	40.0 ± 20.5

Table B.3: Mean absolute error compared to ground truth of ***NSD*** and ***SOL*** measured on the real-world dataset using extracted activity counts from ***IMU*** data.

B.1.4 IMU-based Sleep/Wake Classification

Algorithm	***NSD* [min]**	***MAE NSD* [min]**	***SOL* [min]**	***MAE SOL* [min]**
Ground truth	853.8 ± 163.4	90.8 ± 6.9%	44.8 ± 29.5	24.4 ± 40.1
Actigraphy				
AdaBoost	875.1 ± 162.7	39.9 ± 45.2	**35.2 ± 18.3**	**18.4 ± 22.8**
MLP	**880.5 ± 163.4**	**38.3 ± 45.0**	28.6 ± 16.8	21.3 ± 24.7
Random Forest	873.6 ± 190.2	53.7 ± 109.5	31.7 ± 19.2	20.6 ± 23.7
SVM	882.1 ± 161.3	40.1 ± 45.7	34.8 ± 17.3	18.7 ± 22.2
XGBoost	874.1 ± 188.3	51.9 ± 109.0	30.2 ± 18.0	21.0 ± 23.7
LSTM	842.5 ± 158.5	53.0 ± 21.2	0.0 ± 0.0	35.1 ± 24.1
TCN	829.5 ± 162.0	58.9 ± 26.9	0.5 ± 2.5	34.5 ± 24.6
HRV				
AdaBoost	931.0 ± 167.8	78.4 ± 59.8	0.1 ± 0.4	44.6 ± 29.5
MLP	929.5 ± 168.0	76.5 ± 59.7	0.1 ± 0.9	44.6 ± 29.5
Random Forest	925.8 ± 170.5	77.4 ± 60.6	0.4 ± 1.2	44.4 ± 29.3
SVM	831.6 ± 307.5	147.1 ± 194.3	7.4 ± 49.1	46.6 ± 47.7
XGBoost	931.6 ± 167.0	79.3 ± 60.1	0.2 ± 1.0	44.5 ± 29.5
LSTM	822.6 ± 129.7	53.6 ± 26.2	**0.0 ± 0.0**	**40.1 ± 20.4**
TCN	**825.4 ± 127.7**	**50.2 ± 25.5**	**0.0 ± 0.0**	**40.1 ± 20.4**
Multimodal				
AdaBoost	877.6 ± 160.9	39.6 ± 46.0	**35.7 ± 17.6**	**18.3 ± 22.5**
MLP	**872.9 ± 165.0**	**35.8 ± 42.4**	26.1 ± 18.5	22.1 ± 24.1
Random Forest	872.6 ± 190.3	52.6 ± 109.4	31.4 ± 18.5	20.4 ± 23.7
SVM	882.1 ± 161.3	40.1 ± 45.7	34.8 ± 17.3	18.7 ± 22.2
XGBoost	873.1 ± 189.5	51.9 ± 108.9	30.4 ± 18.2	20.0 ± 23.6
LSTM	818.0 ± 153.0	67.7 ± 22.7	1.0 ± 3.0	35.1 ± 23.8
TCN	832.5 ± 159.0	57.2 ± 24.1	0.5 ± 2.5	34.5 ± 24.6

Table B.4: Mean absolute error compared to ground truth of ***NSD*** and ***SOL*** measured on the real-world dataset using ***IMU*** data.

B.2 Statistical Tests

B.2.1 Benchmarking of algorithms

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	**LSTM**	11869.0	0.0368*	-0.2365
	XGBoost	11069.0	0.0021**	-0.3033
	Sadeh	11516.5	0.0115*	-0.3049
HRV	**LSTM**	11895.0	0.0266*	-0.2361
	XGBoost	10413.5	0.0001***	-0.4678
Multimodal	**LSTM**	11672.0	0.0130*	-0.2629
	XGBoost	10931.0	0.0008***	-0.3372

Table B.5: Results of *Mann-Whitney-U-Test* to assess the influence of good ($n = 128$) or poor ($n = 221$) quality of the actigraph signal on sleep/wake detection accuracy ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	**LSTM**	7296.0	0.0049**	-0.5588
	XGBoost	7376.0	0.0071**	-0.5125
	Sadeh	7699.0	0.0270*	-0.3925
HRV	**LSTM**	8135.0	0.0845	-0.2822
	XGBoost	8043.0	0.0626	-0.3189
Multimodal	**LSTM**	7693.0	0.0176*	-0.5089
	XGBoost	7472.0	0.0071**	-0.4410

Table B.6: Results of *Mann-Whitney-U-Test* to assess the influence of good ($n = 280$) or poor ($n = 69$) quality of the ***PSG*** signal on sleep/wake detection accuracy ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	**LSTM**	21307.5	$< 0.001^{***}$	0.6620
	XGBoost	21217.0	$< 0.001^{***}$	0.6448
	Sadeh	19917.5	$< 0.001^{***}$	0.5007
HRV	**LSTM**	16865.0	0.1118	0.1287
	XGBoost	16576.5	0.2174	0.1311
Multimodal	**LSTM**	19946.0	$< 0.001^{***}$	0.4901
	XGBoost	20873.0	$< 0.001^{***}$	0.6119

Table B.7: Results of *Mann-Whitney-U-Test* to assess the influence of gender ($n_{female} = 192$, $n_{male} = 157$) on sleep/wake detection accuracy ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	**LSTM**	6553.0	> 0.999	0.0272
	XGBoost	6599.0	> 0.999	0.0259
	Sadeh	6537.0	> 0.999	-0.0949
HRV	**LSTM**	7102.5	> 0.999	-0.0010
	XGBoost	6754.0	> 0.999	-0.0649
Multimodal	**LSTM**	6835.5	> 0.999	0.1028
	XGBoost	6687.0	> 0.999	0.0014

