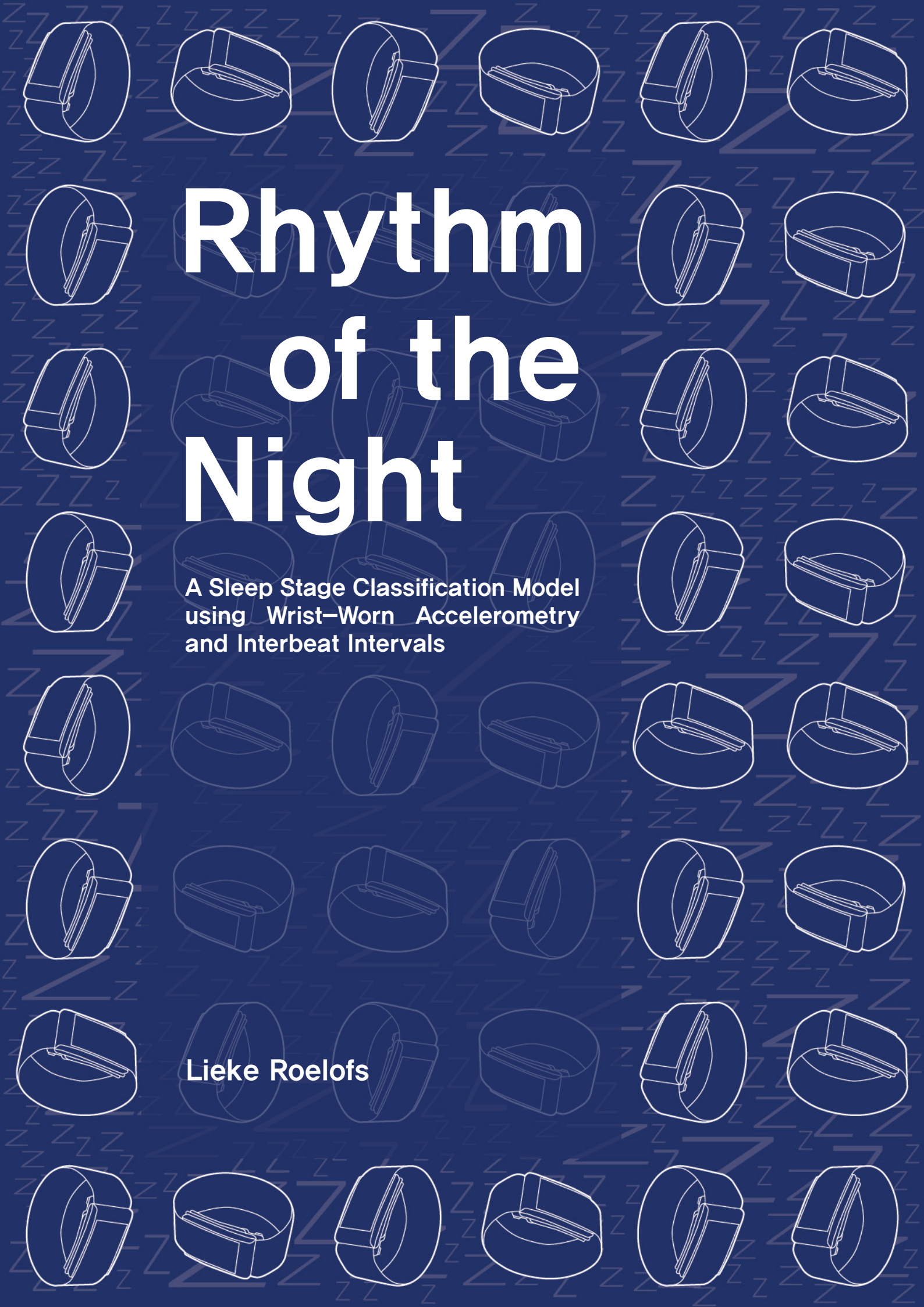


Rhythm of the Night

A Sleep Stage Classification Model using Wrist-Worn Accelerometry and Interbeat Intervals

Lieke Roelofs

Rhythm of the Night

Developing a Sleep Stage Classification Model Using Wrist-Worn Accelerometry and Interbeat Intervals

by

Lieke Roelofs

to obtain the degree of Master of Science in Biomedical Engineering
at the Delft University of Technology,
to be defended publicly on Thursday May 23, 2024 at 10:45 AM.

Student number: 4547403
Thesis committee: Dr. B. Hunyadi, TU Delft, supervisor
Dr. M. Kok, TU Delft
MSc. K. Ebrahimkheil, Corsano Health, supervisor

This thesis is confidential and cannot be made public until June, 2026.

TU Delft

Contents

Nomenclature

AASM American Academy of Sleep Medicine

BMI Body Mass Index

CB CatBoost

CFS Correlation-based Feature Selection

ECG Electrocardiography

EEG Electroencephalography

HF High Frequency

HRV Heart Rate Variability

IBI Interbeat Interval

LF Low Frequency

MAE Mean Absolute Error

MDA Mean Directional Accuracy

METC Medical Ethics Commitee

NREM Non-Rapid Eye Movement

OC Overlap Coefficient

OSA Obstructive Sleep Apnea

PG Polygraphy

PPG Photoplethysmography

PPI Peak-to-Peak Interval

PSG Polysomnography

REM Rapid Eye Movement

RF Random Forest

RLS Restless Leg Syndrome

SPT Sleep Period Time

SVM Support Vector Machine

VLF Very Low Frequency

1

Introduction

Sleep is fundamental to human health and plays a vital role in our physical and cognitive functions. It occupies approximately one-third of an individual's life, serving as a period of recovery and regeneration for both body and mind. The process of sleep is natural but intricate, characterized by alternating stages known as sleep architecture. Each stage is defined by characteristic patterns at cerebral, cardiac, and respiratory levels. Despite its significance, the complexity of sleep and its various stages remains a subject of extensive research. Sleep disorders are known to affect these patterns. For example, a reduction in one of the sleep stages called 'slow-wave sleep' might be the earliest observable symptoms of Alzheimer's disease and tend to surface before or soon after the diagnosis of cognitive impairment [1]. Currently, sleep disorders pose a significant public health challenge, leading to increased risks of cardiovascular and cognitive diseases, increasing the likelihood of workplace accidents, and a decline in overall quality of life. Consequently, monitoring an individual's sleep architecture can play a critical role in tackling these significant public health issues.

1.1. Problem Statement

To monitor and objectively measure sleep architecture, polysomnography (PSG) is used. PSG involves the use of multiple sensors, such as electroencephalography (EEG) and electrocardiography (ECG), and typically measures over the course of a single night. The recording is then split into segments of 30 seconds, each referred to as an epoch. Subsequently, each epoch is visually scored by trainned personnel according to specific criteria. Due to the intrusiveness of the multitude of sensors, PSG setups can interfere with natural sleep patterns. Additionally, the manual scoring of PSG recordings is both labor-intensive and time-consuming. These limitations underscore the need for advancements in sleep monitoring technologies, particularly in developing methods that are both accurate and user-friendly, while overcoming the intrusiveness of the sensors needed.

Figure 1.1: Cartoon showing the downside of PSG

Recent efforts have focused on combining alternative physiological signals and machine-learning techniques for automatic sleep staging to overcome the time-consuming process of manual scoring [2]. Wearable devices have been developed to reduce the number of required sensors, making them less intrusive. Simultaneously, research has been conducted to automate the scoring process and reduce manual interventions [3].

In the current landscape, many commercial devices and applications promise insights into individual sleep architecture. Despite their widespread availability and use, it's crucial to note that none have received medical approval for clinical or healthcare research applications. This gap underscores a significant challenge of using technology for sleep analysis within the medical context, where accuracy and reliability are crucial.

This research is conducted in cooperation with Corsano Health, a MedTech company that develops, produces, and markets a wrist-worn wearable device. Their most recent product, the CardioWatch 287-2, is a wireless monitoring system intended for the continuous collection of vital parameters using wireless, non-invasive technology. The CardioWatch is EU-MDR-CE and FDA certified and can therefore be used in home and healthcare settings. The product is equipped with several algorithms, such as respiration rate and blood pressure. Despite its advanced capabilities, the device's existing sleep algorithm is susceptible to drawbacks: due to the external origin of the current sleep algorithm, challenges arise in addressing atypical results within the algorithm's performance. This underscores the desire to develop an in-house sleep algorithm.

1.2. Research Objective

The primary aim of this thesis is to develop a four-stage sleep classification model for the Corsano CardioWatch 287-2, addressing the critical need for accurate, efficient, and user-friendly sleep stage classification. This research unfolds into several interconnected sub-objectives, designed to tackle specific challenges within the domain of sleep monitoring:

1. **Automated Sleep Period Time Extraction**

   This initial step focuses on using existing methodologies to automatically determine the sleep period time (SPT) window from continuous data. The algorithm automates the sleep data segmentation process, eliminating the need for manual input. This enhances the algorithm's efficiency, and applicability. The development of this automated extraction process sets the foundation for accurate sleep stage classification by ensuring that the algorithm operates within precisely identified periods of sleep.

2. **Development of Four-Stage Sleep Classification Models**

   This objective involves of constructing different classification models using accelerometer and the interbeat interval of the photoplethysmography (PPG) sensor on the CardioWatch to overcome the limitations of PSG. This includes fine-tuning of model hyperparameters and feature selection to improve performance.

3. **Experimental Design and Data Collection**

   This objective is central to the thesis, focusing on the careful planning and execution of the study to gather high-quality data essential for the algorithm's development and evaluation. It includes the essential step of obtaining approval from the Medical Ethics Committee. This phase is crucial for ensuring the reliability of the collected data and, by extension, the validity of the algorithm's performance metrics. By creating a custom dataset, this approach directly addresses potential inconsistencies arising from using data collected by different sensors, thereby enhancing the accuracy and reliability of the classification results.

4. **Evaluating Model Performance**

   The performance of the developed algorithm will be critically evaluated against the gold standard PSG and available commercial algorithms. This evaluation aims to assess the model's accuracy, reliability, and potential superiority or competitiveness in the field of sleep stage classification.

5. **Exploring the Influence of Features on Classification**

   Aiming to deepen our understanding of the complexity of sleep, the significance of certain physiological signals and their variations, represented by features, across different sleep stages are analysed. This exploration directly responds to the challenge of enhancing algorithmic accuracy and interpretability, guiding future improvements and refinements.

Each objective is designed to iteratively build upon the knowledge and outcomes of the previous, facilitating a systematic approach to tackling the overarching research question. The progression from automated SPT window extraction to an adaptable classification framework establishes a robust foundation for the algorithm. Subsequent objectives focus on refining and validating this foundation through careful experimental design, performance evaluation, and feature exploration.

2

Background

This chapter aims at providing more detailed background information where the thesis is build upon. First, it is important to understand the physiology of sleep, how sleep affects your vital parameters and the terminologies that are commonly used. Secondly, the gold standard sleep monitoring method PSG will be discussed, followed by wearable technology for sleep monitoring. Concluding with the explanation of the machine learning methods used in this thesis.

2.1. Sleep

2.1.1. Sleep architecture: understanding the stages

Although we perceive sleep as an uniform state of rest, it is actually a complex, cyclic process crucial for various physiological functions. It is characterized by alternating stages, each characterized by distinct cerebral, cardiac, and respiratory patterns. Sleep can be divided into non-rapid eye movement (NREM) and rapid eye movement (REM) sleep. NREM sleep is further categorized into stages 1, 2, and 3, with each stage representing a deeper level of sleep (Figure 2.1). This figure also shows that the stages are not equally divided: approximately 75-80% of total time spent in sleep is constituted by NREM sleep, with REM sleep covering the remaining 20-25%. The length of a sleep cycle tends to increase over the course of the night, leading to a shift in the distribution of sleep stages. As the night progresses, REM sleep duration expands, while stage 2 begins to account for the majority of NREM sleep, and stages 3 may sometimes even disappear [5].

NREM stages 1 and 2 are considered light sleep, with stage 1 serving as a transition and stage 2 playing a pivotal role in memory consolidation [6]. These stages are often grouped together because they both represent less intensive phases of sleep. Stage 3, or deep sleep, is essential for physical restoration and immune function. The cycling between NREM and REM stages throughout the night is a sophisticated process critical for overall well-being, and is often referred to as sleep architecture. The sleep architecture can be captured in a hypnogram, which displays the alternations of different

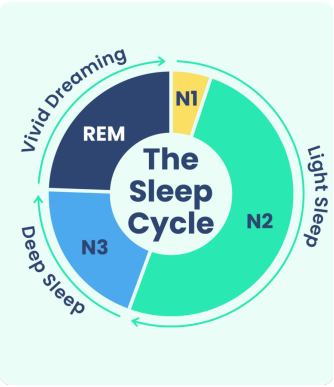


Figure 2.1: Representation of the sleep cycle [4]

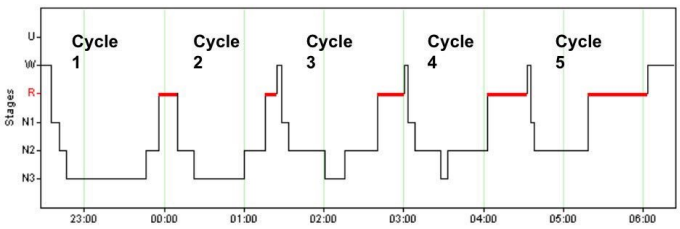


Figure 2.2: Hypnogram

sleep stages over the course of a night (Fig 2.2). Deviations in sleep architecture are linked to various sleep disorders, and can be found while studying one's hypnogram. Disruptions in these stages are associated with an array of health problems, including cognitive impairments, mood disorders, and increased risk of chronic illnesses, emphasizing the need for monitoring sleep using an accurate sleep stage classifier [7].

2.1.2. Traditional Monitoring of Sleep Architecture

PSG is the de facto "gold standard" for objective measurement of sleep. Standard PSG measurements involve multiple sensors, including at least electroencephalography (EEG), electrooculography (EOG), electrocardiography (ECG), and electromyography (EMG), which are usually recorded over the course of a single night. These recordings are split into epochs (30-second segments), and scored using the scoring criteria established by *Rechtschaffen and Kales*, which are modified in 2007 by *American Academy of Sleep Medicine (AASM)* [8]. Each stage is defined by specific brain-wave patterns in combination with additional data on respiratory effort, oxygen saturation and heart rate. These epochs are manually scored and therefore produces potential for variability in interpretation. Research has shown that there still remains an inter-rater reliability agreement of 82% despite adhering to these rules and using skilled personnel [9]. Its use is also hampered by other challenges, such as the equipment's intrusiveness which often affects the natural sleep process.

Due to the complexity of PSG, not all patients suspected of having sleep disorders require this test. Instead, simpler methods are used. Specifically, those suspected of having obstructive sleep apnea (OSA)—a condition where breathing stops temporarily, prompting the brain to awaken the individual to reopen the airway—can be diagnosed using polygraphy (PG). PG, a simplified version of PSG, does not measure brainwaves but focuses on oxygen levels (SpO2), heart rate, and respiratory effort through airflow sensors and belts. Despite OSA being characterized by numerous awakenings, its diagnosis relies on evaluating respiratory effort, airflow, and oxygen saturation, without the need to track sleep architecture disruptions. Although OSA constitutes a substantial proportion of sleep disorders, patients with suspected severe apnea, insomnia or other complex sleep conditions still rely on PSG.

2.1.3. Wearable Technology in Sleep Studies

In response to the limitations of the traditional sleep monitoring method PSG, recent years have seen a significant shift toward wearable technology. Devices equipped with sensors like photoplethysmography (PPG) and accelerometry offer a less intrusive means of gathering data on sleep patterns and vital signs. This technology holds promise of simplifying sleep monitoring, making it more accessible and less burdensome for individuals. Notably, the amount of movement have been widely studied using accelerometers, as periods of inactivity corresponds to periods of sleep. Additionally, research indicates a strong correlation between sleep patterns and vital signs, particularly heart rate and heart rate variability (HRV) [2]. It is well documented that the heart rate slows down and becomes more regular during NREM sleep, whereas during REM sleep, the heart rate is elevated and is more varying. The heart rate can be measured using a PPG sensor. In the next subsections, these two sensors along with their relation to sleep will be explained.

Accelerometry

The first wearable sleep monitoring systems were equipped with accelerometers. These devices, characterized by their ability to measure acceleration due to movement across three axes (x, y, and z), function through a suspended mass that shifts upon motion detection. This displacement is then translated into an electrical signal, typically generated through piezoelectric microcrystals or capacitive elements that convert movement into electrical charges. Positioned commonly on the wrist or ankles, they are optimal for detection of limb movements, and therefore serve as key devices for distinguishing between periods of wakefulness and rest based on activity levels. This capability to detect varying degrees of activity enables accelerometers to identify times of high activity, indicative of wakefulness or engagement in activities, and periods of low or minimal activity, suggesting rest. As such, they can estimate sleep duration by differentiating active and restful states.

Overall, accelerometers are prized for their non-intrusive nature, cost-effectiveness, and simplicity, making them an ideal choice for both research applications and home monitoring in sleep studies.

The analysis of accelerometer data for sleep studies involves both linear features, such as movement intensity, and nonlinear features, derived from calculations over time windows with thresholds. This mix demands a modeling approach capable of handling the diversity of data characteristics. Therefore, the selection of an appropriate model is crucial, requiring an algorithm versatile enough to interpret both the straightforward and complex patterns inherent in the data to accurately assess sleep patterns.

Photoplethysmography

Photoplethysmography (PPG) is an advanced, non-invasive optical technique used to monitor changes in blood volume within the skin's microvascular bed. The PPG sensor emits light—typically infrared or green—into the skin and detects the amount of light that is absorbed or reflected by the blood vessels with a photodetector. As blood pulsates through the vessels with each heartbeat, the sensor captures the variations in light absorption, producing a waveform that reflects cardiovascular activity. The highest peak corresponds with the systolic peak, and the time interval between successive systolic peaks corresponds to the time between two consecutive heartbeats, referred to as peak-to-peak intervals (PPI) or interbeat-intervals (IBI) (Figure 2.3). In this study we use the term IBIs. As the variations in reflected light intensity directly correlate with fluctuations in blood volume, valuable insights into cardiovascular health are provided.

The Corsano CardioWatch 287-2 PPG sensor contains infrared, near-infrared, and green light sources. Infrared and near-infrared light are preferred for deeper tissue penetration and is crucial in measuring blood oxygen levels. Green light has more superficial tissue penetration, and therefore presents a lower susceptibility to motion artifacts. This enhances the signal-to-noise ratio and ensures more reliable data collection in sleep studies and other applications where minimizing interference from user movement is crucial. For the purposes of this thesis, the focus will be exclusively on the green light component.

From the raw PPG signal, the IBIs are calculated. These IBIs are used to determine heart rate variability (HRV), which will be further explained in the following subsection.

2.1.4. Heart Rate Variability

HRV stands as a critical physiological marker, intricately linked with various stages of sleep. Unlike the more pronounced fluctuations observed in brainwave activity, changes in heart rate are considerably more subtle, making manual visual scoring of sleep stages by human observers nearly impossible. However, these minute variations can be effectively captured and analyzed through HRV metrics.

The application of PPG for HRV analysis, specifically through the examination of IBIs, offers a non-invasive and practical approach to understanding heart rate fluctuations. In sleep research, HRV offers insights into the autonomic nervous system's activity during sleep. It is captured through two types of features: time domain and frequency domain. Time domain features assess the variations in time intervals between successive heartbeats. Meanwhile, frequency domain features analyze the distribution of power across various frequency bands.

Given the inherent complexity of HRV data, capturing both linear and nonlinear physiological dynamics, selecting an analytical model that can adeptly handle these multifaceted relationships becomes crucial. Time domain features may lean towards linear characteristics, suitable for traditional modeling techniques, while frequency domain and nonlinear metrics demand models capable of navigating complex nonlinear interactions. This underscores the significance of choosing a model that can handle both linear and nonlinear features.

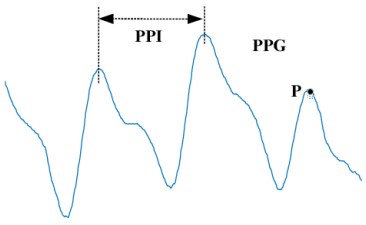


Figure 2.3: A PPG signal illustrating the meaning of IBIs, referred to as PPI in this figure. [10].

2.2. Machine Learning

In the selection of machine learning models for sleep stage classification, Support Vector Machines (SVM), Random Forest (RF), and CatBoost emerged as leading candidates from my preceding literature review [11]. Each of these methods offers unique advantages for handling the complexities inherent in sleep data analysis and are explained below:

- **Support Vector Machine** stand out in the realm of supervised machine learning for their proficiency in identifying the optimal line or plane that maximizes the distance between differing classes. This capability renders them highly effective for complex classification tasks, such as sleep stage classification, by adeptly handling high-dimensional spaces. The strategic use of kernel functions allows SVMs to excel in environments where data points are nonlinearly separable. Renowned for robust performance across varying dataset sizes, SVMs exhibit resistance to overfitting and demonstrate exceptional efficiency in managing noisy data.
- **Random Forest** is an ensemble model, meaning multiple decision trees are created to predict one outcome. Each decision tree is made from a subset of the samples in the training data, leveraging the combined predictions of numerous decision trees to boost accuracy and robustness. It is particularly good at dealing with nonlinear relationships and high-dimensional data. RF's resilience to overfitting, capability to handle outliers and missing values, and insights into feature importance make it a valuable tool for understanding sleep data patterns.
- **CatBoost** is an algorithm for gradient boosting on decision trees. It's name is coined from *"Category"* and *"Boosting"*, representing its ability to perform well on categorical data using a boosting technique. Boosting is an iterative, sequential, and adaptive technique that fixes its predecessor's error. CatBoost's robustness against noisy data and its efficacy in dealing with missing values are particularly beneficial for sleep stage classification tasks that involve heterogeneous datasets.

These models were specifically chosen for their demonstrated capabilities in existing research and their alignment with the dual goals of achieving high accuracy and computational efficiency in sleep stage classification. Each model contributes uniquely to the algorithm's design: SVM for its boundary precision, RF for its robustness and interpretability, and CatBoost for its innovative approach to data diversity. This selection process ensures a comprehensive analytical framework all with their own way to tackel the challenges for sleep stage classification.

3

Methodology

In this chapter the methodology is explained to achieve the objectives outlined in section 1.2. This chapter begins with a description of the study design in 3.1, aimed at creating a dataset appropriate for this thesis. This is followed by an overview of the data collection and preprocessing methods for sensor data, detailed in Section 3.2, which together cover Research Objective 3. Subsequently, the algorithm design is discussed which is divided in methods for extracting the SPT window (3.3.1) (Research Objective 1) and methods for the four-stage classification model (3.3.2), of which its feature selection method will be described in 3.3.3 and the hyperparameter tuning in 3.3.4 (Research Objective 2). This chapter further delves into the evaluation of these methodologies, covering the analysis of automatic SPT window extraction (3.4.1) and the four-stage clasisifcation model (3.4.2), along with a comparitive assessments against commercially available algorithms (3.4.3) (Research Objective 4). Lastly, an analysis of the relationships between various sleep stages and corresponding physiological signals will be described in 3.4.4 (Research Objective 5).

3.1. Study Design

This study was initiated with the objective of assembling a dataset through simultaneous recordings from the Corsano CardioWatch and PSG. The simultaneous data acquisition from the Corsano CardioWatch and PSG sets the stage for utilizing supervised machine learning methods to examine the correlations among HRV, activity levels, and sleep phenomena.

The collaboration was initiated with Slaapapneu Service, a healthcare provider specializing in the diagnosis and treatment of sleep-related disorders, including insomnia and sleep apnea, with multiple facilities across the Netherlands.

The study's inclusion criteria were specifically limited to participants who had undergone a PSG assessment at Slaapapneu Service in South-Holland during the months of January to March 2024, who could give informed consent. Exclusion criteria were recordings with missing or having poor quality data. For PSG data quality data was judged by the somnologist of Slaapapneu Service. For the Corsano CardioWatch, quality was defined using the signal quality index, or missing data in excess of 50% of the nocturnal data.

Within the Slaapapneu Service facilities not every patient is subject to PSG, as delineated in Chapter 2.1.2, individuals suspected of having OSA typically receive a PG assessment. PSG is reserved for those diagnosed with severe OSA or when there is a suspicion of additional sleep disorders. Consequently, the dataset will be biased towards patients with significant sleep disorders, and thus contains a comparatively small sample of healthy individuals.

The methodological framework is outlined in Figure 3.1. Subjects scheduled for PSG at Slaapapneu Service were asked during their visit to participate in this research. Participation included wearing the CardioWatch over the course of one night in addition to the PSG recording. Subjects were provided with detailed instructions and an informational document. This preparatory material emphasized that the integration of the CardioWatch into their PSG assessment would not deviate from their original treat-

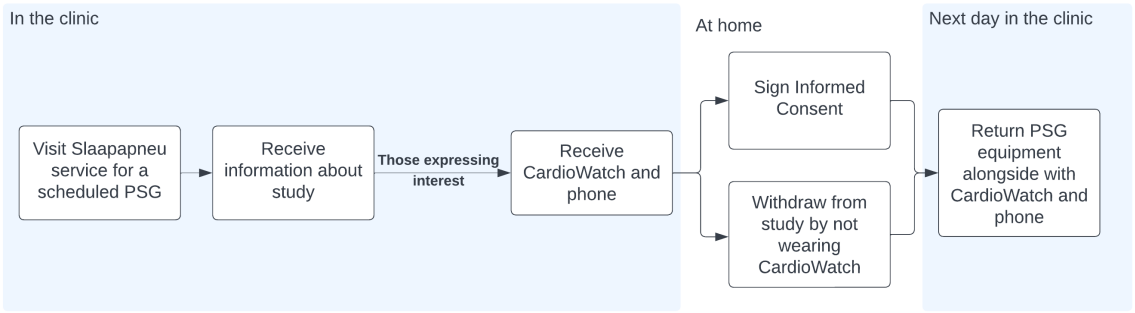


Figure 3.1: Flowchart of the study design, detailing the participant road map. The process starts with a scheduled PSG at the clinic, follows with information dissemination and device distribution at home, and concludes with the return of study equipment at the clinic the next day.

ment protocol, ensuring that the study's procedures aligned seamlessly with standard care practices. Subjects who expressed interest were equipped with the CardioWatch. The CardioWatch requires connection to a smartphone for data collection. To streamline this process and ensure compatibility, each CardioWatch was paired with a provided smartphone and pre-configured to facilitate immediate use and data synchronization. Participants were instructed to return both the CardioWatch and the paired smartphone, along with their signed informed consent and the PSG equipment, the following day. It was explicitly stated that participants retained the right to withdraw from the study at any point by removing the CardioWatch and refraining from signing the informed consent upon equipment return.

The study's protocol, along with the consent form, underwent rigorous review by the Medical Ethics Committee (METC) of Leiden-Rotterdam-Den Haag, which affirmed the study's compliance with ethical standards, ensuring the safety and rights of participants while evaluating the potential risks and benefits of participation. Detailed information on the procedures and required documents is available in the "Onderzoekers-checklist voor WMO/MDR/IVDR onderzoeksdossier" [12].

I personally managed the entire process required to secure METC approval, a task that extended over a considerable period of the time. In addition to the METC application, I oversaw the coordination of the sleep study, which involved multiple responsibilities. These included direct communication with Slaapapneu Service to ensure seamless coordination. My responsibilities also extended to the inclusion of participants. This role necessitated frequent travel across South-Holland to meet with potential participants, facilitate the distribution of the CardioWatch devices and phones, and ensure proper handling and storage of data.

3.2. Data Collection and Preprocessing

Data collection was conducted using two methods: via a PSG device and via the Corsano CardioWatch..

The PSG devices used in this study were the Löwenstein Miniscreen Pro and the Löwenstein Sonata, and is performed using standard procedure [13], including the following sensors:

- A chest and abdominal belt
- A pulse oximeter
- An air cannula to measure airflow and snoring
- 5x EEG electrodes
- 4x EMG electrodes, two next to the eyes and two under the chin

The analysis of PSGs were conducted by a somnologist at Slaapapneu Service, adhering to the AASM standards for sleep classification [8]. One somnologist was responsible for scoring all the PSG recordings, ensuring uniformity in the analysis. The analysis was done using the Miniscreen Viewer

5.2.3 software, after which the sleep stages, accompanied by their respective timestamps, were manually exported for further analysis. PSG data categorizes sleep into five stages. To match this to a four-stage analysis, stages N1 and N2 were combined into the singular category *"light sleep"*, while the remaining classes (*"wake", N3 (= "deep sleep"*), and "REM") remained unchanged.

The Corsano CardioWatch was configured to collect raw accelerometer and PPG data at a sampling rate of 32 Hz, which were then stored on the cloud. An existing algorithm processed the raw PPG signal to calculate the IBIs, which were also uploaded to the cloud. Subsequently, the raw accelerometer data along with the IBIs were later manually extracted as .csv files from this cloud. The IBIs were further filtered to exclude values below 300 ms and above 1650 ms, corresponding to heart rates outside the normal range of 30 to 200 beats per minute.

3.2.1. Demographic Characteristics of the Dataset

In the period from January to March 2024, 27 participants were included in the study. Three of them did not meet the quality requirements for inclusion. The remaining 24 participants constituted the cohort for this study.

For each participants, the following basic demographic information is collected: gender, age, weight, and height, after which the Body Mass Index (BMI) is calculated for each participant using the formula: $\frac{\text{weight (kg)}}{\text{height (m)}^2}$.

The distribution across gender is skewed towards males, with 14 male participants compared to 10 female participants. Age distribution varied, with the largest group (12 participants) falling within the 30-49 group, followed by 6 participants in the 50-64 group. The study also included younger and older age groups, with 1 participant under 18, 2 between 18-19, and 3 aged 65 or older. Regarding BMI, a majority of the subjects (14 participants) have a BMI of less than 30, indicating a prevelance of non-obesity, while 10 participants had a BMI greater than 30, which is indicative of obesity. Assessing the different sleep disorders that got diagnosed after the PSG, the study found 5 participants without apnea, 9 with light apnea, 7 with mild apnea, and 3 with severe apnea. Additionally, the prevelance of Restless Leg Syndrome was low, with only 1 participant having a moderate level and 2 participants experiencing severe levels, while the remaining 21 participants did not exhibit this syndrome. This information is outlined in Table 3.1.

Table 3.1: Demographic Characteristics of Participants

Total participants	**24**
Gender	
Female	10
Male	14
Age Range	
18-	1
18-29	2
30-49	12
50-64	6
65+	3
BMI	
< 30	14
30 or higher	10
Apnea	
No apnea	5
Light	9
Mild	7
Severe	3
Restless Leg Syndrome	
Moderate	1
Severe	2
No RLS	21

3.2.2. Balancing the Dataset

The distribution among the different sleep stages is not equal, as explained in Section 2.1.1. To address class imbalance before model training, a balancing approach was employed for the training set. Specifically, the sleep stage with lowest epoch count was identified for each participant. The epoch count across all stages was then standardized by extracting an equivalent number of epochs from the

stages with higher counts. This method ensured that each participant's dataset had a uniform distribution of stages, resulting in a balanced training set. However, it may also lead to a scenario where the data predominantly reflects the characteristics of certain individuals more than others.

Unlike the training set, the validation and test sets were not balanced. Instead, to accommodate this imbalance, weighted metrics were employed, detailed in the subsequent subsection.

3.2.3. Dividing the Dataset

The dataset will be divided into a training set and a validation set. The first 19 participants form the training set, where the remaining 5 participants form the validation set. For the application for the four-stage classification model, the training set will be split using a four-fold cross-validation strategy for feature selection and hyperparameter tuning. The 4-fold cross-validation technique provides a good balance between having a meaningful and diverse size of the test set and allowing the model to learn from a sufficiently large portion of the data.

The division into training, test and validation subsets was executed on participant level, ensuring the epochs of individual patients were exclusively allocated to a singular group.

The validation set comprised five participants: three females and two males, aged 19, 43, 59, 61, and 65. Among these participants, three have mild apnea, and two do not have apnea. The validation set is thus skewed towards older individuals.

3.3. Algorithm Design

The methodology of the proposed algorithm is subdivided into various substeps, outlined in Figure 3.2. Initially, the algorithm calculates the SPT window using continuous accelerometer data. This calculated window is then utilized to segment the accelerometer and continuous PPI data. Features are calculated over predefined segments within this window, which are then processed by the classifier. This classifier is responsible for translating the feature values into a predictive hypnogram. The following sections will comprehensively explore each substep.

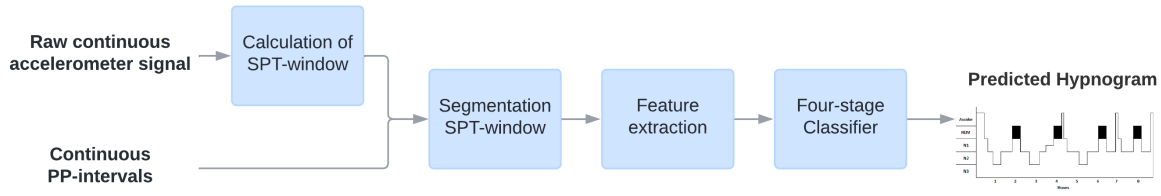


Figure 3.2: Schematic overview of the algorithm design. Starting with processing raw accelerometer data to determine the SPT window. Subsequently, this window is segmented from the continuous accelerometer and PP-intervals data. Feature extraction is then performed on these segments before they are inputted into the four-stage classifier. Finally, the output yields a hypnogram.

3.3.1. Automatic SPT window Extraction

The SPT window is the time window starting at sleep onset and ending when waking up after the last sleep episode of the night. Automatic SPT window extraction is desirable to minimize the amount of data input to the classification model. Automatic extraction of the SPT window can be seen as a preprocessing step. Two promising methods have been selected based on prior literature research [11], being the *angle*$_z$ method and the method described by the company Actigraph, which I will refer to as the *Actigraph counts* method. These methods will be explained in the following subsections.