Table B.8: Results of *Mann-Whitney-U-Test* to assess the influence of sickness ($n_{sick} = 54$, $n_{healty} = 295$) on sleep/wake detection accuracy ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	**LSTM**	16391.5	0.0467^{*}	0.2721
	XGBoost	16137.5	0.0971	0.2551
	Sadeh	15791.5	0.2354	0.1866
HRV	**LSTM**	15295.5	0.4491	0.1070
	XGBoost	15533.0	0.2801	0.1264
Multimodal	**LSTM**	15163.0	0.5702	0.1410
	XGBoost	16000.0	0.0934	0.2125

Table B.9: Results of *Mann-Whitney-U-Test* to assess the influence of sleep quality according to WHIIRS scoring ($n_{good} = 129$, $n_{poor} = 220$) on sleep/wake detection accuracy ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	**LSTM**	5307.5	0.0276*	-0.3182
	XGBoost	5161.0	0.0138*	-0.3632
	Sadeh	5143.5	0.0126*	-0.4156
HRV	**LSTM**	5411.0	0.0291*	-0.3359
	XGBoost	5075.5	0.0060**	-0.3338
Multimodal	**LSTM**	5513.5	0.0450*	-0.3388
	XGBoost	4979.0	0.0036**	-0.4309

Table B.10: Results of *Mann-Whitney-U-Test* to assess the influence of more than 5 h extra workload per week($n_{work>5} = 46$, $n_{normal} = 303$) on sleep/wake detection accuracy ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Kruskal-Wallis test			
Modality	**Algorithm**	**p**	**H**
Actigraphy	**LSTM**	0.6267	1.7460
	Sadeh	0.4901	2.4191
	XGBoost	0.5055	2.3377
HRV	**LSTM**	0.1224	5.7886
	XGBoost	0.0188*	9.9717
Multimodal	**LSTM**	0.4537	2.6215
	XGBoost	0.3309	3.4228

Posthoc pairwise ttests						
Modality	**Algorithm**	**A**	**B**	**U-value**	**p**	**Hedges' g**
HRV	**XGBoost**	black	chinese	1602.0	$>$ 0.999	-0.3614
		black	hispanic	3074.0	0.7389	-0.3267
		black	white	5416.0	0.2755	-0.3310
		chinese	hispanic	1480.0	$>$ 0.999	0.0398
		chinese	white	2680.5	$>$ 0.999	0.0427
		hispanic	white	4902.0	$>$ 0.999	0.0022

Table B.11: Results of *Kruskal Wallis test* to assess the influence of race on sleep/wake detection accuracy (top) and *Mann-Whitney-U* posthoc tests (bottom) ($n_{white} = 130$, $n_{black} = 105$,$n_{chinese} = 40$, $n_{hispanic} = 74$) ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Kruskal-Wallis test			
Modality	**Algorithm**	**p**	**H**
Actigraphy	**LSTM**	0.2472	7.8776
	Sadeh	0.6825	3.9572
	XGBoost	0.4812	5.5021
HRV	**LSTM**	0.9868	0.9678
	XGBoost	0.8217	2.8970
Multimodal	**LSTM**	0.7644	3.3458
	XGBoost	0.7802	3.2244

Table B.12: Results of *Mann-Whitney-U-Test* to assess the influence of different sickness on sleep/wake detection ($n_{healthy} = 295$, $n_{\boldsymbol{INS}} = 19$, $n_{\boldsymbol{RLS}} = 7$, $n_{apnea} = 21$, $n_{apnea+\boldsymbol{INS}} = 3$, $n_{apnea+\boldsymbol{RLS}} = 2$, $n_{\boldsymbol{RLS}+\boldsymbol{INS}} = 2$) ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

B.2.2 Real-World Dataset

Modality	**Algorithm**	**U-value**	**p**	**Hedges' g**
Actigraphy	**AdaBoost**	217.0	> 0.999	-0.1305
	XGBoost	256.0	> 0.999	0.1147
HRV	**AdaBoost**	239.5	> 0.999	-0.1293
	XGBoost	300.5	> 0.999	0.0629
Multimodal	**AdaBoost**	260.0	> 0.999	0.1098
	XGBoost	256.0	> 0.999	0.0927

Table B.13: Results of *Mann-Whitney-U-Test* to assess the influence of alcohol ($n_{alcohol} = 13$, $n_{no\_alcohol} = 37$) on sleep/wake detection measured with κ ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	AdaBoost	138.0	0.0158*	-0.9142
	XGBoost	198.0	0.4233	-0.5875
HRV	AdaBoost	202.5	0.4391	-0.6464
	XGBoost	212.0	0.6048	-0.7078
Multimodal	AdaBoost	186.0	0.2433	-0.6401
	XGBoost	221.0	> 0.999	-0.4643

Table B.14: Results of *Mann-Whitney-U-Test* to assess the influence of subjective sleep quality ($n_{good} = 17$,$n_{bad} = 33$) on sleep/wake detection measured with κ ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	AdaBoost	843.0	> 0.999	0.2877
	XGBoost	888.0	> 0.999	0.3473
HRV	AdaBoost	888.5	> 0.999	0.4103
	XGBoost	907.0	0.6835	0.5472
Multimodal	AdaBoost	929.0	0.5520	0.4586
	XGBoost	879.0	> 0.999	0.3251

Table B.15: Results of *Mann-Whitney-U-Test* to assess the influence of the sleep quality score acquired via ***PSQI*** ($n_{good} = 32$, $n_{bad} = 47$) on sleep/wake detection measured with κ ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	AdaBoost	164.0	0.4474	-0.5089
	XGBoost	166.0	0.4913	-0.4887
HRV	AdaBoost	241.0	> 0.999	-0.0982
	XGBoost	188.5	> 0.999	-0.1051
Multimodal	AdaBoost	157.0	0.3184	-0.4901
	XGBoost	186.0	> 0.999	-0.3431

Table B.16: Results of *Mann-Whitney-U-Test* to assess the influence of the wake-up mode ($n_{alarm} = 11$,n_{no_alarm}=39) on sleep/wake detection measured with κ ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Modality	Algorithm	U-value	p	Hedges' g
Actigraphy	**AdaBoost**	232.0	> 0.999	-0.1210
	XGBoost	243.0	> 0.999	-0.1778
HRV	**AdaBoost**	152.0	0.1145	-0.7661
	XGBoost	151.5	0.0920	-0.6424
Multimodal	**AdaBoost**	216.0	> 0.999	-0.2138
	XGBoost	247.0	> 0.999	-0.1249

Table B.17: Results of *Mann-Whitney-U-Test* to assess the influence of the profession the participant had $n_{student} = 37$, $n_{employee} = 13$ on sleep/wake detection measured with κ (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

B.3 Additional Results from Real-World Evaluation

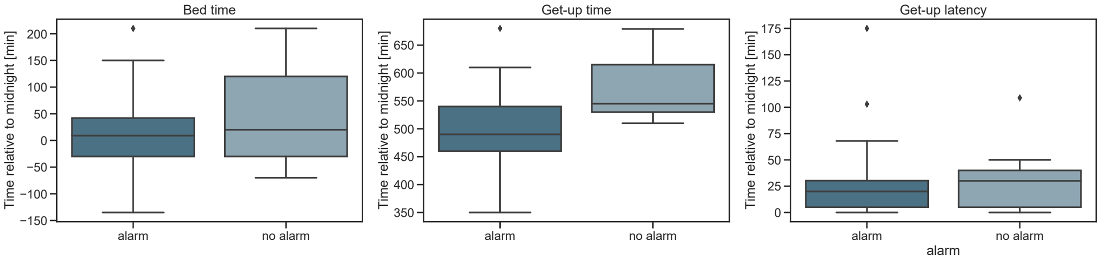

Figure B.1: Sleep habits of participants waking up with or without alarm.

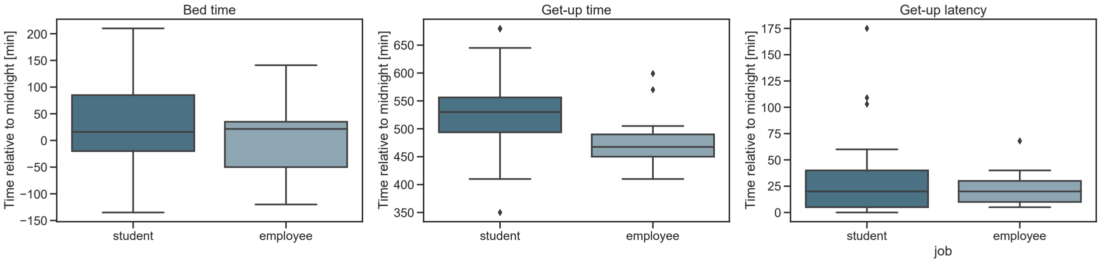

Figure B.2: Sleep habits of students and employees.

Appendix C

Questionnaires

Morning Questionnaire

Subjects ID: Vp_____ **Date:**

Night 1	
Bedtime	:
Estimated time to fall asleep	:
Wake-up time	:
Stand-up time (got out of bed)	:
Were you woken up by an alarm clock?	Yes / No
Have you consumed two or more alcoholic beverages before going to bed?	Yes / No
How would you rate the overall quality of your sleep last night?	

very bad	1	2	3	4	5	6	7	very good

Night 2	
Bedtime	:
Estimated time to fall asleep	:
Wake-up time	:
Stand-up time (got out of bed)	:
Were you woken up by an alarm clock?	Yes / No
Have you consumed two or more alcoholic beverages before going to bed?	Yes / No
How would you rate the overall quality of your sleep last night?	

very bad	1	2	3	4	5	6	7	very good

Night 3	
Bedtime	:
Estimated time to fall asleep	:
Wake-up time	:
Stand-up time (got out of bed)	:
Were you woken up by an alarm clock?	Yes / No
Have you consumed two or more alcoholic beverages before going to bed?	Yes / No
How would you rate the overall quality of your sleep last night? very bad 1 2 3 4 5 6 7 very good	

Sleep Quality Questionnaire (PSQI)

The following questions relate to your usual sleeping habits *only during the past four weeks.* Your answers should be as specific as possible and refer to the majority of the days and nights during the past four weeks. Please answer all questions.

1. During the past four weeks, when have you usually gone to bed at night?	**Usual time:**
2. During the past four weeks, how long has it usually taken you to fall asleep at night?	**in minutes:**
3. In the past four weeks, when did you usually get up in the morning?	**Usual time:**
4. How many hours have you actually slept per night during the past four weeks? **(This doesn't have to be the same as the number of hours you spent in bed.)**	**Effective sleep time (hours) per night:**

Please tick the answer that applies to you for each of the following questions.
Please answer every question.

5. How often have you slept badly in the last four weeks ...

a) ... because you couldn't fall asleep within 30 minutes?	**- Not at all for the past four weeks** **- Less than once a week** **- Once or twice a week** **- Three or more times a week**
b) ... because you woke up in the middle of the night or early in the morning?	**- Not at all for the past four weeks** **- Less than once a week** **- Once or twice a week** **- Three or more times a week**
c) ... because you had to get up to go to the toilet?	**- Not at all for the past four weeks** **- Less than once a week** **- Once or twice a week** **- Three or more times a week**

d) ... because you had difficulties to breathe?