Angle$_z$

The angle$_z$ method, developed by van Hees et al. [14], is used to estimate the arm angle called angle$_z$. The z-axis corresponds to the axis perpendicular to the skin surface, specifically the dorsal-ventral direction when the wrist is in the anatomical position. The angle$_z$ is estimated using the following formula:

$$angle_z = \tan^{-1} \frac{a_z}{a_x^2 + a_y^2} \cdot 180/\pi \tag{3.1}$$

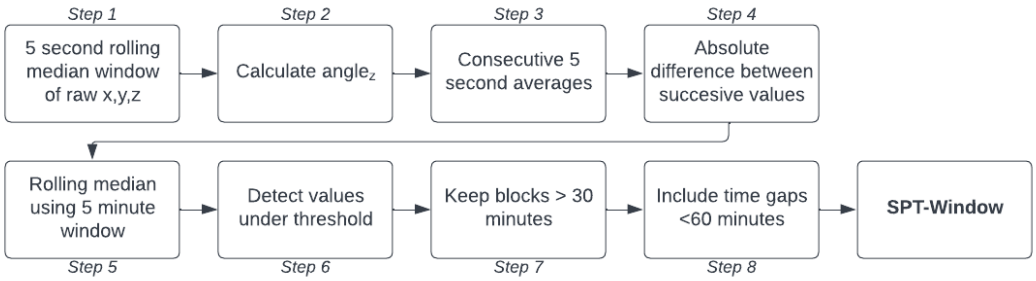


Figure 3.3: Schematic representation of the algorithm based on calculating angle$_z$

Where a_x, a_y and a_z are the median values of a five-second rolling window of the three orthogonally positioned raw acceleration sensors in gravitational units. The estimated arm angles are then averaged per five-second epoch. Subsequently, the absolute difference between successive values are calculated, which is smoothened using again a rolling five-minute window. A threshold is set to distinguish periods of time with many posture changes from those with few posture changes.

Observation blocks are identified when the output is below this threshold and are kept if they last longer than 30 minutes. If the gap between consecutive observation blocks is less than 60 minutes, the blocks are merged. This 60-minute time period is chosen because a second sleep onset within 60 minutes can be assumed as being the same nocturnal bed time. Sleep separated by awake periods greater than 60 minutes should be treated as two distinct sleep episodes. The steps of the angle$_z$ method are visualised in figure 3.3.

Threshold optimization will be done by exploring different thresholds for the angle$_z$ values mentioned in step 6 of Figure 3.3. These thresholds will be compared using the performance metrics that will be discussed in section 3.4.1.

Actigraph Counts

Actigraph is a company that offers wearable devices using accelerometry for monitoring physical activity and sleep. These devices are commonly used in the context of sleep disorders and are proven useful for measuring sleep periods. Their method is based on *activity counts*. The counts unit is a measure that quantifies acceleration within a specific epoch. This reliance on epoch-based counts was a necessity in earlier days due to the limitations of on-board storage and battery capacity. By quantifying acceleration within epochs, the devices could store and process data more efficiently. The algorithms used to transform raw accelerometer data into counts vary across devices and companies, of which many have been held proprietary. However, recently, the counts algorithm of Actigraph has been made available for the research community [15] in the Python library called *agcounts*. Actigraph uses this count-metric for sleep classification.

Two different approaches are deployed for their sleep-wake classification, called Cole-Kripke and Sadeh, which are also extensively used in other sleep classification research [11]. The Cole-Kripke algorithm is primarily used to score adult populations, while the Sadeh algorithm is mostly used for younger adolescents. These algorithms help determine whether a given epoch is considered asleep based on the activity counts. Both algorithms employ a weighted summation method for each window. In the Cole-Kripke algorithm, weights are derived from a linear regression model, whereas in the Sadeh algorithm, weights are based on statistical measures derived from the distribution of activity counts, such as the *mean* (μ) or *standard deviation* (σ).

Because the participants in this study are solely young adults or adults, the Cole-Kripke algorithm will be used in context of this thesis. The Cole-Kripke algorithm is displayed in equation 3.2, where A represents the *activity counts* recorded during a specific epoch *T*.

$$Cole-Kripke = 0.001 \cdot (106 \cdot A_{T-4} + 54 \cdot A_{T-3} + 58 \cdot A_{T-2} + 76 \cdot A_{T-1} + 74 \cdot A_T + 67 \cdot A_{T+1} + 67 \cdot A_{T+2} \quad (3.2)$$

Through this weighted summation method, the algorithm deferentially weights the activity counts from distinct epochs. A sum result below the threshold of 1 is commonly used in sleep research, and indicates a sleep epoch. This thesis will explore various threshold values to find the optimum for

71	**Total Features**
3	**Demographic Features**
1	Gender
1	Age
1	BMI
46	**Time-domain features of IBI values**
4	Means and median of HR and IBIs (both detrended and absolute)
6	SDNN, pNN50, and SDSD (both detrended and absolute)
2	RR range, MAD of IBIs
28	Percentiles (5%, 10%, 25%, 50%, 75%, 90%, and 95%) of detrended and absolute HR/IBIs
2	DFA and PDFA over an 11-minute window
7	**Frequency domain features of PPI values**
5	VLF, LF, and HF power, total power and LF-to-HF ratio
2	Normalized LF and HF
18	**Accelerometer features**
2	Cole-Kripke value, angle_z value
16	Mean, median SD, of x, y, z and magnitude values
1	**Other**
1	epoch nr

Table 3.2: Table with all the input features measured over a 4,5-minute window unless specified otherwise. Features are based on demographic characteristics, IBI-values or accelerometer data.

the potential use for SPT window extraction. The thresholds will be compared using the performance metrics discussed in section 3.4.1.

3.3.2. Four-Stage Classification Model

The ability to accurately identify and monitor each of the sleep stages is crucial for understanding sleep patterns and monitor one's sleep architecture.

This subsection begins with the explanation of the features from accelerometry and IBIs. This step is essential for capturing the physiological signals that differentiate between the stages of sleep. Subsequently, the methodology for feature selection is discussed, along with hyperparameter tuning. Lastly, the methodology for comparing various classifiers is presented.

3.3.2.1. Feature Extraction

The algorithm employs a comprehensive set of features categorized into four primary domains: demographic features, time-domain features, frequency-domain features, and accelerometer-based features.

Demographic features, including age, gender, BMI, are recorded due to their influence on sleep architecture and potential sleep disorders.

Both time-domain and frequency-domain features are derived from IBIs, which themselves are calculated from the raw PPG signal. This signal processing step is crucial for capturing HRV, which is indicative of various sleep stages and overall cardiovascular health. To calculate time-domain and frequency-domain features, a Python toolbox is used called hrv-analysis 1.0.4.

Accelerometer features are extracted directly from the raw signals captured by the accelerometer, providing insights into the physical movements and orientation of the body during sleep.

A detailed enumeration of the extracted features across these categories is presented in Table 3.2.

1. Demographic Features

The demographic features encapsulate the demographic characteristics displayed in Table 3.1. Age and BMI are identified as key factors in the study of sleep. It is observed that infants and young children spend a greater proportion of their sleep in the REM phase, which tends to decrease with age. Conversely, an increase in the amount of light sleep and a decrease in deep sleep are noted among older adults. Furthermore, a correlation between BMI and the occurrence of certain sleep disorders, such as sleep apnea, underscores the importance of monitoring BMI. Therefore, collecting these demographic features are essential.

2. Accelerometer Features

The generated features represent aspects of the change in the time series, for instance the moving average and the moving standard deviation, and the magnitude. Mean, median and standard deviation

values of the x, y, z and magnitude of the accelerometer signal have been calculated over a 4.5-minute window. Also the values of the previously mentioned $angle_z$ and *Actigraph Counts* algorithm have been included as features.

3. Time-domain Features based on IBIs

The fundamental and primary features extracted from the IBIs pertain to the time domain and are calculated over a moving 4.5-minute window, which is common to capture HRV features [16]. These encompass both central tendency and variability measures of IBIs and heart rate (HR). All time-domain features are calculated using the hrv-analysis package, with the exception of DFA and PDFA which are calculated using the antropy 0.1.6 package. The features will be explained below:

- **Mean and Median IBIs:** These statistical measures represent the central tendencies of heartbeat intervals, illustrating the average and middle values of the time intervals between consecutive heartbeats.
- **Mean and Median HR:** While heart rate (HR) is inversely related to IBIs via the formula HR = 60000/IBI, both are analyzed as distinct features. This distinction is important because they function on different scales, which can influence their statistical properties and applicability.
- **Percentiles of Detrended and Absolute HR and IBI Values:** Employed to delineate the distribution of heart rate and IBI values within the dataset, these percentiles help in understanding the spread and outliers in the data.
- **Standard Deviation of IBIs (SDNN):** The standard deviation of the time interval between successive normal heart beats. It elucidates how much the intervals vary from the average interval of the window.
- **Standard Deviation of Succesive Difference (SDSD):** calculates the standard deviation but than over the succesive differences.
- **Mean Absolute Difference (MAD), IBI Range:** The MAD computes the mean of the absolute differences between each pair of successive IBIs. The IBI Range describes the total width between the shortest and longest intervals, encapsulating the extreme points of heart rate variability within the recording.
- **Proportion of IBI Differences Exceeding 50ms (pNN50):** This feature calculates the ratio of the number of times IBIs differ by more than 50 milliseconds to the total count of IBIs, offering insight into the frequency of significant short-term variations in heart rate.
- **Detrended Fluctuation Analysis (DFA):** DFA quantifies the long-term correlations in IBIs by cumulatively summing the intervals, segmenting them, and measuring the deviation of each segment from its linear trend using the sum of squares of residuals. This process is repeated with segments of varying lengths to assess the heart rate dynamics over different scales. DFA is a well-established method for determining the scaling behavior of noisy data in the presence of trends without knowing their origin and shape
- **Petrosian Fractal Dimension Analysis (PDFA):** PDFA evaluates the complexity or irregularity of a sign. It computes the Petrosian Fractal Dimension, which is derived from the number of sign changes in the first derivative of the RR-interval series. This feature provides a quantitative measure of the signal's fractal properties, with higher values indicating increased irregularity or complexity in the RR-interval pattern [17].

4. Frequency-domain Features based on RR-intervals

Spectral analysis of IBIs is predicated on the determination of the power spectral density (PSD) of the signal. The PSD encapsulates the power distribution of a signal across various frequency bands and can be computed utilizing methods such as the Fast Fourier Transform or wavelet transforms. These techniques facilitate the decomposition of the IBIs into its constituent frequency components, essential for HRV analysis. However, IBIs are not evenly spaced since the intervals are measured

at non-regular times. Therefore it is crucial to first interpolate these IBIs. This is all done within the hrv-analysis package.

Within the context of HRV, three principal spectral components are distinguished in the cardiac signal [18]:

- **Very Low Frequency (VLF):** Represents the lowest frequency band in HRV analysis, typically ranging from 0.003 to 0.04 Hz. While its physiological underpinnings are not fully understood, VLF power is often associated with long-term regulatory mechanisms, including hormonal regulation, thermoregulation, and circadian rhythms.

- **Low Frequency (LF):** Represents the sympathetic nervous system activity, with contributions from the parasympathetic nervous system. In HRV analysis, LF power typically ranges from 0.04 to 0.15 Hz. In normalized units, LF reflects the balance between sympathetic and parasympathetic influences.

- **High Frequency (HF):** Primarily reflects parasympathetic (vagal) activity and respiratory sinus arrhythmia. HF power, typically ranging from 0.15 to 0.4 Hz, is strongly influenced by respiratory patterns, with heart rate variability increasing during inhalation and decreasing during exhalation.

The LF and HF components, especially when analyzed in their normalized forms (LFnu and HFnu), elucidate the dynamics between the sympathetic and parasympathetic branches of the autonomic nervous system. The relative power distribution across these frequencies and the LF/HF ratio are metrics for interpreting the autonomic state. Specifically, this ratio can provide insights into the autonomic nervous system's modulation during different sleep stages, aiding in the distinction between wakefulness and deep sleep states. The parasympathetic nervous system predominates in resting conditions, while the sympathetic nervous system drives the "fight or flight" response in stressful situations. Therefore the parasympathetic nervous system, representing HF, is more active during NREM sleep, where the sympathethic nervous system, represented by LF is more active during wake and REM [19]. The normalised LF and HF can be derived from each other using the formulae $LFnu = 1 - HFnu$.

Total power (TP), encompassing the sum of power in the VLF, LF, and HF bands, represents the aggregate variability within the heart rate signal. This measure is indicative of the overall heart rate variability, offering a holistic view of the heart's rhythmical fluctuations over the spectrum of analyzed frequencies.

3.3.2.2. Feature Selection

Features in a dataset play a crucial role in predictive modeling as they encapsulate the relationships between physiological signs and sleep stages, which are critical for making accurate predictions. However, not all features contribute positively to this task. Instead, they may add unnecessary complexity, are very similar or introduce noise that jeopardizes the prediction. To create accurate predictions, it is important to identify the most relevant and informative features from the original feature sets.

For this study, the Correlation-based Feature Selection (CFS) method will be employed. CFS is an effective method for identifying features that are highly correlated to the target, but eliminates features that exhibit high correlation amongst themselves. A potential limitation of this method is that it might not always effectively identify features that are individually weak, but collectively strong predictors.

The procedure starts by assessing the correlation scores between each feature and the target. Features demonstrating a correlation score of 0.1 or lower are deemed insufficiently related and are subsequently removed. This filtered group of features is then designated as the initial subset, referred to as CFS1.

Subsequently, inter-correlation scores among remaining features are examined to identify and remove redundancy, thereby enhancing computational efficiency. For each feature, other features displaying an inter-correlation of 0.8 or higher are identified and compiled into a list. This procedure is replicated for each feature. Subsequently, a representative is selected from each group of highly correlated features, effectively reducing the total number of features. This process forms the second subset, referred to as CFS2.

In the analysis, the model performance of all features, along with the model performance for both subsets will be analyzed.

3.3.2.3. Hyperparameter Tuning

Hyperparameter tuning is important because the performance of machine learning models can vary significantly depending on the values of certain parameters. By tuning the hyperparameters, the best combination of values can be found that maximizes the model's performance or generalization ability. Furthermore, with the relatively small size of the database, a four-fold cross-validation method is used. Importantly, the division was executed at the participant level, guaranteeing that no participant data appears simultaneously in both training and testing phases. For hyperparameter tuning, the first 19 participants were used. The hyperparameters to be tuned per model are discussed below:

- **Support Vector Machine:** Hyperparameter tuning for two critical parameters is done: the kernel function and the regularization parameter (C-value). The kernel function is responsible for transforming the input data into a higher-dimensional space, enabling the model to capture more complex relationships between the data points. The kernel functions *'linear', 'radial basis function'* and *'sigmoid'* were evaluated. The regularization parameter C plays a vital role in the trade-off between the model complexity and the deree to which it can generalize to unseen data. A high C-value focuses on classifying all training examples correctly, while a lower C-value encourages a margin between the classes to enhance generalization. The values of [1, 10 and 100] are evaluated. For the kernels *'radial basis function'* and *'sigmoid'* also the gamma values of [0.001, 0.0001] have been considered. It should be noted that prior to inputting the dataset into the SVM model, the data is normalized using *StandardScaler* from the Scikit-Learn library in python.

- **Random Forest:** The hyperparameters to be tuned for the Random Forest model include the number of estimators, which determines the number of decision trees in the ensemble. Generally, having more trees reduces the risk of overfitting, but there is a limit where further improvements become insignificant. The max_depth parameter controls the maximum depth of each decision tree, while $min\_samples\_split$ sets the minimum number of samples a node must contain in order to consider splitting. These hyperparameters play a crucial role in the performance and generalization ability of the Random Forest model. The values [100, 300, 500] will be evaluated for n_estimators, the values [2, 5, 10] will be evaluated for $min\_samples\_split$ and the values [10, 20] will be evaluated for max_depth .

- **CatBoost:** In the CatBoost model, the hyperparameters to tune include the *max depth*, *number of iterations*, and *learning rate*. The depth parameter controls the maximum depth of each decision tree in the ensemble, influencing the complexity of the model. The number of iterations determines the number of boosting rounds, essentially the number of decision trees in the ensemble, which impacts model performance and computational cost. The learning rate controls the step size during optimization, affecting the speed and quality of learning. The values [3, 6] will be evaluated for *max depth*, the values [300, 200, 100] will be evaluated for *number of iterations*, and the values [0.01, 0.1, 1] will be evaluated for *learning rate*.

The selection of the optimal set of hyperparameters for each model will be based on the performance metrics, that will be explained in the following section. Model comparisons will be conducted by tallying how frequently a model achieves either the highest or second-highest performance score. Based on these results, the model consistently demonstrating superior performance across various metrics will be chosen. Each performance metric will be treated with equal importance during this evaluation process.

3.4. Evaluation of Methodology

3.4.1. Data Analysis of Automatic SPT window Extraction

To evaluate the performance of the SPT window calculation, the mean absolute error (MAE), mean directional accuracy (MDA) and a version of the overlap coefficient (OC) will be calculated. Additionally, a Bland-Altman plot will be made to visualize the differences between PSG and the predicted SPT window.

The MAE is a metric of prediction accuracy between paired observations expressing the same phenomenon. In this case it represents the average absolute difference between the predicted values and the actual values of the PSG. The MAE will be used to assess both start and end times of the

SPT window. MAE is a straightforward measure that tells you the average size of the errors in a set of predictions, however it does not considering their direction. The formula for MAE is detailed in equation 3.3.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\text{Predicted Time}_i - \text{Actual Time}_i| \tag{3.3}$$

The MDA assesses the accuracy of predictions by evaluating how frequently the predicted direction aligns with the actual direction of change. Ranging between 0 and 1, a score of 1 signifies flawless alignment between predicted and actual directions. Specifically, in this context the formula is simplified and quantifies the frequency with which predicted errors are either 0 or precede the start time or are 0 and succeed the endpoint times. The formula can be found in 3.4, where N represents the total number of predictions, PSG-start and Predicted-Start represents the predicted start time according to either the PSG or prediction method, and $\mathbf{1}(\cdot)$ is the indicator function, which returns 1 if the condition inside the parentheses is true, and 0 otherwise. For MDA_{end} it should match $\textit{PSG-Start}_i - \textit{Predicted-Start}_i \geq 0$.

$$MDA_{\text{start}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\,(\text{PSG-Start}_i - \text{Predicted-Start}_i \leq 0) \tag{3.4}$$

The OC is calculated to measure the similarity between the predicted SPT window and the actual sleep time recorded by the PSG. It quantifies the extent to which the two time periods overlap by comparing the size of their intersection to the size of the smaller sets of the two. However, in this variation the denominator is set to the size of the PSG, as this is the standard that we are trying to meet. The equation can be found in 3.5.

$$OC = \frac{|SPTwindow \cap PSG|}{|PSG|} \tag{3.5}$$

These metrics will be used to fine-tune the thresholds in both methods. MAE evaluates the magnitude of error, MDA assesses the directional accuracy per prediction, and the OC metric gauges the extent to which the error and direction influence overall agreement. Ideally, achieving an MDA of one across all methods is desirable. However, if the MDA is less than one but the error is small and the OC is close to one hundred, it indicates close proximity between prediction and actual value.

Following threshold adjustments, performance metrics will be recalculated for the validation set. Additionally, a Bland-Altman plot will be made to visualise the two measurement techniques by plotting the differences against averages of the two techniques. It shows the agreement between two quantitative measurements by constructing limits of agreement. In the Bland Altman plots the direction of the error is visualised.

3.4.2. Data Analysis of Four-Stage Classification Model

Evaluating the performance of the four-stage classification models necessitates a multi-metric approach, given the complexity of model assessment. The evaluation process begins with the construction of a confusion matrix (Fig. 3.4). This matrix facilitates an epoch-by-epoch comparison between the model's predictions and the sleep stages predicted by the PSG recording.

The first performance metrics to be calculated are accuracy and Cohen's κ. Accuracy quantifies the proportion of 30-second epochs correctly classified by the model (Equation 3.6), while Cohen's κ (Equation 3.7 evaluates the agreement level beyond chance. Where accuracy represents a percentage, κ is usually interpreted as follows: below 0.20 'slight agreement', between 0.20 and 0.40 "fair agreement", between 0.40 and 0.60 "moderate agreement", between 0.60 and 0.80 "substantial agreement", and above 0.80 "almost perfect agreement" [20].

Furthermore, precision (Equation 3.8) and recall (sensitivity) (Equation 3.9) are calculated for each sleep stage seperately.

The abbreviations used in these equations correspond to those defined in Fig. 3.4.

$$Accuracy = \frac{T1 + T2 + T3 + T4}{Total\ amount\ of\ samples} \tag{3.6}$$

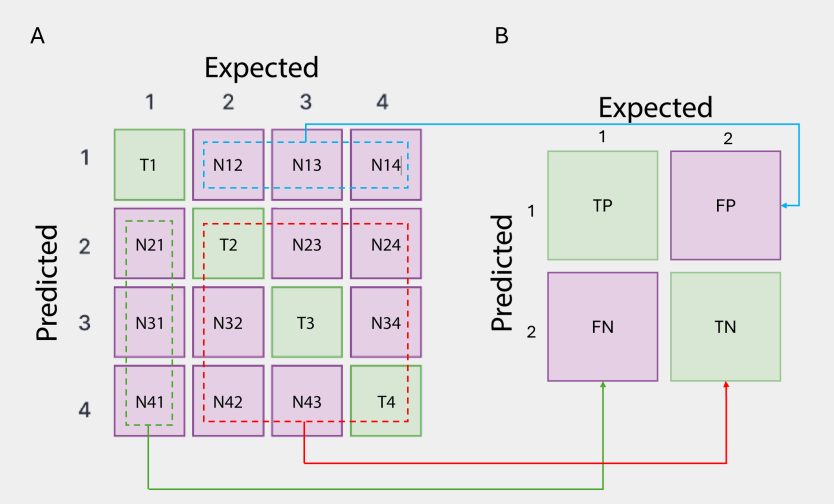


Figure 3.4: Confusion Matrices and how to convert from a multi-class confusion matrix (A) to a binary class confusion matrix (B) of which the performance metrics are calculated.

$$Cohen's\ \kappa = \frac{observed\ accuracy - expected\ accuracy}{1 - expected\ accuracy} \tag{3.7}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \tag{3.9}$$

3.4.3. Evaluation Against Commercially Available Algorithms

The four-stage classification model's effectiveness will be benchmarked against two leading commercial sleep stage classification algorithms: the Philips algorithm, integrated within the Corsano CardioWatch, and the Night-train algorithm, available through Corsano Health's cloud service. Both algorithms leverage accelerometer and PPG data for sleep analysis and will be subjected to the same set of performance metrics in comparison with PSG data as mentioned in the previous section. This comparative analysis aims to contextualize the model's performance within the landscape of existing solutions, providing a clear perspective on its relative accuracy, reliability, and potential advantages or limitations.

3.4.4. Exploring the Influence of Features on Classification

To comprehensively understand how specific features impact the classification of various sleep stages, an analysis of feature importance will be conducted. This analysis focuses on using a one-vs-rest classification approach for each sleep stage seperately. By isolating each sleep stage, this analysis aims to elucidate the features crucial for distinguishing that particular stage from the rest. For each machine learning model a different approach is used:

- **SVM:** In the context of SVM, permutation importance is employed. Permutation importance involves randomly shuffling the values within each feature in the dataset and observing the impact on model performance, thereby identifying features with the most significant influence on predictive accuracy. This model-agnostic approach quantifies the importance of each feature by measuring the decrease in model accuracy after permutation.
- **RF:** For RF, the built-in feature importance mechanism is utilized, assessing the impact of each feature based on its contribution to the model's predictive accuracy and the reduction of variance.
- **CB:** CB also offers proprietary feature importance scores, quantifying the contribution of each feature to the model's performance.

Integration of these feature importance metrics across different models will result in a table, presenting the influential features per stage in sleep stage classification.

4

Results

This section delineates the results, structured into four main subsections. Initially, in section 4.1 threshold tuning will be done for the automated SPT-window extraction, followed by a comparison of the two applied methods. Secondly, the performance of the four-stage classification model is examined in section 4.2, offering insights into its capabilities and limitations in distinguishing between different sleep stages. In section 4.3 the comparative analysis against commercial algorithms makes it possible to compare the algorithm against existing solutions for sleep stage classification. Lastly, in section 4.4 the influence of features on the classification are explored and sheds light on the significant predictors, underpinning the classifier's decision-making process.

4.1. Automated SPT window Extraction

Two methods found in literature have been compared for their effectiveness for SPT window extraction based on the outcome measures MAE, MDA and OC against the PSG data. Optimization involved tuning the threshold of either the angle$_z$ value or the Actigraph counts value to classify periods as sleep (below threshold) or awake (above threshold). These thresholds were then used to compare the performance of both methods using the validation set.

4.1.1. Threshold Tuning

Figure 4.1 shows the performance metrics per threshold value, where Angle$_z$ is visualised in subgraph 4.1a and the Activity Counts in 4.1b. The Figures highlight how varying the threshold affects the MAE, MDA and OC of start and end times.

In both methods, the mean OC (OCm) starts already quite high, and plateaus beyond a specific value. Simultaneously, the standard deviation of the OC (OCstd) decreases as the threshold value increases, indicating more precision in SPT window extraction. Once the OCm stabilizes, further increasing the threshold does not substantially enhance overlap, making MAE a critical factor for minimizing unnecessary data processing. The exact values can be found in Appendix A.

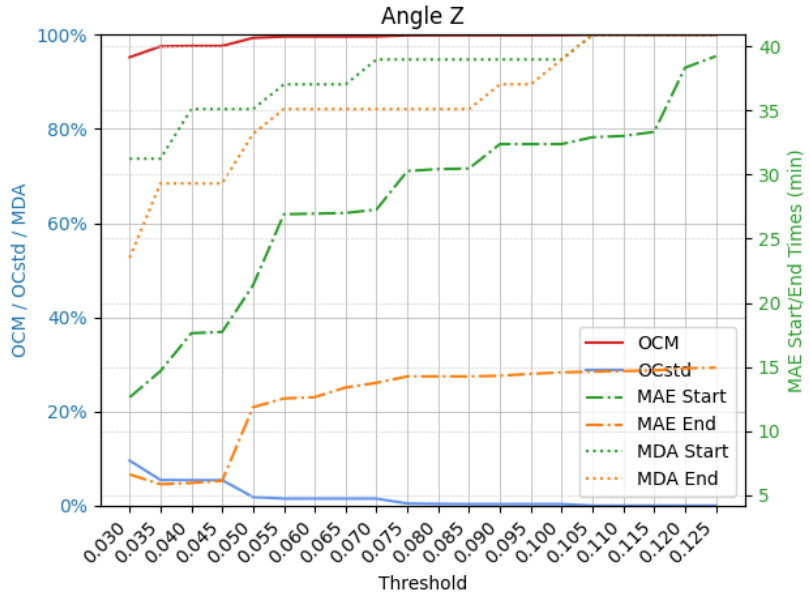
Angle$_z$

For the angle$_z$ method, OCm±OCstd values range from 95.22% ±9.57% to 100% ±0.0%. The OCm plateaus at a threshold of 0.050, after which it slowly inclines to its peak value of 100% ±0.0% around a threshold of 0.105. The marginal OCm difference between threshold 0.050 and 0.105 is 0.67% with a difference of 1.81% in OCstd.

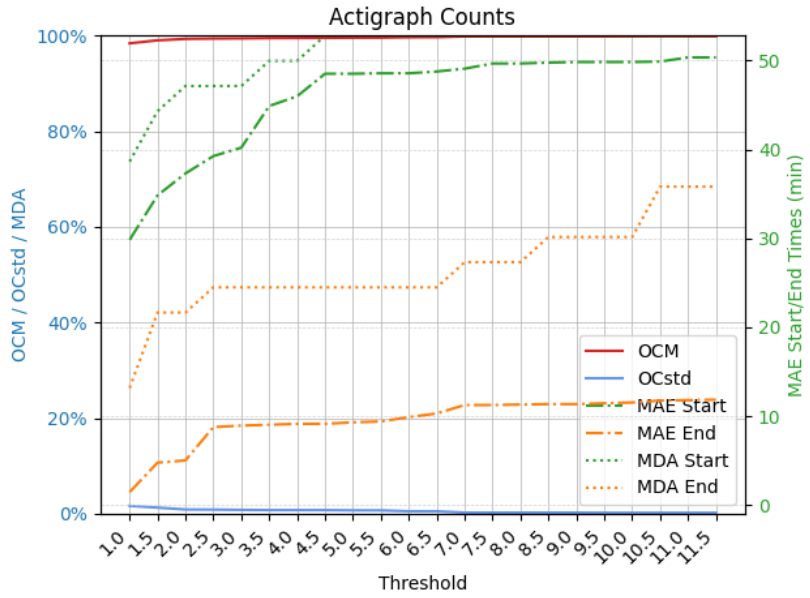
Examining at the MAE, denoted by the 'dash-dot'line, both lines plateau around a threshold of 0.055, where the rate of change for MAE start increases more compared to the MAE of the end time. The maximum difference in MAE for start and end times between thresholds 0.055 and 0.105—00:06:00 and 00:02:06, respectively— indicates that the rate of change in MAE decreases after reaching a threshold of 0.055. This suggests that raising the threshold beyond this point doesn't affect MAE to the same extent for both start and end times.

Lastly, examining the MDA, denoted by the dotted line, the rate of change for both is highest before the threshold of 0.055, and both reach a score of 100% at threshold 0.105.

Considering all of this, the threshold of 0.105 is preferred, as this gives an OCm score and a MDA score for both start and end times of 100%, indicating that the entire sleep window is extracted. The corresponding errors for this are 00:32:54 for the start time and 00:14:38 for the end time, indicating that it precedes the PSG start time by approximately 32 minutes and succeeds the PSG end time by approximately 15 minutes.



(a) Angle$_z$ method



(b) Actigraph Counts method

Figure 4.1: Graphs display the mean Overlap Coefficient (OCm) and standard deviation of OC (OCstd), along with Mean Absolute Error (MAE) and Mean Directional Accuracy (MDA) for start and end times across various thresholds. Subgraph (a) showcases results obtained using the Angle$_z$ method, while subgraph (b) presents those from the Actigraph Counts method. The x-axis represents threshold values, with the left y-axis denoting OCm and OCstd as percentages, while the right y-axis represents MAE in minutes. For detailed numerical data, refer to Appendix A.

Actigraph Counts

The Actigraph Counts method shows OCm values ranging from 98.40% ±1.68% to 99.88% ±0.92%, where the range is smaller in comparison to the difference in threshold for the anlge$_z$ method. The OCm plateau occurs at a threshold of 2, showing an OCm of 99.33% and an OCstd of 0.92%. Ultimately, the maximum OCm value is reached at a threshold of 10.5, reaching an OCm ±OCstd of 99.88% ±0.23%. The negligible difference in OCm between these thresholds is 1.48%, with a 0.69% variation in OCstd.

Examining the MAE, denoted by the 'dash-dot'line, it is seen that the MAE of start time plateaus around a value of 4.5, wheras the MAE of end time plateaus earlier at a value of 2.5. The difference in MAE for start and end times between this plateau value and the highest value is marginal, 00:1:22 and 00:02:59 respectively.

Lastly, examining the MDA, denoted by the dotted line, the MDA Start (green) reaches a value of 100% already around a threshold of 4.5, whereas the MDA of end time (orange) is only reaching a maximum value of 68,42% around the threshold of 10.5.

Considering all of this together, it becomes evident that with these thresholds, the OCm did not reach 100%, indicating a lack of complete overlap of the PSG window. Examination of the MDA reveals that approximately a third of the predictions precede the PSG end time, with an MAE for end time of approximately 00:10:00. However, this doesn't significantly influence the OCm, as it reaches almost a perfect score with 99.88%. Therefore, threshold 10.5 is chosen, as this shows the highest MDA end value.

4.1.2. SPT window extraction using Validation Set

The Angle$_z$ method with threshold 0.105 and the Actigraph Counts method with threshold 10.5 are used to calculate the OC metrics together with the MAE and MDA for start and end times. The results can be found in Table 4.1.

The OCm and MDA of the start time using the Angle$_z$ method, both being less than 100%, indicate that the Angle$_z$ method struggles to capture the entire SPT window. With an MDA of 80%, it is evident that for one participant, being participant 27, this method failed to extract the complete window. However, the Actigraph Counts method exhibits a significantly higher MAE for the start time, almost four times as high as that of the Angle$_z$ method.

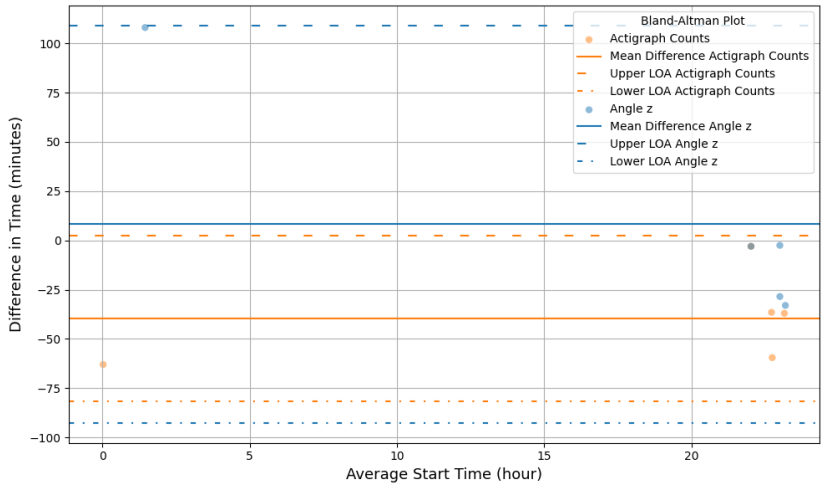
The Bland-Altman plots visualised in Figure 4.2 showcases the analysis of start and end times of the predicted SPT window in comparison to PSG. Bland-Altman plots relates the mean of (x-axis) and difference between (y-axis) the predicted and actual start and end time, with the actual start time being the PSG SPT window. The blue dots present the Angle$_z$ method and the orange dots the Actigraph counts method.