- Not at all for the past four weeks
- Less than once a week
- Once or twice a week
- Three or more times a week

e) ... because you coughed or snored loudly?

- Not at all for the past four weeks
- Less than once a week
- Once or twice a week
- Three or more times a week

f). . . because you felt too cold?

- Not at all for the past four weeks
- Less than once a week
- Once or twice a week
- Three or more times a week

g) ... because you felt too warm?

- Not at all for the past four weeks
- Less than once a week
- Once or twice a week
- Three or more times a week

h) ... because you had a bad dream?

- Not at all for the past four weeks
- Less than once a week
- Once or twice a week
- Three or more times a week

i). . . because you were in pain?

- Not at all for the past four weeks
- Less than once a week
- Once or twice a week
- Three or more times a week

j). . . for other reasons?

Please describe:

And how many times in the past month have you had poor sleep because of this?

- Not at all for the past four weeks
- Less than once a week
- Once or twice a week
- Three or more times a week

6. Overall, how would you rate the quality of your sleep over the past four weeks?

- very good
- Pretty good
- Pretty bad
- Very bad

7. How often have you taken sleeping pills (doctor-prescribed or over-the-counter) in the past four weeks?

- **Not at all for the past four weeks**
- **Less than once a week**
- **Once or twice a week**
- **Three or more times a week**

8. How often have you had difficulty staying awake in the past four weeks, such as driving, eating, or at social gatherings?

- **Not at all for the past four weeks**
- **Less than once a week**
- **Once or twice a week**
- **Three or more times a week**

9. During the past four weeks, have you had trouble doing normal day-to-day tasks with enough momentum?

- **No problem**
- **Hardly any problems**
- **Some problems**
- **Big problem**

10. Do you sleep alone in your room?

- **Yes**
- **Yes, but a partner / roommate sleeps in a different room**
- **No, the partner sleeps in the same room, but not in the same bed**
- **No, the partner sleeps in the same bed**

If you have a roommate / partner, please ask him / her whether and how often he / she has noticed the following.

a) Loud snoring

- **Not at all for the past four weeks**
- **Less than once a week**
- **Once or twice a week**
- **Three or more times a week**

b) Long pauses in breathing during sleep

- **Not at all for the past four weeks**
- **Less than once a week**
- **Once or twice a week**
- **Three or more times a week**

c) twitching or jerking movements of the legs while sleeping

- **Not at all for the past four weeks**
- **Less than once a week**
- **Once or twice a week**
- **Three or more times a week**

d) Nocturnal phases of confusion or disorientation during sleep

- **Not at all for the past four weeks**
- **Less than once a week**
- **Once or twice a week**
- **Three or more times a week**

e) Or other forms of restlessness during sleep

Please describe:

Please provide the following information about yourself:

Age:	**____ Years**	**Height: ....**	**Weight:...................**
Gender:	**- Female** **- masculine**	**Job:** **- student** **- Worker**	**- Pensioner)** **- independent** **- Employee)** **- unemployed / housewife (husband)**

Glossary

AASM American Academy of Sleep Medicine.

AdaBoost Adaptive Boosting.

Adam Adaptive Moment Estimation.

AHI Apnea–hypopnea Index.

AI Artificial Intelligence.

ANN Artificial Neural Network.

BLE Bluetooth Low Energy.

BMI Body Mass Index.

CNN Convolutional Neural Network.

DIM Digital Integration Mode.

DiPsyLab Digital Psychology Lab.

ECG Electrocardiogram.

EEG Electroencephalogram.

EMG Electromyography.

EOG Electrooculography.

FN False Negative.

FP False Positive.

GMM Gaussian Mixture Model.

HMM Hidden Markov Model.

HRV Heart Rate Variability.

IMU Inertial Measurement Unit.

INS Insomnia.

LF-LSTM Local Feature-Based LSTM.

LSTM Long Short-Term Memory.

MAE Mean Absolute Error.

MESA Multi-Ethnic Study of Atherosclerosis.

MLP Multi Layer Perceptron.

MSLT Multiple Sleep Latency Test.

NN Normal-to-Normal.

NREM Non-Rapid Eye Movement.

NSD Net Sleep Duration.

PDSC Pediatric Daytime Sleepiness Scale.

PPG Photoplethysmography.

PSG Polysomnography.

PSQI Pittsburgh Sleep Quality Index.

REM Rapid Eye Movement.

RLS Restless legs syndrome.

RNN Recurrent Neural Network.

SE Sleep Efficiency.

SOL Sleep Onset Latency.

SVM Support Vector Machine.

TAT Time Above Threshold.

TCN Temporal Convolutional Network.

TN True Negative.

TP True Positive.

TSD Total Sleep Duration.

UCSD University of California San Diego.

WASO Wake after Sleep Onset.

WHIIRS Women's Health Initiative Insomnia Rating Scale.

XGB XGBoost.

ZCM Zero Crossing Mode.

List of Figures

List of Tables

Bibliography

[Aki19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[Bai18] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*, 2018.

[Ban07] Siobhan Banks and David F. Dinges. Behavioral and Physiological Consequences of Sleep Restriction. *Journal of Clinical Sleep Medicine : JCSM : official publication of the American Academy of Sleep Medicine*, 3(5):519–528, August 2007.

[Bar02] G. M. Barthlen. Schlafdiagnostik (Polysomnographie). In Heinrich Matthys and Werner Seeger, editors, *Klinische Pneumologie*, pages 103–111. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[Beh13] Soroor Behbahani, Nader Jafarnia Dabanloo, Ali Motie Nasrabadi, Cesar A. Teixeira, and Antonio Dourado. Pre-ictal heart rate variability assessment of epileptic seizures by means of linear and non-linear analyses. *Anadolu kardiyoloji dergisi: AKD = the Anatolian journal of cardiology*, 13(8):797–803, December 2013.

[Ber97] L. R. Berney and D. B. Blane. Collecting retrospective data: Accuracy of recall after 50 years judged against historical records. *Social Science & Medicine*, 45(10):1519–1525, November 1997.

[Bla08] Terri Blackwell, Susan Redline, Sonia Ancoli-Israel, Jennifer L. Schneider, Susan Surovec, Nathan L. Johnson, Jane A. Cauley, and Katie L. Stone. Comparison of Sleep Parameters from Actigraphy and Polysomnography in Older Women: The SOF Study. *Sleep*, 31(2):283–291, February 2008.

[Bon17] Giuseppe Bonaccorso. *Machine Learning Algorithms*. Packt Publishing Ltd, July 2017.

[Bor14] Marko Borazio, Eugen Berlin, Nagihan Kucukyildiz, Philipp Scholl, and Kristof Van Laerhoven. Towards Benchmarked Sleep Detection with Wrist-Worn Sensing Units. In *2014 IEEE International Conference on Healthcare Informatics*, pages 125–134, Verona, September 2014. IEEE.

[Brø16] Jan Christian Brønd and Daniel Arvidsson. Sampling frequency affects the processing of Actigraph raw acceleration data to activity counts. *Journal of Applied Physiology*, 120(3):362–369, February 2016.

[Brø17] Jan Christian Brønd, Lars Bo Andersen, and Daniel Arvidsson. Generating actiGraph counts from raw acceleration recorded by an alternative monitor. *2351-2360*, 2017.

[Bug21] Paulo Bugalho, Manuel Salavisa, Filipa Serrazina, Marco Fernandes, Gonçalo Cabral, André Sobral Pinho, and Rita Ventura. REM sleep absence in patients referred to polysomnography for REM sleep behavior disorder. *Journal of Neural Transmission*, 128(2):191–198, February 2021.

[Buy89] D. J. Buysse, C. F. Reynolds, T. H. Monk, S. R. Berman, and D. J. Kupfer. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Research*, 28(2):193–213, May 1989.

[Can13] Alberto Cano, Amelia Zafra, and Sebastián Ventura. Weighted Data Gravitation Classification for Standard and Imbalanced Data. *IEEE Transactions on Cybernetics*, 43(6):1672–1687, December 2013. Conference Name: IEEE Transactions on Cybernetics.

[Car16] David W. Carley and Sarah S. Farabi. Physiology of Sleep. *Diabetes Spectrum : A Publication of the American Diabetes Association*, 29(1):5–9, February 2016.

[Ces19] Ambra Cesareo, Ylenia Previtali, Emilia Biffi, and Andrea Aliverti. Assessment of Breathing Parameters Using an Inertial Measurement Unit (IMU)-Based System. *Sensors*, 19(1):88, January 2019. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[Che15] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline.

Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, June 2015.

[Che20a] Z. Chen, M. Wu, W. Cui, C. Liu, and X. Li. An Attention Based CNN-LSTM Approach for Sleep-Wake Detection with Heterogeneous Sensors. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2020. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[Che20b] Z. Chen, M. Wu, K. Gao, J. Wu, J. Ding, Z. Zeng, and X. Li. A Novel Ensemble Deep Learning Approach for Sleep-Wake Detection Using Heart Rate Variability and Acceleration. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–10, 2020. Conference Name: IEEE Transactions on Emerging Topics in Computational Intelligence.

[Cho10] Sudhansu Chokroverty. Overview of sleep & sleep disorders. *The Indian journal of medical research*, 131:126–40, February 2010.

[Col92] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin. Automatic Sleep/Wake Identification From Wrist Activity. *Sleep*, 15(5):461–469, 1992.

[Col11] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.

[Daf18] Eliran Dafna, Ariel Tarasiuk, and Yaniv Zigel. Sleep staging using nocturnal sound analysis. *Scientific Reports*, 8(1):13474, September 2018.

[DC11] Philip De Chazal, Niall Fox, Emer O'Hare, Conor Heneghan, Alberto Zaffaroni, Patricia Boyle, Stephanie Smith, Caroline O'Connell, and Walter T. Mcnicholas. Sleep/wake measurement using a non-contact biomotion sensor: Non-contact biomotion sensing of sleep. *Journal of Sleep Research*, 20(2):356–366, June 2011.

[Dev10] S. Devot, R. Dratwa, and E. Naujokat. Sleep/wake detection based on cardiorespiratory signals and actigraphy. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 5089–5092, August 2010. ISSN: 1558-4615.

[DiP20] Robert DiPietro and Gregory D. Hager. *Chapter 21 - Deep learning: RNNs and LSTM*. The Elsevier and MICCAI Society Book Series. Academic Press, 2020.

[Dom14] Alexandre Domingues, Teresa Paiva, and J. Miguel Sanches. Sleep and Wakefulness State Detection in Nocturnal Actigraphy Based on Movement Information. *IEEE Transactions on Biomedical Engineering*, 61(2):426–434, February 2014.

[Dou82] N J Douglas, D P White, C K Pickett, J V Weil, and C W Zwillich. Respiration during sleep in normal man. *Thorax*, 37(11):840–844, November 1982.

[Dra03] C. Drake, Chelsea Nickel, E. Burduvali, T. Roth, Catherine Jefferson, and Badia Pietro. The pediatric daytime sleepiness scale (PDSS): sleep habits and school outcomes in middle-school children. *Sleep*, 2003.