Figure 4.2a presents the start times. The solid lines represents the difference between the two methods, which is positive for the Angle$_z$ and negative for the Actigraph Counts method. A positive difference indicates that the predicted start time exceeds the actual start time, whereas a negative value indicates the start time precedes the actual start time. Thus, a negative value is more favorable as it signifies that the prediction is earlier than the actual start time. The positive value of the Angle$_z$ line is due to one outlier visible in the left corner of the figure. The other blue dots are negative, meaning that the predicted window starts somewhat before the actual SPT window. The orange dots are all negative, but spaced further from the zero line, indicating they start well before the actual SPT window.

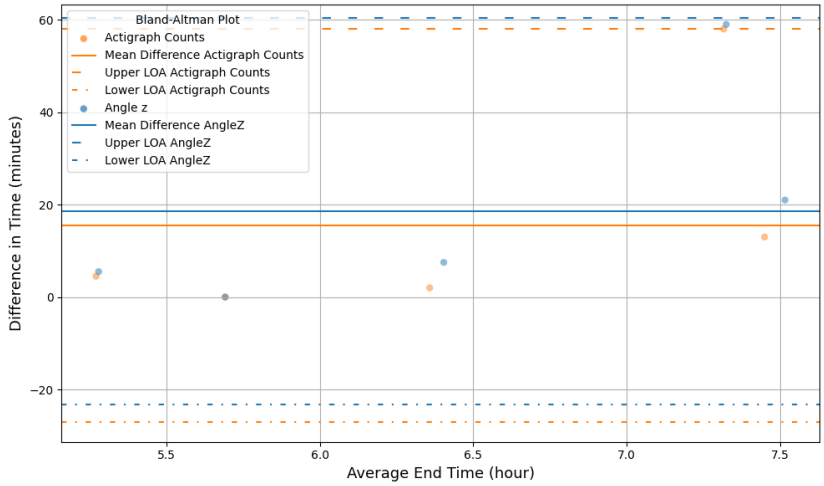
Figure 4.2b presents the end times. Both mean differences are above zero, with the Angle$_z$ method having a slightly higher value than the Actigraph Counts method. The dots appear in pairs, where the yellow dot is often closer to the zero-line, indicating a more precise prediction of the end time.

Table 4.1: Performance metrics of angle$_z$ (threshold = 0.105) and Actigraph Counts method (threshold = 10.5) for SPT window prediction using the validation set. Performance metrics include the Mean Absolute Error (MAE) and Mean Directional Accuracy (MDA) of start and end times, and mean Overlap Coefficient (OCm), and the standard deviation of OC (OCstd).

	Start Time		End Time			
	MAE	MDA	MAE	MDA	OCm	OCstd
Angle Z	0:08:12	80%	0:18:36	100%	94,71%	10,59%
Actigraph Counts	0:39:48	100%	0:15:30	100%	100%	100%



(a) Bland Altman plot for start time.



(b) Bland Altman plot for end time.

Figure 4.2: Bland Altman plots for the start and end times using Angle_z method (threshold 0.105) or Actigraph Counts method (threshold 6.5).

To zoom in on the outlier of the Angle$_z$ method in the start-time Blant-Altman plot, the predicted SPT windows and the PSG sleep-wake classification are shown in Figure 4.3. For the Angle$_z$ method, the sleep segments resulting from step 7 in figure 3.3 are shown. In this figure, significant fluctuations between wake and sleep states are observed throughout the night in the PSG signal. The Actigraph Counts method, depicted by the yellow line, interprets the entire period as a single continuous sleep period. However, the Angle$_z$ method, represented by the blue line, identifies two separate sleep segments due to the prolonged duration between consecutive sleep periods. As the time gap between these sleep segments exceeds 60 minutes, the first segment is considered a distinct sleep window.

It's important to note that the calculation of the SPT window only considers one window per night. Consequently, the Angle$_z$ method misses the first sleep segment, leading to an incomplete extraction of the SPT window.

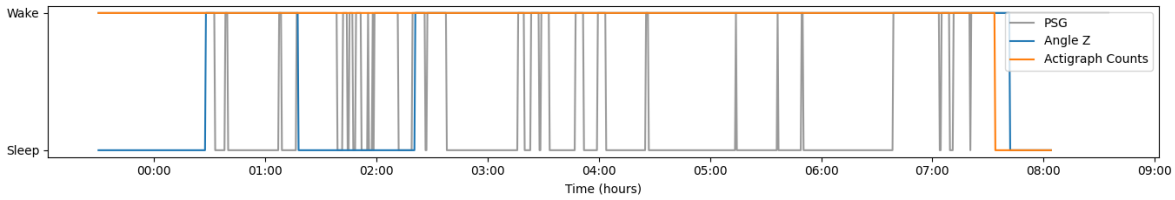


Figure 4.3: Comparison between PSG sleep-wake prediction and sleep period segments of the Angle$_z$ method (after step 7 of the Angle Z calculation 3.3.

4.2. Four-Stage Classifier

This section outlines the performance metrics for four-stage sleep classification. It begins by examining the distribution of sleep stages among participants (Section 4.2.1) and addresses the handling of class imbalance (Section 4.2.2). The focus then shifts the machine learning models by feature selection (Section 4.2.3) and hyperparameter tuning (Section 4.2.4).

4.2.1. Sleep Stage Distribution among Participants

The bar plot in Figure 4.4 illustrates the distribution of epoch counts across sleep stages for all participants. The bar for each participant consists of several colors corresponding to the sleep stages, with the height of each color reflecting the number of epochs recorded for that sleep stage.

Light sleep tends to have the highest count across most participants, which is consistent with the typical structure of a sleep cycle where light sleep constitutes the largest proportion. Deep sleep bars, which are crucial for physical restoration and memory consolidation, are notably shorter than light sleep, reflecting its smaller share of the sleep cycle. However, there is considerable variability among partic-

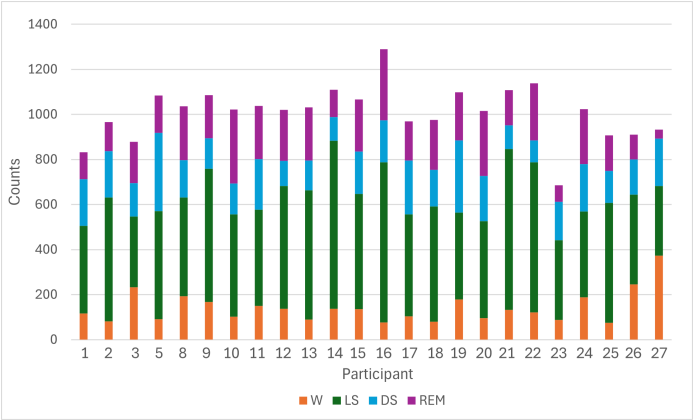


Figure 4.4: A bar plot diagram showcasing the amount of epochs per stage per participant. W = Wake, LS = Light Sleep, DS = Deep Sleep

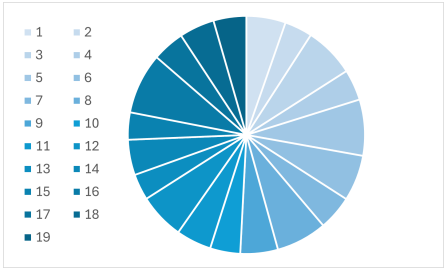


Figure 4.5: A pie chart showing the representation of each participant in the balanced dataset.

ipants in deep sleep counts, possibly indicating differences in sleep quality or age-related factors, as younger individuals tend to have more deep sleep.

The wake stages, represented by the orange bars, denote wakefulness periods within the sleep cycle, anticipated to be minimal in cases of consolidated sleep. Despite this expectation, the height of the orange bars remains noticeable. Particularly noteworthy is participant 27, whose elevated wake count surpasses even that of light sleep, indicating a night of limited rest.

REM sleep counts, essential for cognitive processes such as memory and learning, display variation but generally fall below light sleep counts and above wake stages, aligning with typical proportions of a standard sleep cycle.

4.2.2. Handling Class-Imbalance

To address class imbalance prior to model training, an equalization approach was employed for each participant's data as explained in Section 3.4.2. Shortly, this entailed identifying the sleep stage with the lowest epoch count for each individual and then extracting an equivalent number of epochs randomly from the other stages.

The resulting allocation among the various participants' data is visualized in the pie chart depicted in Figure 4.5. The contribution varied significantly among participants, with participant 16 providing the fewest epochs per stage of 77 epochs, while participant 19 provided the most, at 179 epochs per stage—more than double that of participant 16.

Following this equalization method, the proportions of the original stage data utilized were as follows: 89% of wake, 22% of light sleep, 63% of deep sleep, and 53% of REM Sleep epochs.

4.2.3. Feature Selection

Initially, a correlation matrix was constructed and features demonstrating an absolute correlation of 0.1 or greater with the 'sleep stage' feature were shortlisted. From this analysis, a total of 30 features remained and were visually represented in a heat map to assess inter-feature redundancy. The heat map is visualised in Figure 4.6.

Highly inter-correlated areas, indicated by bright colors on the heatmap, were examined to address

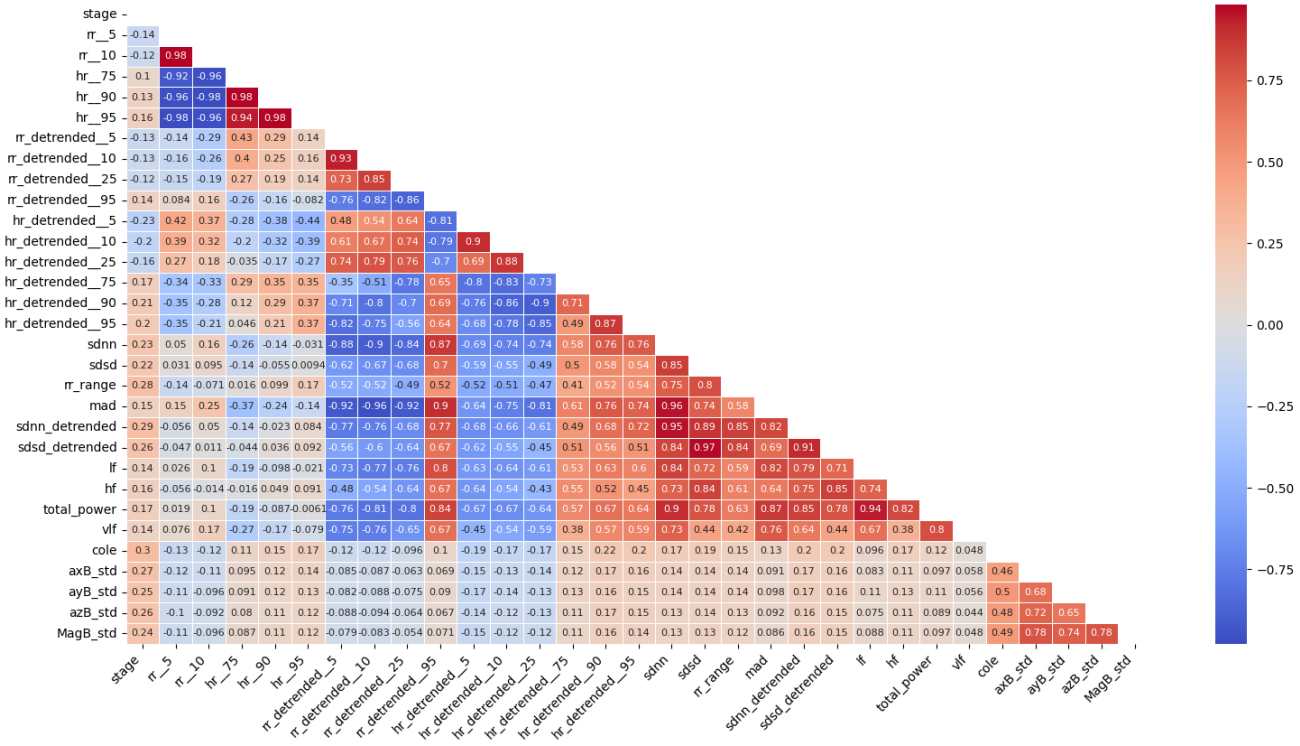


Figure 4.6: Heat map of 30 features that have a correlation of 0.1 or higher to the target label 'stage'.

multicollinearity concerns. The analysis identified several feature subgroups and determined the most suitable features to retain based on their individual correlations to the target label and inter-correlations with other features.

1. **['total_power', 'vlf']**: 'vlf' was selected despite 'total_power' having a marginally higher correlation with the target label. This decision was made due to 'total_power's significant inter-correlation with other features.

2. **['rr_detrended__95', 'hr_detrended__10', 'hr_detrended__5']**: 'hr_detrended__5' was chosen for its lower inter-correlation and stronger correlation with the target label.

3. **['hr_detrended__10', 'hr_detrended__75']**: 'hr_detrended__75' got chosen as 'hr_detrended__10' was closely related the feature selected in the second group.

4. **['rr__10', 'hr__75', 'hr__90', 'hr__95', 'rr__5']**: 'hr__95' was selected for its highest correlation with the target label without inter-correlation concerns.

5. **['rr_detrended__95', 'sdnn', 'mad', 'total_power', 'lf']**: Since 'total_power' and 'rr_detrended__95' were excluded in previous subgroup decisions, the remaining candidates are 'MAD', 'sdnn', and 'lf'. Despite 'lf' displaying the lowest correlation to the target label among its peers, its documented importance in sleep stage analysis justifies its selection.

6. **['sdnn', 'sdsd', 'sdsd_detrended', 'rr_range', 'sdnn_detrended', 'hf']**: 'hf' was chosen due to its theoretical relevance to specific sleep stages. A trade-off between 'rr_range' and 'sdnn_detrended' needed to be made, where 'sdnn_detrended' was selected because it is less susceptible to potential noise and has a higher correlation to the target variable. Notably, while 'hf' and "sdnn_detrended' are part of the same group, they do not exhibit significant inter-correlation with each other, only with other members of this group.

7. **['hr_detrended__10', 'hr_detrended__90', 'hr_detrended__95', 'mad', 'hr_detrended__25']**: 'hr_detrended__90' was selected for its strong correlation with the target label and lack of significant inter-correlation.

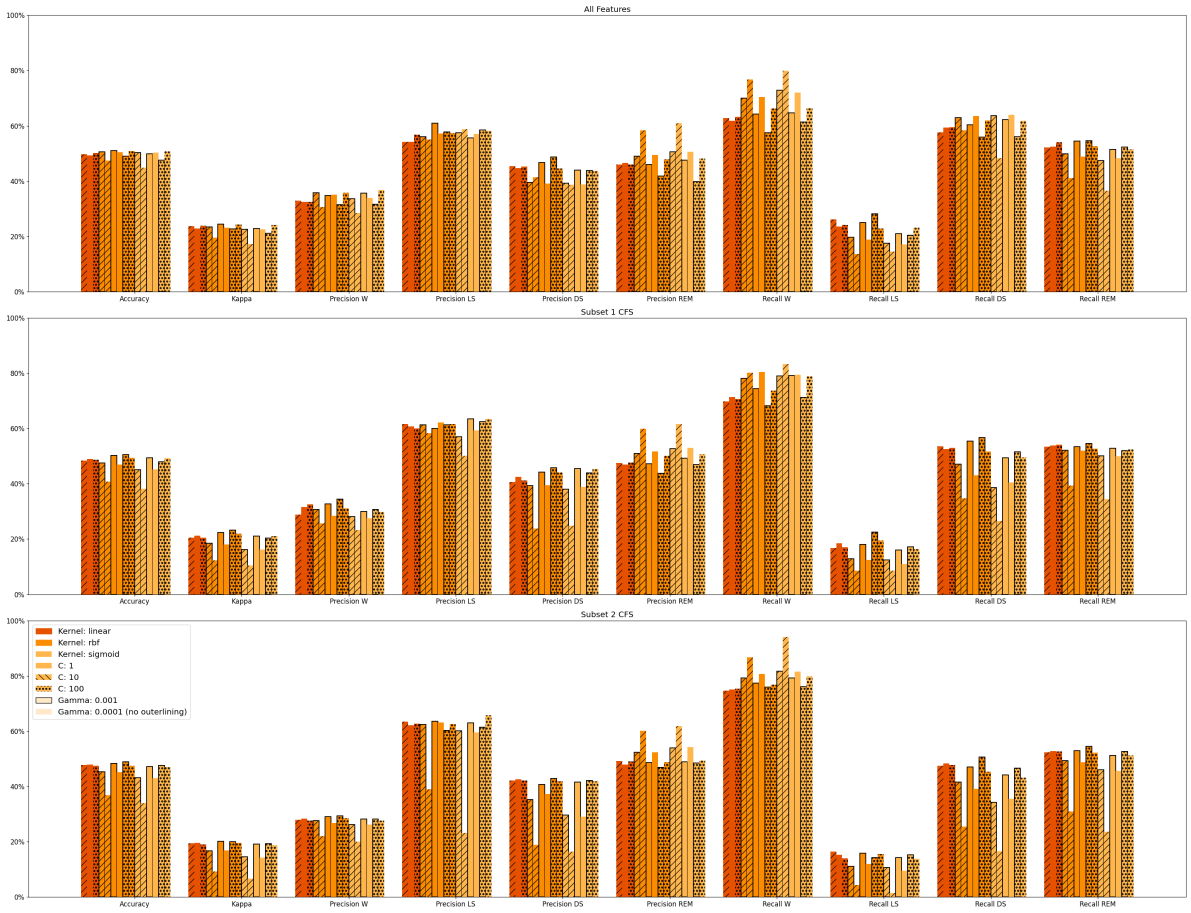
8. **['rr_detrended__10', 'rr_detrended__95', 'sdnn', 'mad', 'total_power', 'rr_detrended__25']** and **['rr_detrended__10', 'hr_detrended__95', 'sdnn', 'mad', 'rr_detrended__5']** in another: these groups are similar. Given the recurring appearance of 'rr_detrended__10', and 'mad', 'total_power'and 'sdnn' already got discarded in previous groups, the decisions were made in favor of 'rr_detrended__5' and 'rr_detrended__25' to maximize the diversity of features without high inter-correlation.