[dZ15] Massimiliano de Zambotti, Stephanie Claudatos, Sarah Inkelis, Ian M. Colrain, and Fiona C. Baker. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiology International*, 32(7):1024–1028, August 2015.

[dZ16] Massimiliano de Zambotti, Fiona C. Baker, Adrian R. Willoughby, Job G. Godino, David Wing, Kevin Patrick, and Ian M. Colrain. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiology & Behavior*, 158:143–149, May 2016.

[Ebr13] Irshaad O. Ebrahim, Colin M. Shapiro, Adrian J. Williams, and Peter B. Fenwick. Alcohol and Sleep I: Effects on Normal Sleep. *Alcoholism: Clinical and Experimental Research*, 37(4):539‘–549, 2013.

[Edo21] Paul Edouard, David Campo, Pierre Bartet, Rui-Yi Yang, Marie Bruyneel, Gabriel Roisman, and Pierre Escourrou. Validation of the Withings Sleep Analyzer, an under-the-mattress device for the detection of moderate-severe sleep apnea syndrome, February 2021.

[Fer94] F. Ferraris, I. Gorini, U. Grimaldi, and Marco Parvis. Calibration of three-axial rate gyros without angular velocity standards. *Sensors and Actuators A-physical - SENSOR ACTUATOR A-PHYS*, 42:446–449, April 1994.

[Hag20] Shahab Haghayegh, Sepideh Khoshnevis, Michael H. Smolensky, and Kenneth R. Diller. Application of deep learning to improve sleep scoring of wrist actigraphy. *Sleep Medicine*, 74:235–241, October 2020.

[Hag21] Shahab Haghayegh, Sepideh Khoshnevis, Michael H. Smolensky, Kenneth R. Diller, and Richard J. Castriotta. Deep Neural Network Sleep Scoring Using Combined

Motion and Heart Rate Variability Data. *Sensors*, 21(1):25, January 2021. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[Hoc98] Sepp Hochreiter. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116, April 1998. Publisher: World Scientific Publishing Co.

[Ibe04] Conrad Iber, Susan Redline, Adele M. Kaplan Gilpin, Stuart F. Quan, Lin Zhang, Daniel J. Gottlieb, David Rapoport, Helaine E. Resnick, Mark Sanders, and Philip Smith. Polysomnography Performed in the Unattended Home Versus the Attended Laboratory Setting—Sleep Heart Health Study Methodology. *Sleep*, 27(3):536–540, May 2004.

[Ibe07] Conrad Iber, Sonia Ancoli-Israel, Andrew Chesson, and Stuart Quan. *The AASM manual for the scoring of sleep and associated events: rules, terminology, and technical specification*. Publisher: American Academy of Sleep Medicin, Westchester, IL, 2007.

[Imt21] Syed Anas Imtiaz. A Systematic Review of Sensing Technologies for Wearable Sleep Staging. *Sensors*, 21(5):1562, February 2021.

[Jep14] Jesper Jeppesen, Sandor Beniczky, Peter Johansen, Per Sidenius, and Anders Fuglsang-Frederiksen. Using Lorenz plot and Cardiac Sympathetic Index of heart rate variability for detecting seizures for patients with epilepsy. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2014:4563–4566, 2014.

[JL01] Girardin Jean-Louis, Daniel F. Kripke, William J. Mason, Jeffrey A. Elliott, and Shawn D. Youngstedt. Sleep estimation from wrist movement quantified by different actigraphic modalities. *Journal of Neuroscience Methods*, 105(2):185–191, 2001.

[Kha16] Aftab Khan, Nils Hammerla, Sebastian Mellor, and Thomas Ploetz. Optimising sampling rates for accelerometer-based human activity recognition. *Pattern Recognition Letters*, 73, 01 2016.

[Kin17] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017.

[Kot07] S.B. Kotsiantis. Supervised machine learning: A review of classification techniques. *IOS Press*, pages 3–24, 2007.

[Kre07] Nancy F. Krebs, John H. Himes, Dawn Jacobson, Theresa A. Nicklas, Patricia Guilday, and Dennis Styne. Assessment of Child and Adolescent Overweight and Obesity. *Pediatrics*, 120(Supplement_4):S193–S228, December 2007.

[Kri10] Daniel F. Kripke, Elizabeth K. Hahn, Alexandra P. Grizas, Kep H. Wadiak, Richard T. Loving, J. Steven Poceta, Farhad F. Shadan, John W. Cronin, and Lawrence E. Kline. Wrist actigraphic scoring for sleep laboratory patients: algorithm development: Wrist actigraphic algorithm development. *Journal of Sleep Research*, 19(4):612–619, December 2010.

[Kü21] A. Küderle, N. Roth, and R. Richer. imucal - a python library to calibrate 6 dof imus (version 2.0.2) [computer software]. https://github.com/mad-lab-fau/imucal, 2021.

[Lau20] Timo Lauteslager, Stylianos Kampakis, Adrian J. Williams, Michal Maslik, and Fares Siddiqui. Performance Evaluation of the Circadia Contactless Breathing Monitor and Sleep Analysis Algorithm for Sleep Stage Classification. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5150–5153, July 2020. ISSN: 2694-0604.

[Len15] Gustavo Lenis, Felix Conz, and Olaf Dössel. Combining different ECG derived respiration tracking methods to create an optimal reconstruction of the breathing pattern. *Current Directions in Biomedical Engineering*, 1(1):54–57, September 2015.

[Lev03] Douglas Levine, Daniel Kripke, Robert Kaplan, Megan Lewis, Michelle Naughton, Deborah Bowen, and Sally Shumaker. Reliability and validity of women's health initiative insomnia rating scale. *Psychological assessment*, 15:137–48, 07 2003.

[Lew04] A.T. Lewicke, E.S. Sazonov, and S.A.C. Schuckers. Sleep-wake identification in infants: heart rate variability compared to actigraphy. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 442–445, September 2004.

[Li18] Qiao Li, Qichen Li, Chengyu Liu, Supreeth P. Shashikumar, Shamim Nemati, and Gari D. Clifford. Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram. *Physiological Measurement*, 39(12):124005, December 2018. Publisher: IOP Publishing.

[Li20] Xinyue Li, Yunting Zhang, Fan Jiang, and Hongyu Zhao. A novel machine learning unsupervised algorithm for sleep/wake identification using actigraphy. *Chronobiology International*, 37(7):1002–1015, July 2020.

[Li21] Li Li, Toru Nakamura, Junichiro Hayano, and Yoshiharu Yamamoto. Age and gender differences in objective sleep properties using large-scale body acceleration data in a Japanese population. *Scientific Reports*, 11:9970, May 2021.

[LT19] Tatjana Loncar-Turukalo, Eftim Zdravevski, José Machado da Silva, Ioanna Chouvarda, and Vladimir Trajkovik. Literature on Wearable Technology for Connected Health: Scoping Review of Research Trends, Advances, and Barriers. *Journal of Medical Internet Research*, 21(9):e14017, September 2019.

[Mal96] Marek Malik. Heart rate variability. *Annals of Noninvasive Electrocardiology*, 1(2):151–181, 1996.

[Man01] Renee L. Manser, Peter Rochford, Robert J. Pierce, Graham B. Byrnes, and Donald A. Campbell. Impact of Different Criteria for Defining Hypopneas in the Apnea-Hypopnea Index. *Chest*, 120(3):909–914, 2001.

[Mar11] Jennifer L. Martin and Alex D. Hakim. Wrist Actigraphy. *Chest*, 139(6):1514–1527, June 2011.

[McK10a] Patrick E. McKight and Julius Najab. *Kruskal-Wallis Test*, pages 1–1. John Wiley & Sons, Ltd, 2010.

[McK10b] Patrick E. McKnight and Julius Najab. *Mann-Whitney U Test*, pages 1–1. John Wiley & Sons, Ltd, 2010.

[Mor07] Timothy I. Morgenthaler, Teofilo Lee-Chiong, Cathy Alessi, Leah Friedman, R. Nisha Aurora, Brian Boehlecke, Terry Brown, Andrew L. Chesson, Vishesh Kapur, Rama Maganti, Judith Owens, Jeffrey Pancer, Todd J. Swick, and Rochelle Zak. Practice Parameters for the Clinical Evaluation and Treatment of Circadian Rhythm Sleep Disorders. *Sleep*, 30(11):1445–1459, November 2007.

[Mos09] Doris Moser, Peter Anderer, Georg Gruber, Silvia Parapatics, Erna Loretz, Marion Boeck, Gerhard Kloesch, Esther Heller, Andrea Schmidt, Heidi Danker-Hopfe, Bernd Saletu, Josef Zeitlhofer, and Georg Dorffner. Sleep Classification According to AASM and Rechtschaffen & Kales: Effects on Sleep Scoring Parameters. *Sleep*, 32(2):11, 2009.

[Ore14] G. Orellana, C. M. Held, P. A. Estevez, C. A. Perez, S. Reyes, C. Algarin, and P. Peirano. A balanced sleep/wakefulness classification method based on actigraphic data in adolescents. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4188–4191, Chicago, IL, August 2014. IEEE.

[oSDRU93] National Commission on Sleep Disorders Research (U.S.) and United States, editors. *Wake up America: a national sleep alert: report of the National Commission on Sleep Disorders Research*. The Commission, Washington, D.C., 1993.

[Pal19] Joao Palotti, Raghvendra Mall, Michael Aupetit, Michael Rueschman, Meghna Singh, Aarti Sathyanarayana, Shahrad Taheri, and Luis Fernandez-Luque. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *npj Digital Medicine*, 2(1):50, December 2019.

[Ped11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[PH13] T. Porkka-Heiskanen, K.-M. Zitting, and H.-K. Wigren. Sleep, its regulation and possible mechanisms of sleep disturbances. *Acta Physiologica*, 208(4):311–328, 2013.

[Resa] National Sleep Research Resource. Multi-ethnic study of atherosclerosis - actigraphy introduction. https://sleepdata.org/datasets/mesa/pages/actigraphy-introduction.md. Accessed: 2021-11-07.

[Resb] National Sleep Research Resource. Multi-ethnic study of atherosclerosis - dateset introduction. https://sleepdata.org/datasets/mesa/pages/dataset-introduction.md. Accessed: 2021-11-07.

[Resc] National Sleep Research Resource. Multi-ethnic study of atherosclerosis - polysomnography introduction. https://sleepdata.org/datasets/mesa/pages/polysomnography-introduction.md. Accessed: 2021-11-07.

[Ric21] Robert Richer, Arne Küderle, Martin Ullrich, Nicolas Rohleder, and Bjoern M. Eskofier. Biopsykit: A python package for the analysis of biopsychological data. *Journal of Open Source Software*, 6(66):3702, 2021.

[Rob21] Robin Champseix. Heart rate variability analysis. https://pypi.org/project/hrv-analysis/, 2021.

[Rot18] Nils Roth, Christine F. Martindale, Bjoern M. Eskofier, Heiko Gaßner, Zacharias Kohl, and Jochen Klucken. Synchronized Sensor Insoles for Clinical Gait Analysis in Home-Monitoring Applications. *Current Directions in Biomedical Engineering*, 4(1):433–437, September 2018. Publisher: De Gruyter.

[Rus16] Nicole Rusk. Deep learning. *Nature Methods*, 13(1):35–35, January 2016.