The features that did not show an inter-correlation were the accelerometer features 'cole', 'axB_std','ayB_std', 'azB_std', 'MagB_std'. By combining these with abovementioned selected features, the final feature set was formed, comprising the 15 features: 'hr__95', 'vlf', 'hf', 'lf', 'sdnn_detrended', 'hr_detrended__75', 'hr_detrended__5', 'hr_detrended__90', 'rr_detrended__5', 'rr_detrended__25' 'cole', 'axB_std', 'ayB_std', 'azB_std', and 'MagB_std'.

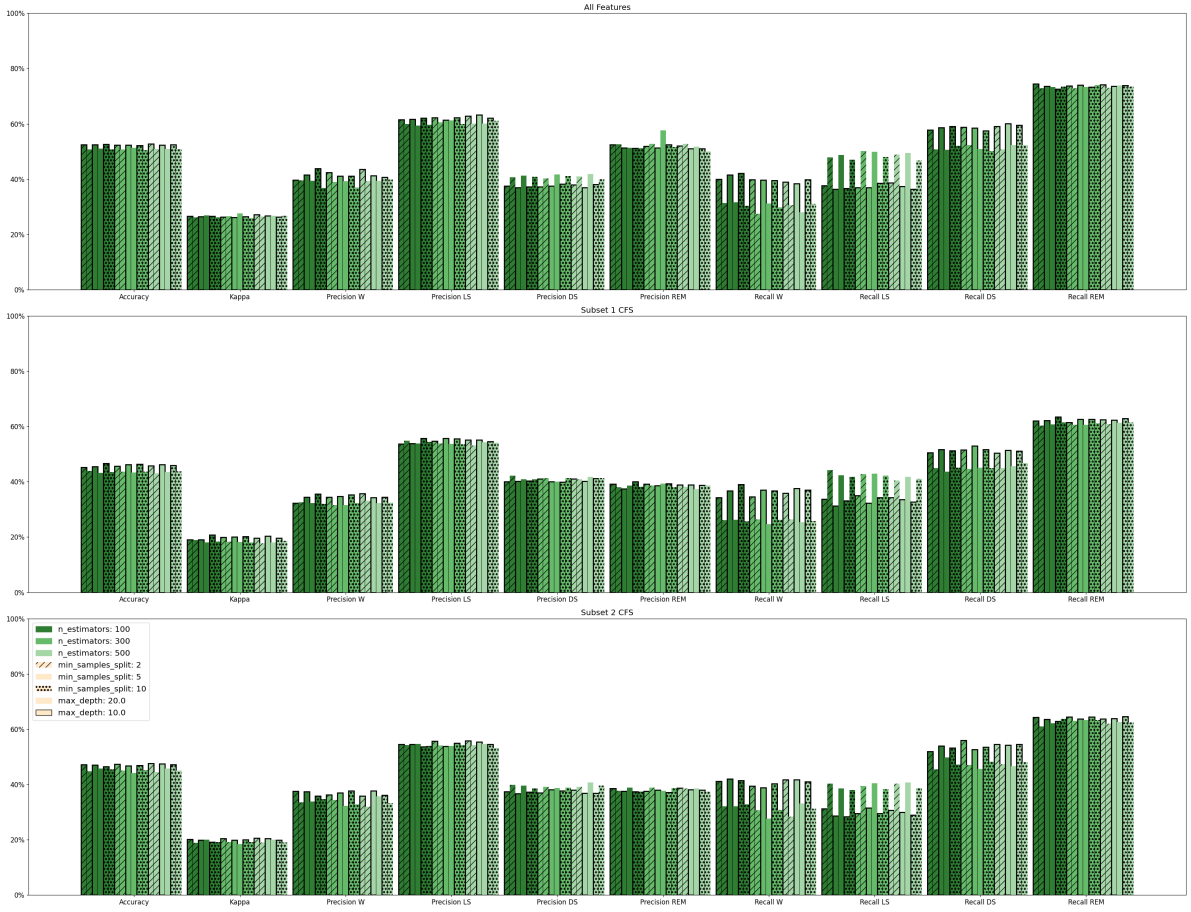
4.2.4. Hyperparameter tuning

This section delves into the hyperparameter tuning process and the resultant performance metrics for the SVM, RF, and CB models employed in this study.

Hyperparameter tuning is done for each feature subset, being all features, the first subset of CFS (30 features) and the second subset of features (15 features). The predictions by the models were smoothened using a smoothing array. This adjustment aimed to reduce fluctuations in the hypnogram over short timeframes, The function is designed to apply a mode-based smoothing technique to a time-series data array, utilizing a sliding window approach to reduce noise and enhance data consistency. By requiring an odd-sized window to ensure a central element, the function iterates through the input array, replacing each element within the defined window with the mode of that window. This method effectively mitigates short-term fluctuations and preserves the integrity of significant trends in the data, unless a clear repetitive pattern is observed.

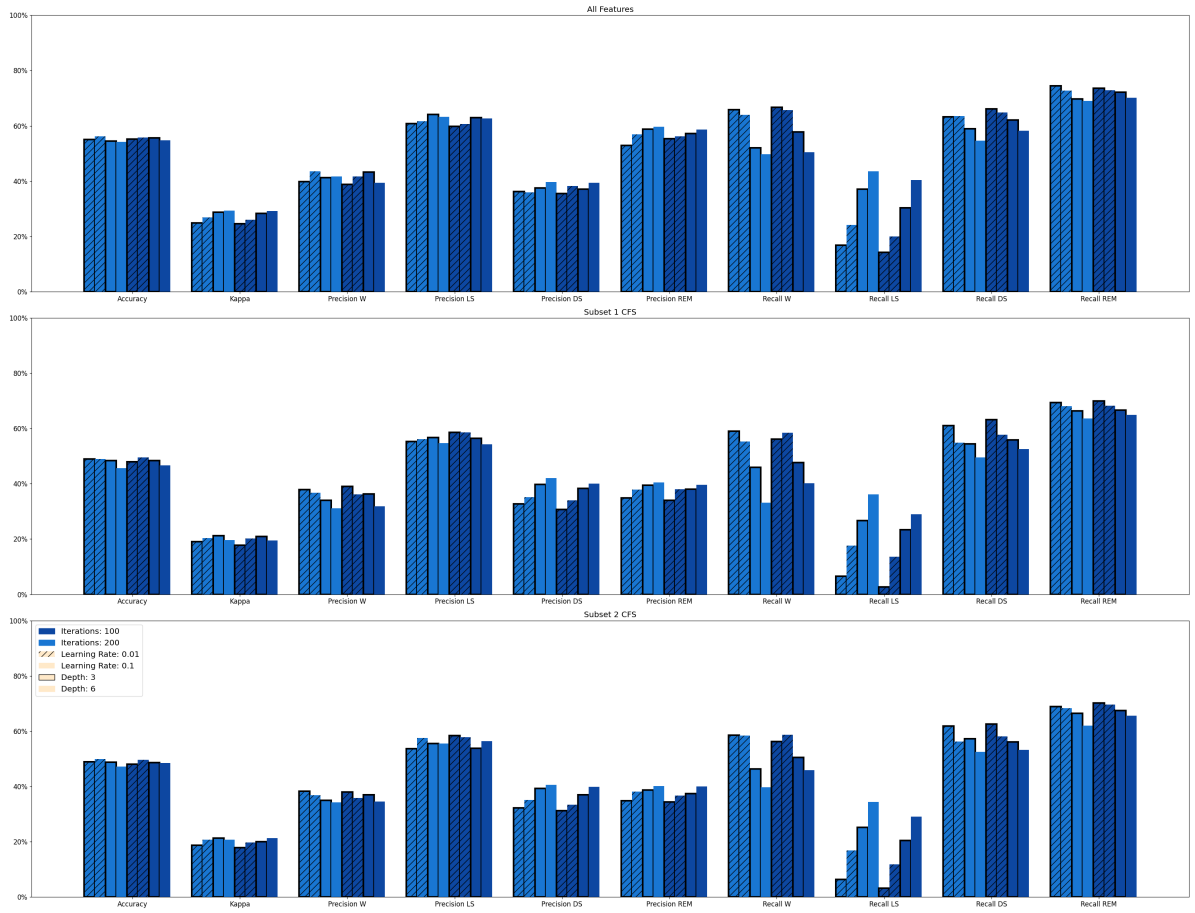


(a) SVM models



(b) Hyperparameter tuning RF models

Figure 4.7: Hyperparameter tuning of SVM and RF models



(c) CatBoost models

Figure 4.7: Hyperparameter tuning of CatBoost models (continued)

Figure 4.7 illustrates the performance metrics of SVM, RF, and CB models under various hyperparameter settings with different feature subsets, as evaluated with a 4-fold-cross validation method. Each subfigure represents the respective model's performance across accuracy, Cohen's kappa, and precision and recall per stage, where SVM is visualised in (a), RF in (b) and CB in (c). The best set of hyperparameters was chosen by tallying how frequently a model achieved either the highest or second-highest performance across the performance metrics. The exact values can be found in Appendix B.

Overall it can be seen that the heights per performance metric are approximately of the same height per feature subset (each graph within a subgraph). However, notable variations in performance metrics are observed among different hyperparameter sets, where the deviation is the highest in the SVM model, especially with the recall scores. Whereas the heights per performance metric per hyperparameter subsets are more subtle in the RF and CB model. In both SVM and CB there is a lower recall score for deep sleep, with a lot of variation per hyperparameter set.

- **SVM:** In Figure 4.7a, the performance metrics across various feature sets are consistently represented. The hyperparameter combination ['rbf', 10, 0.001] yields superior performance when all features are considered, achieving the highest scores in accuracy, kappa, and precision for light sleep, and the second highest in precision for deep sleep and recall for REM sleep. For CFS subset 1 and subset 2, ['rbf', 100, 0.001] emerges as the top performer, with notable results in seven metrics under subset one, albeit the lowest in precision for REM sleep. The visualization utilizes a mid-orange color to denote the 'rbf' kernel, a striped pattern for C=10, a dotted pattern for C=100, and a black outline for a gamma value of 0.001.

- **RF:** As depicted in Figure 4.7b, the RF model shows consistent performance shapes across metrics in each subgraph, particularly excelling in precision for wake and recall for REM sleep. There

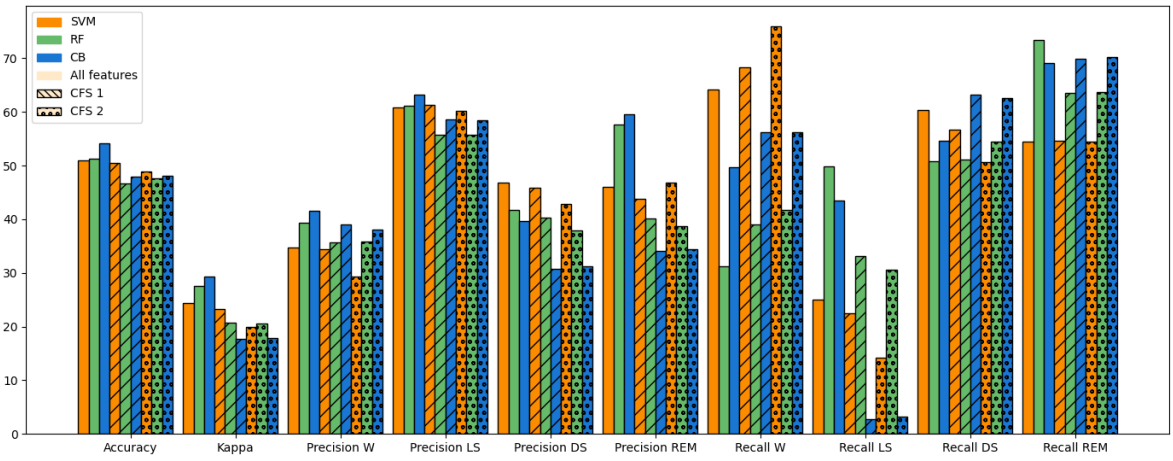


Figure 4.8: The models with the highest performed hyperparameter settings in comparison to each other using the test set

is a marked contrast between the performances of bars outlined in dark (representing a max depth of 10) and those without outline (max depth of 20). Bars with an outline consistently score higher in accuracy and precision across various sleep stages, with the exemption of recall for light sleep where those without score consistently higher. Optimal performance using all features, considering computational efficiency, is achieved with the hyperparameters [300, 5, 20] for n_estimators, min_samples_split and max depth respectively, denoted by a middle green color, absence of pattern, and no outline. For CFS subset 1 the model with values [100, 10, 20], denoted by a dark green color, dotted pattern and no outline, where it scores highest on 5 performance metrics, and for CFS subset 2 the values [500, 2, 10] gets chosen, denoted by the light green color, dotted pattern and black outline.

- **CB:** Figure 4.7c illustrates a uniformity in bar heights across most metrics, irrespective of the feature subset used. A notable exception of the uniformity is the recall for light sleep, where models with a higher max depth and a learning rate of 0.1 achieve the best results but perform lower in recall for wake. The hyperparameters [200, 0.1, 6], represented by light blue color without outline or pattern, are most effective overall, securing the top position in four metrics and second in one. For both the first and second feature subsets, the parameters [100, 0.1, 3] (marked by dark blue color, black outline, and no pattern) consistently deliver the highest scores across most metrics.

Notably, the best-performing models across SVM, RF, and CB algorithms demonstrate similar results, suggesting that, despite different algorithmic approaches and configurations, there is a level of performance parity when the hyperparameters are optimally tuned.

In Figure 4.8 the performance metrics are visualised per feature set per model with the highest scoring hyperparameter set, calculated with the four-fold cross-validation. SVM, RF, and CB, denoted by the different colors, perform similarly across most metrics, with some variations where SVM scores significantly higher as can be seen for recall wake, but also significantly lower for recall REM for example.

Feature selection seems to have a different impact on different models. For some metrics, one feature selection method outperforms the other, which is the case with recall for wake for SVM models: the fewer features the higher it scores. Overall, using all features is mostly the best. The impact of feature selection is not consistent across different metrics. For example, in recall for class deep sleep, CFS 1 outperforms all features for CB but not for RF and CB.

4.2.5. Model Performance using Validation Set

In Figure 4.9 the performance metrics of each subset per model can be seen. The exact numbers can be found in Appendix D. Overall you see the pattern that with a decrease in features, also a decrease in

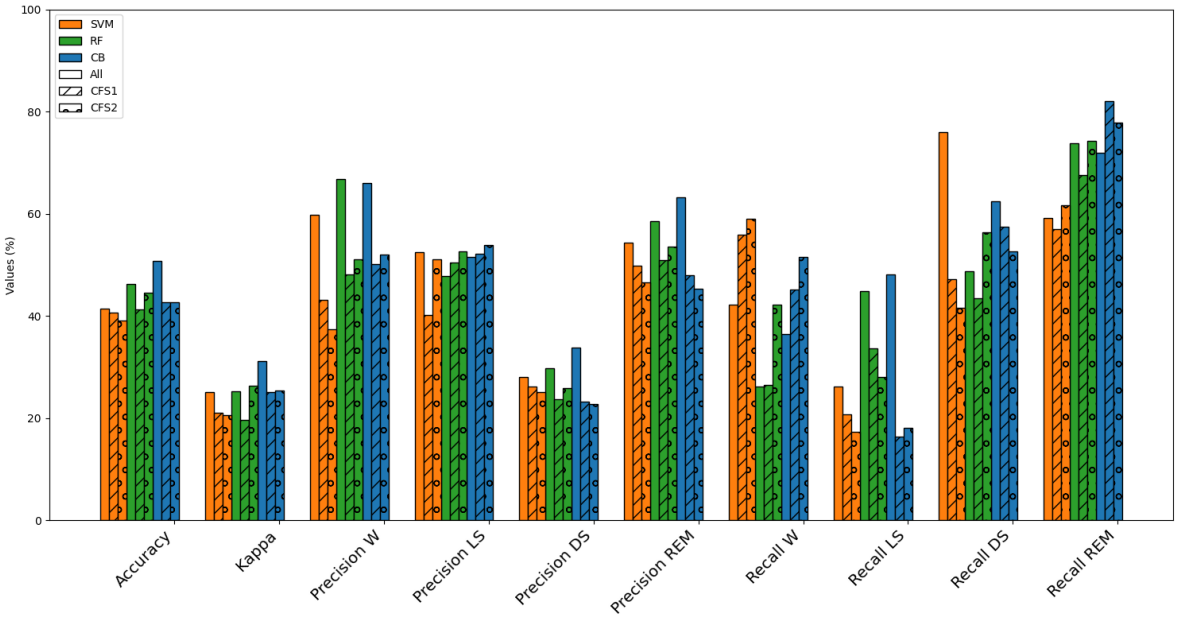


Figure 4.9: The mean performance metrics per model and feature set of the validation set

accuracy and cohen's kappa and precision scores can be seen. For recall is that the other way around, the less features, the recall score is often higher, with the exception of recall for wake.