[Sad94] A. Sadeh, K. M. Sharkey, and M. A. Carskadon. Activity-Based Sleep-Wake Identification: An Empirical Test of Methodological Issues. *Sleep*, 17(3):201–207, 1994.

[Sad11] Avi Sadeh. The role and validity of actigraphy in sleep medicine: An update. *Sleep Medicine Reviews*, 15(4):259–267, August 2011.

[San19] A. Sano, W. Chen, D. Lopez-Martinez, S. Taylor, and R. W. Picard. Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1607–1617, July 2019. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[Saz02] N.A. Sazonova, E.S. Sazonov, and S.A.C. Schuckers. Activity-based sleep-wake identification in infants. In *Computers in Cardiology*, pages 525–528, Memphis, TN, USA, 2002. IEEE.

[Saz04] Edward Sazonov, Nadezhda Sazonova, Stephanie Schuckers, and Michael Neuman. Activity-based sleep? Wake identification in infants. *Physiological measurement*, 25:1291–304, November 2004.

[Sch08] Axel Schäfer and Karl W. Kratky. Estimation of breathing rate from respiratory sinus arrhythmia: comparison of various methods. *Annals of Biomedical Engineering*, 36(3):476–485, March 2008.

[Seh11] Amita Sehgal and Emmanuel Mignot. Genetics of sleep and sleep disorders. *Cell*, 146(2):194–207, July 2011.

[Sil08] A. Silva, M. L. Andersen, M. T. De Mello, L. R. A. Bittencourt, D. Peruzzo, and S. Tufik. Gender and age differences in polysomnography findings and sleep complaints of patients referred to a sleep laboratory. *Brazilian Journal of Medical and Biological Research*, 41:1067–1075, December 2008. Publisher: Associação Brasileira de Divulgação Científica.

[Sil13] Alessandro Silvani and Roger A. L. Dampney. Central control of cardiovascular function during sleep. *American Journal of Physiology-Heart and Circulatory Physiology*, 305(12):H1683–H1692, December 2013. Publisher: American Physiological Society.

[Sin15] Shirshendu Sinha, Ronak Jhaveri, and Alok Banga. Sleep Disturbances and Behavioral Disturbances in Children and Adolescents. *Psychiatric Clinics of North America*, 38(4):705–721, December 2015.

[Sou17] Jane Souza, Maria Luiza Cruz de Oliveira, Ivanise Sousa, and Carolina Azevedo. Gender differences in sleep habits and quality and daytime sleepiness in elementary and high school teachers. *Chronobiology International*, 35:1–13, December 2017.

[Sta05] Neil Stanley. The physiology of sleep and the impact of ageing. *European Urology Supplements*, 3(6):17–23, January 2005.

[Sto17] Katie L. Stone and Sonia Ancoli-Israel. Chapter 171 - actigraphy. In Meir Kryger, Thomas Roth, and William C. Dement, editors, *Principles and Practice of Sleep Medicine (Sixth Edition)*, pages 1671–1678.e4. Elsevier, sixth edition edition, 2017.

[Sui19] Yu Sui, Mengze Yu, Haifeng Hong, and Xianxian Pan. Learning from Imbalanced Data: A Comparative Study. In Weizhi Meng and Steven Furnell, editors, *Security and Privacy in Social Networks and Big Data*, Communications in Computer and Information Science, pages 264–274, Singapore, 2019. Springer.

[Tan01] H. Tanaka, K. D. Monahan, and D. R. Seals. Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology*, 37(1):153–156, January 2001.

[Til09] Joëlle Tilmanne, Jérôme Urbain, Mayuresh V. Kothare, Alain Vande Wouwer, and Sanjeev V. Kothare. Algorithms for sleep-wake identification using actigraphy: A comparative study and new results. *Journal of Sleep Research*, 18(1):85–98, 2009.

[tL13] Bart H. W. te Lindert and Eus J. W. Van Someren. Sleep Estimates Using Microelectromechanical Systems (MEMS). *Sleep*, 36(5):781–789, May 2013.

[Tra14] Travis Cooper. Understanding sleep for optimal recovery & productivity. https://www.catalystathletics.com/articles/images/2014-04-14-cooperSleep.jpg, 2014. Accessed: 2021-11-01.

[Val21] Raphael Vallat and Matthew P Walker. An open-source, high-performance tool for automated sleep staging. *eLife*, 10:e70092, October 2021.

[Web82] John B. Webster, Daniel F. Kripke, Sam Messin, Daniel J. Mullaney, and Grant Wyborney. An Activity-Based Sleep Monitor System for Ambulatory Use. *Sleep*, 5(4):389–399, September 1982.

[WF83] Johanna Wilde-Frenz and Hartmut Schulz. Rate and Distribution of Body Movements during Sleep in Humans. *Perceptual and Motor Skills*, 56(1):275–283, February 1983. Publisher: SAGE Publications Inc.

[Wid18] Edita Rosana Widasari, Koichi Tanno, and Hiroki Tamura. Automatic Sleep Stage Detection Based on HRV Spectrum Analysis. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 869–874, October 2018. ISSN: 2577-1655.

[Yer97] Vikram K Yeragani, Edward Sobolewski, Jerald Kay, V.C Jampala, and Gina Igel. Effect of age on long-term heart rate variability. *Cardiovascular Research*, 35(1):35–42, July 1997.

[Zha18] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association : JAMIA*, 25(10):1351–1358, May 2018.

[Zha20] Bing Zhai, Ignacio Perez-Pozuelo, Emma A. D. Clifton, Joao Palotti, and Yu Guan. Making Sense of Sleep: Multimodal Sleep Stage Classification in a Large, Diverse Population Using Movement and Cardiac Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–33, June 2020.