Overall, the best performing model is the CatBoost model, using all the features. This model shows five times the best score (in Accuracy, Cohen's kappa, Precision for Deep Sleep and REm and Recall for Light Sleep), and two times the second highest score (Precision Wake and Recall Deep Sleep). The exact scores can be found in Appendix C.

4.3. Commercial Algorithms

This section delves into an in-depth comparative evaluation of two leading commercially available sleep stage classification algorithms, Philips and Night Train. The evaluation encompasses an analysis of algorithm performance across 19 participants for the Philips algorithm and 16 for the Night Train algorithm, excluding participants 2, 5, and 12 due to connectivity issues during data acquisition.

Detailed outcomes of these analyses are summarized in the graphs depicted in Figure 4.10. The exact results can be found in Appendix D.

4.3.1. Philips Performance

The Philips algorithm's classification performance showcased substantial inter-individual variability, a pattern that raises questions about the uniformity of the algorithm for a heterogeneous cohort. The Philips model reached a mean accuracy of 61.81% with a Cohen's κ of 0,39 indicating a fair to moderate agreement, reaching same values as mentioned in literature (59.3% with κ of 0.42) [21], with lowest accuracy of 49,82% with κ of 0,23 and highest value of 75.60% with κ of 0,60. The Philips algorithm, represented by the darker color in Graph 4.10 show a different shape per participant, indicating that the algorithms performance on the specific sleep stages differs per patient. The precision and recall scores per sleep stage are discussed below:

- **Wake:** A notable difference in precision was observed among participants, with a peak precision of 100%, contrasted starkly by a complete failure to predict wake periods in others (participants 10, 16, 24). This inconsistency, together with the low recall rates, highlights a significant shortcome in detecting actual wake periods.
- **Light sleep:** Exhibiting the highest mean precision and recall among the stages, with a mean precision of 66.68% and a mean recall of 70.14%, light sleep nonetheless reveals variability, hinting at potential outliers that could skew the performance metrics.

- **Deep sleep and REM sleep:** Both stages show similar results with precision values of 55.50% and 55.30% respectively, and recall values of 58.25% and 59.98% respectively. However, they ranges greatly on participant level.

4.3.2. Night Train

Initially, the Night Train model performed worse then expected, resulting in all performance metrics of around 30%. Upon visual inspectation, it looked like there was a timeshift between the two hypnograms, which made me wonder whether there is a correlation. Therefore, it is explored if there might be a time offset that occurs. Therefore, a correlation-based time shift analysis was conducted to identify the time shift with the highest correlation. A preferable time shift of 34 was discovered, which is used in the performance metrics shown in Table D.1. However, Night Train still provides classification for the entire SPT window of the PSG, ruling out that missing data does not influence the performance metrics. After consulting with the company, it was revealed that this time shift is due to how the API processes data. Further analysis is ongoing but falls outside the scope of this research.

The Night train algorithm reached a mean accuracy of 56.00% and a Cohen's κ of 0,32 representing a fair agreement between predicted and actual values. These values are lower than what they report in their benchmark test (n=21) of 74% with a Cohen's κ of 0.62.

The bars of the Night Train algorithm, represented by the ligher color in Graph 4.10, show a varied shapes per participant. This variability suggests that the algorithm's efficacy in identifiying certain stages differ among participants. The precision and recall scores per sleep stage are discussed below:

- **Wake:** The algorithm's precision in wake detection was markedly low, with a mean precision of 27.48%. The relatively high std value (17%) indicates that there is a big difference among patients (lowest 1.56% and highest 71.43%). The recall scores are relatively higher of 57.50%, but also shows big variation across individuals.
- **Light sleep:** Light Sleep scores overall the highest for both precision and recall with scores of 64.19% and 64.13%. The results do not differ as much per participant in comparison to the wake predictions.
- **Deep sleep:** The prediction of deep sleep is 56.11%. For participant 1, 8, 22 and 27 it shows extreme low values, implying that the prediction is not exact, while for the other patients the values are around 65% or higher. Same holds for recall, which has a mean value of 48.12% where for some it shows particularly good results (88.67%) and for some particularly bad (0.00%).
- **REM:** Reporting the highest precision score of 65.90%. Overall it shows generally higher results, with some exceptions with values around 40%. On the other hand, the recall score of REM is lowest, with a value of 43.32%, highly variating between 12.07% to 77.42%.

This comparative analysis reveals a crucial insight: both commercially available sleep stage classification algorithms, Night Train and Philips, exhibit significant variability in their performance across different participants and sleep stages. Notably, for light sleep stages, Night Train and Philips algorithms demonstrate comparable outcomes, as represented by the lines which are closely together in the graph with the Mean results. However, their performance does not consistently align on the same participants, indicating variability.

A standout observation is the pronounced discrepancy in performance at the individual participant level for both algorithms. Specifically, there is no consistent pattern where one algorithm outperforms the other across all sleep stages for the same participant. In instances where one model exhibits minimal efficacy (0.00% accuracy), the counterpart tends to achieve considerably better results.

The accuracy and Cohen's kappa metrics (Philips 61.81% with a Cohen's κ of 0,39, and Night Train (56.00% and a Cohen's κ of 0,32) indicating a fair to moderate agreement underscore a vital point: the capability of both algorithms to accurately classify all sleep stages fluctuates markedly across different individuals. This fluctuation, mirrored in the precision and recall metrics for specific sleep stages, suggests that while the algorithms may perform well for certain individuals or in identifying specific sleep stages, their overall generalizability and reliability across a broad spectrum of the population remain questionable.

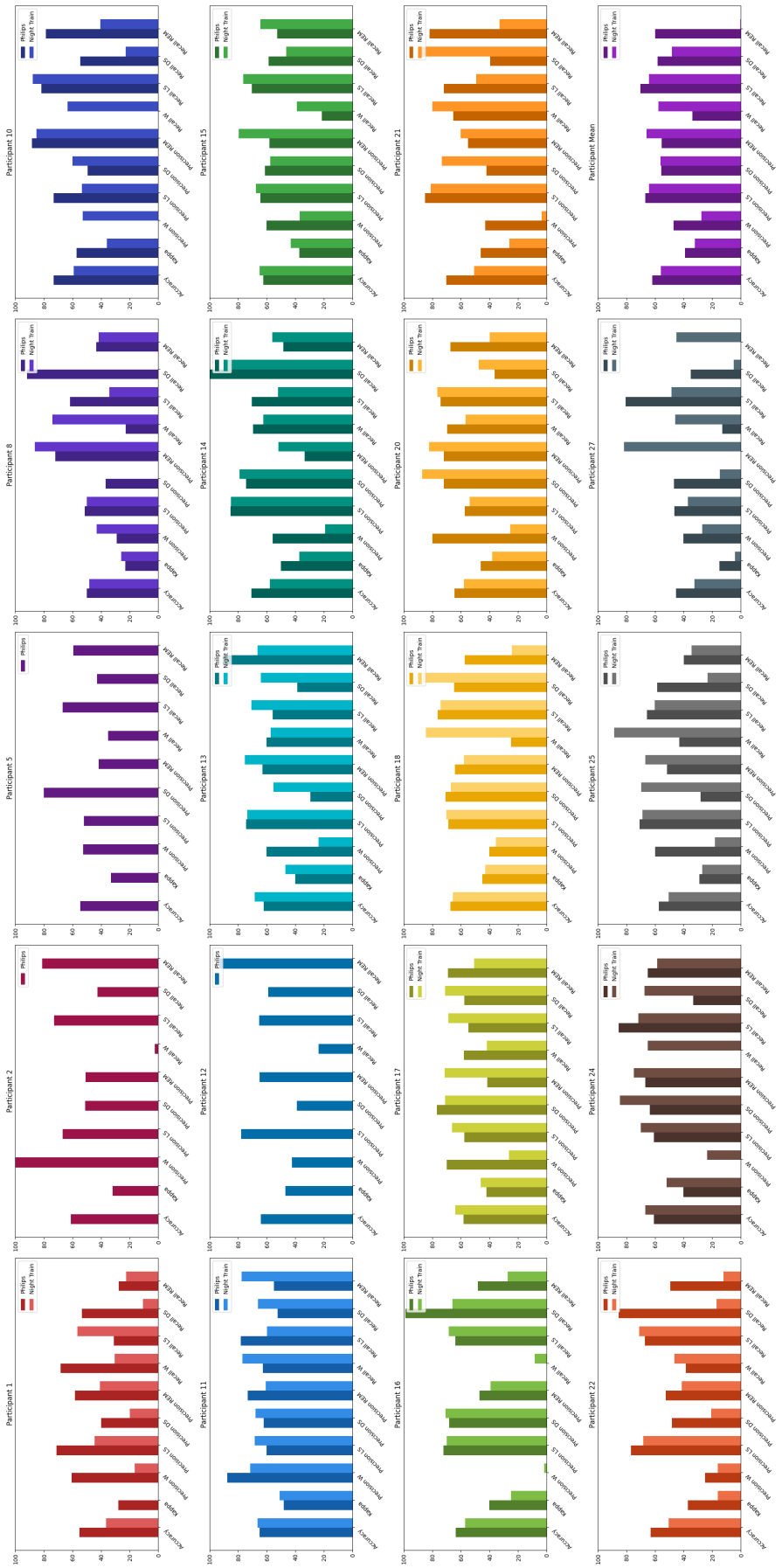


Figure 4.10: Summary of performance metrics for the Philips and Night Train algorithms across participants, including participant-specific results and the aggregate mean for all patients per algorithm. The darker color represents Philips, whereas the lighter color represents Night Train. Kappa value has been done times 100 to better visualise it. Exact results can be found in Appendix D

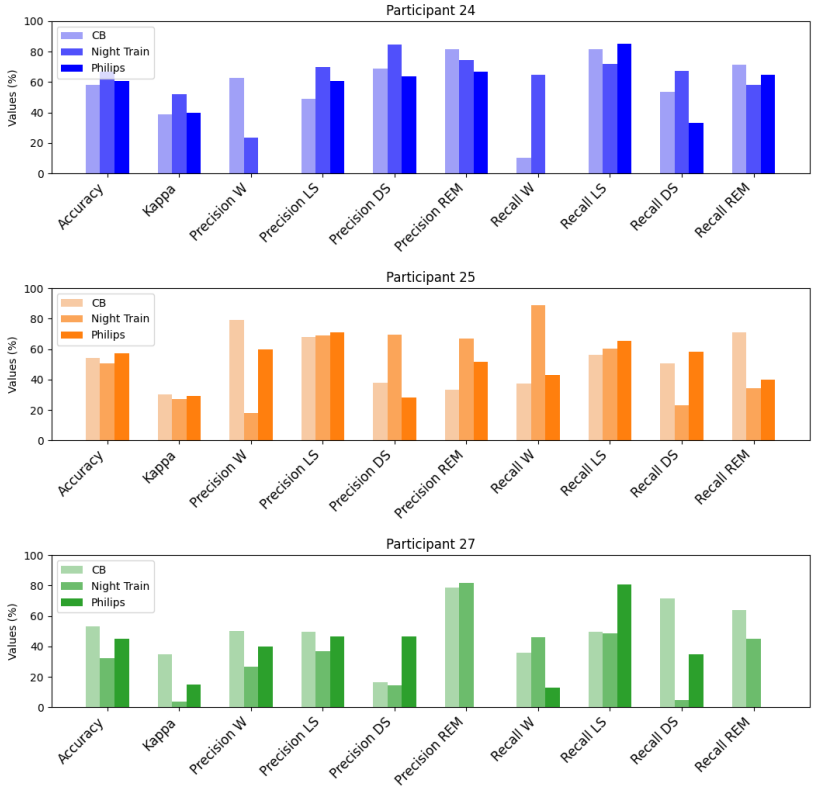


Figure 4.11: Performance metrics of participant 24, 25, 27 with CB model using all features against Philips and Night Train

A direct comparison of our CB model is depicted in Figure 4.11, where the participants 24, 25 and 27 are compared (exact numbers can be found in Appendix E). These participants are within the validation set and have a prediction by both Philips and Night Train. It can be seen that the CB model exhibits performance levels comparable to those of commercially available alternatives. Notably, for participant 27, the CB model achieves the highest scores across various metrics, showcasing its robustness for certain individuals. In this specific individual, there is only one REM period, which explains why the precision and recall scores are either very low, or either very high.

For the other participants CB still scores comparible to the commercial alternatives. Looking at participant 24, it often avoids ranking the lowest. Only for precision for light sleep the CB is doing worse. The same holds for participant 25, where CB is exceptional well in precision for wake and recall for REM. The CB model is at least capable of showing values higher than 0% for all performance metrics. Something that Philips is not able to do, as precision and recall for wake is absent for participant 24 and recall REM is absent for participant 27.

4.4. Feature Influence on Classification

The feature influence analysis consisted of employing permutation importance for the SVM model, alongside the inherent feature importance mechanisms of both the RF and CB models.

Per model, the top ten features were noted to identify commonalities and can be found in Table 4.2.

In the wake-vs-all classification, 'ayB std', 'cole', 'azB std', and 'MagB std' emerged as four common features across all models, all of which are accelerometer-based. Some HRV features also appeared but were not consistently present across the models. 'rmssd detrended' and 'sdsd detrended' were noted in two of the three methods.

For the light-sleep-vs-all classification, the featured factors varied more across the models. Age scores high for both RF and CB. Whereas pdfa is also represented in both. This classification showcased the most pronounced variation in feature importance across in comparison to the other stages.

For SVM mostly accelerometer features are important, where for RF percentiles of IBIs and HR show up often, and for CB it is a mix between all different categories.

The deep-sleep-vs-all classification revealed 'vlf' as highly important for both CB and RF, where the normalized hf and lf features are important for SVM. 'vlf' is lower during NREM sleep compared to wakefulness or REM sleep. This observation can be attributed to the differences in autonomic nervous system activity during these states. Also accelerometer-based features such as 'cole' ranked highly. An interesting appearance is again the top-feature of CB being 'Age', which seem to be highly important.

In the REM-vs-all classification, frequency features like 'vlf', 'hfnu', 'lfnu', and 'lf-hf-ratio' were consistently highlighted in the CB and RF models, with 'dfa' and 'pnn50' also represented. 'MagB std' figured prominently in the top ten features for these models' importance metrics. For SVM, 'vlf' and 'dfa' were also present, but the top important features were accelerometer-based.

It can be seen that the 'cole' metric scores high for wake-vs-all classification over all the models. This metric is typically utilized in commercial alternatives for a sleep-wake classification model solely based on accelerometry data. In the following paragraph, we will delve into the feasibility and accuracy of employing the 'cole' metric for sleep-wake classification within the context of this study. It's important to note that this approach differs from the calculation of Sleep Period Time (SPT) windows, as sleep-wake classification involves identifying not only entire sleep windows but also short periods of wakefulness within the sleep period itself.

4.4.1. Sleep-Wake Classification Using the 'cole' Feature

The exploration of feature influence revealed that accelerometer metrics hold important information for sleep-wake classification. This study investigates the effectiveness of the Actigraph Counts method for sleep-wake differentiation.

The 'cole' feature was evaluated for sleep-wake classification accuracy. In the SPT window analysis, sleep segments were typically merged if the interruption between them was under 60 minutes; however, this step was omitted for sleep-wake analysis. Classification criteria were established as follows: a period registers as sleep when scoring below a threshold of one, and as wakefulness when exceeding this threshold. Subsequently, a smoothing algorithm was applied, requiring a span of five consecutive minutes to confirm sleep and a duration of ten consecutive minutes to verify wakefulness.

This method achieved an accuracy of 90.49% ± 6.15% across participants, with a sensitivity of 93.87% ± 5.13% and a specificity of 59.08% ± 29%. The scores per participants can be found in Figure 4.12. It can be seen that the specificity for participants 8 and 16 is specifically low. These are not the participants with RLS, which are participants 3, 9 and 15. The lower specificity observed in participants 8 and 16 can be attributed to the method's requirement of ten consecutive minutes to verify wakefulness. Both participants exhibited shorter movement periods, leading to a decrease in sensitivity.

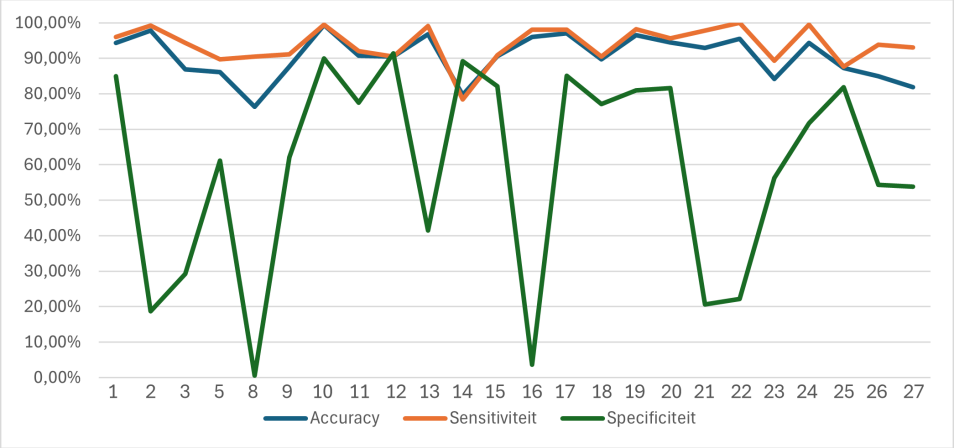


Figure 4.12: Sleep-wake classification using ActiGraph Counts

Table 4.2: Top 10 most important features per sleep stage. For SVM permutation importance is used, while for RF and CB the feature importance functionality within the model is used.

	SVM			RF		CB	
Sleep Stage	**Feature**	**Importance**	**Std**	**Feature**	**Importance**	**Feature**	**Importance**
Wake	cole	0,019	0,001	azB_std	0,104	BMI	11,254
	azB_std	0,007	0,001	cole	0,099	cole	8,518
	axB_median	0,006	0,001	MagB_std	0,086	anglez	8,367
	BMI	0,004	0,001	ayB_std	0,084	dfa	5,976
	axB_std	0,004	0,001	axB_std	0,068	azB_std	5,324
	ayB_median	0,002	0,001	anglez	0,054	pdfa	4,985
	ayB_std	0,002	0,000	sdsd_detrended	0,043	ayB_std	4,286
	hr_detrended__10	0,002	0,001	dfa	0,022	Age	4,072
	azB_median	0,002	0,001	rr_detrended__75	0,018	MagB_std	2,814
	hr_detrended__5	0,001	0,001	rr__25	0,016	axB_std	2,370
Light Sleep	axB_std	0,002	0,002	pdfa	0,038	Age	8,636
	rmssd_detrended	0,001	0,001	Age	0,037	BMI	8,449
	sdsd_detrended	0,001	0,001	rr__75	0,031	anglez	7,051
	axB_diff	0,001	0,000	hr__25	0,030	pdfa	6,177
	ayB_diff	0,001	0,001	hf	0,030	pnn50	4,192
	axB_mean	0,001	0,001	rr__50	0,030	hf	4,159
	MagB_diff	0,001	0,001	hr_mean	0,028	rr_detrended__75	3,967
	sdnn	0,001	0,001	pnn50_detrended	0,026	cole	3,168
	azB_median	0,001	0,001	hr_median	0,026	rr__75	3,016
	cole	0,000	0,002	pnn50	0,025	hr_median	2,228
Deep sleep	hfnu	0,014	0,006	vlf	0,083	Age	10,906
	lfnu	0,014	0,006	sdnn	0,053	vlf	8,469
	sdnn_detrended	0,012	0,009	rr_detrended__5	0,051	cole	7,470
	sdnn	0,009	0,006	cole	0,045	azB_std	5,125
	dfa	0,009	0,006	hr_detrended__95	0,045	anglez	5,038
	sdsd_detrended	0,008	0,004	sdnn_detrended	0,042	BMI	4,541
	mad	0,007	0,005	mad	0,039	pdfa	4,495
	axB_std	0,007	0,009	axB_std	0,033	axB_std	4,303
	azB_std	0,007	0,007	azB_std	0,031	MagB_std	3,568
	lf	0,006	0,004	MagB_std	0,030	hfnu	2,225
REM	ayB_std	0,025	0,002	lfnu	0,071	pnn50	10,981
	MagB_std	0,018	0,002	lf_hf_ratio	0,068	vlf	7,934
	axB_median	0,015	0,003	hfnu	0,058	anglez	6,093
	rr_detrended__90	0,014	0,002	dfa	0,054	lf_hf_ratio	4,354
	MagB_median	0,011	0,002	vlf	0,038	cole	4,218
	vlf	0,011	0,002	azB_std	0,036	lfnu	3,846
	dfa	0,009	0,002	rr_detrended__95	0,034	dfa	3,744
	hr_detrended__10	0,009	0,002	pnn50	0,034	hfnu	3,490
	rr_detrended__10	0,007	0,001	anglez	0,030	rr_detrended__50	2,955
	hr_detrended__90	0,006	0,001	rr_detrended__50	0,029	rr_detrended__95	2,887

5

Discussion

This chapter delves into the broader implications of the study's results. The discussion will encompass a comprehensive analysis of sleep stage classification for wrist-wearables, diving deeper into PSG, comperative analysis and challenges of classification models and automated SPT detection, and if HRV and accelerometry seem fit for sleep classification.

5.1. Error in Ground Truth

In the evaluation of this thesis, the predicted sleep stages were benchmarked against the gold-standard PSG. It is essential to consider that, although the PSG assessments adhere to the AASM criteria, they are not infallible, evidenced by an inter-rater reliability of 82%. This highlights a degree of error within the ground truth used as a comparative measure. Notably, discrepancies often arise between light and deep sleep classifications, suggesting that even the 'gold standard' bears limitations. This inherent uncertainty within PSG readings must be accounted for when analyzing the efficacy of the predictive models developed in this study and implies a need for cautious interpretation of the results. Therefore, it is not neccesarily aimed to reach an accuracy of 100%.

5.2. Automated SPT window Extraction

This study compared two methods for automated SPT window extraction: the $angle_z$ method and the Actigraph counts approach. When examining the OC metric, Actigraph Counts achieved a perfect score of 100%, suggesting complete extraction of the SPT window. In contrast, Angle$_z$ achieved 94.71%, indicating a slight shortfall in fully capturing the SPT window. The MDA metric shows that this is due to a prediction in the wrong direction for the start times using the Angle$_z$ method.

The Bland-Altman also showcases this. For the start time of the SPT window, Angle$_z$ is not able to capture the exact start time of one of the participants. This was due to the fact that the Angle$_z$ method tracked two distinct sleep periods, as the gap between the two periods was more than 60 minutes. Zooming in on the exact predictions for this participant, it was highlighted that the PSG also showed a long wake period with short fluctuations between light sleep and deep sleep. For the other participants, the Angle$_z$ method showed a smaller error in comparison to the Acitgraph Counts method.

The Bland-Altman plot for the end time show that the Actigraph Count method shows a smaller error in comparison to the Angle$_z$ method.

In summary, Angle$_z$ appears to be the preferable method for SPT window extraction, Although the Actigraph Counts method exhibits a fourfold increase in MAE for the start time. The differential effectiveness of the two methods may be attributed to their underlying principles: Actigraph counts rely on a bandpass filter to process raw accelerometer data, which may capture a broader spectrum of movements. In contrast, Angle$_z$ focuses on a specific type of movement—rotation around the z-axis—which may provide a more precise reflection of the patient's sleep movements.

This method focuses an accelerometric signals, presupposes that periods of inactivity correspond to sleep episodes. This analysis has revealed that there remains a disrepancy with an average error

of around 40 minutes for start times and 15 minutes for end times. This variance underlines a key limitation in using accelerometer data for sleep analysis, as periods of inactivity do not necessarily indicate sleep onset.

It is observed that since participants were permitted to remove the wrist-worn device immediately upon awakening, conducting a precise threshold analysis for end time became challenging. If participants directly take it of while laying in bed, minimum movement is measured and therefore the ending of the window is hard to determine.

5.3. Four Stage Classification Model

In analyzing the distribution of sleep stages among our participants, it was evident that class imbalance presented a substantial challenge. Such imbalances can significantly influence model training and performance, as machine learning algorithms may become biased toward the more prevalent classes. The primary cohort showed a sleep stage distribution that aligns with normative sleep architecture, with light sleep predominating, followed by deep and REM sleep, and minimal wake periods, reflective of consolidated sleep patterns.

However, participant 27's data notably diverged from this pattern, with wake epochs exceeding those of light sleep, indicating a night of poor sleep quality or the presence of sleep disturbances. Furthermore, the significant variability in deep sleep across participants suggests heterogeneity in sleep health, possibly influenced by undiscussed factors such as age, lifestyle, or a sleep disorder. The high representation of sleep disorders in the dataset could skew the models ability in classifying 'normal' sleep periods.

5.3.1. Balancing Method

The balancing method employed in this study involved equalizing the number of samples across all sleep stages for each participant by using the stage with the lowest sample count as the standard. This approach aimed to utilize as much available data as possible. However, an unintended consequence of this method was the uneven representation of participants in the dataset. Most participants contributed a comparable number of samples, but participant 19 was overrepresented due to having a higher minimum stage count.

The uneven representation within the dataset poses a potential risk of introducing bias into the training model, especially if the overrepresented data align with specific demographic characteristics or particular sleep disorders. These biases have the potential to distort the model's learning process, resulting in a final predictive model that excels in predicting overrepresented patterns but struggles to generalize effectively across the broader population. Considering the relatively small size of the dataset, consisting of only 24 participants in total, the influence of each individual participant becomes significant. Therefore, a larger dataset is needed.

5.3.2. Classification Models

The fine-tuning of hyperparameters demonstrated a marginal enhancement in performance metrics across the studied models. This observation suggests that while hyperparameter optimization is crucial, its impact on model performance has its limits. Notably, all models converged to similar performance levels post-tuning, underscoring the potential of each model to achieve comparable accuracy given optimal configurations.

Analysis of mean performance metrics, as depicted in Figure 4.9, reveals that the models exhibit parallel trends across the dataset. This parallelism indicates a shared ability among the models to discern patterns within the sleep stage data effectively, despite the inherent differences in their algorithmic approaches.

Interestingly, the reduction of features did not lead to a proportionate decline in performance metrics. This outcome highlights the efficacy of the selected features in capturing the essential characteristics required for sleep stage classification. However, the significance of specific features varied across models, as detailed in Table 4.2.

5.3.3. Variability Amongst Different Folds

Because of the small dataset, I explored the performance metrics using different folds of the four-fold cross-validation. Figure 5.1 presents these metrics per participant for the three different predictive models using the feature subset by CFS1.

Here it can also be seen that when you compare the model performance per participant, the lines overall flow in the same direction.

Notable differences in the lines within one graph suggest model sensitivity to the data subsets they were trained on. Specifically, the RF model for participant 27 demonstrates a significant outlier with one fold showing markedly higher precision for REM sleep compared to the other folds.

These deviations between folds seem more attributable to participant-specific factors rather than to any particular model. For instance, participant 27 exhibits a large variance across all models, while participant 25 displays minimal deviation.

For the SVM model, performance metrics peak variably across participants: accuracy is relatively stable, but precision for wake stages is notably higher for participants 24, 26, and 27, and lower for participants 23 and 25. In the SVM models, recall for wake consistently scores the lowest in comparison to the other metrics.

The RF model maintains a high recall for wake stages without showing consistency in other metrics over the test participants. The accuracy and kappa values remain stable across different folds and participants.

CB's performance pattern is consistent over de test participants, especially in recall scores, except for participant 27, who shows an anomalous low recall for REM sleep. This model generally performs lowest in identifying light sleep but consistently exhibits the highest recall for wake stages.

Figure 5.1: Variability in Performance Metrics per participant by Cross-Validation Fold for SVM, RF, and CB Models using all features.

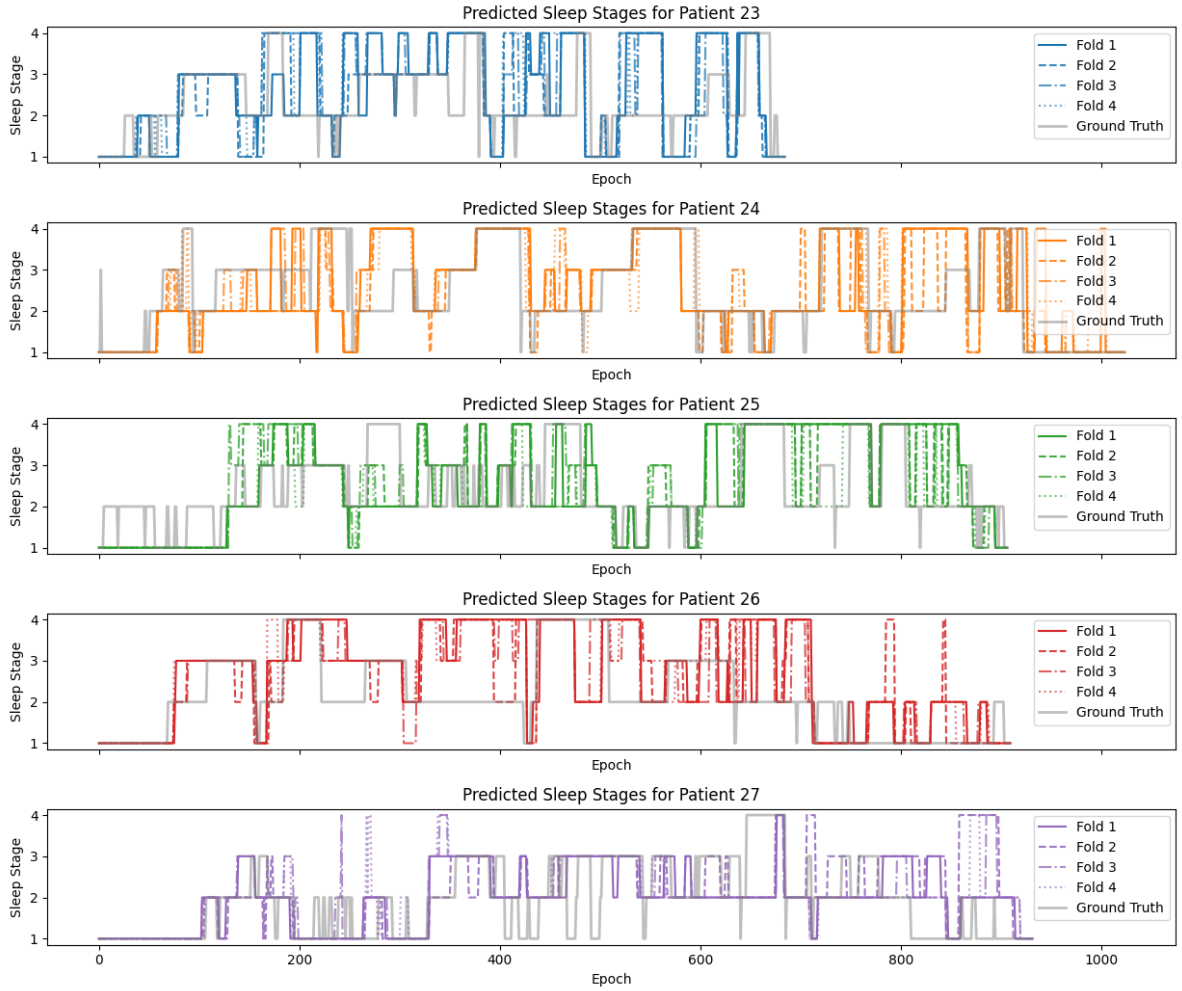


Figure 5.2: Predicted hypnograms for participants 23 to 27 as determined by a random forest model using the CBFS subset of features. Each colored line within the hypnogram corresponds to predictions from a distinct fold of the four-fold cross-validation process. The grey line represents the PSG Hypnogram.

The analysis of model predictions, particularly through the visualization of predicted hypnograms for participants 23 to 27 as shown in Figure 5.2, adds depth to this discussion. Notably, there is a visible variability in the predictions across different folds. This observation underscores the essential role of input data in determining model outcomes. The comparison of model predictions with the ground truth PSG hypnograms—depicted by the grey line—serves as visual representation of the fluctuating accuracy in sleep stage classification.

Furthermore, the observed discrepancies in model predictions across folds highlight the complex nature of sleep data and the influence of data selection on model training and validation.

5.3.4. Size Training Set

The size of the training set used in this study, while robust for preliminary investigations, is modest when benchmarked against other research endeavors in the field, which often incorporate datasets comprising approximately 100 patients. This discrepancy in dataset size is not merely a quantitative matter but has qualitative implications for the study's outcomes and its generalizability.

A larger dataset could potentially offer a richer diversity of sleep patterns, thereby enhancing the model's ability to learn and generalize across a broader spectrum of sleep behaviors. The nuanced variability inherent in a larger sample size would likely contribute to refining the algorithm's accuracy, particularly in distinguishing between more subtle distinctions in sleep stages or in identifying sleep disorders with greater precision. This approach holds potential for significantly reducing the observed

variations during cross-validation, thereby strengthening the reliability of the sleep stage classification models.

5.4. Commercially Available Algortihms

The observed variability in performance between the two algorithms, as depicted in Figure 4.10, underscores the influence of participant-specific characteristics on model outcomes. This variability indicates that neither model uniformly excels across all individuals, suggesting a significant sensitivity to the unique sleep patterns and potential disorders of each participant. It is particularly noteworthy that for certain participants, one algorithm outperforms the other and vice versa, without a consistent pattern emerging across the dataset.

The comparison reveals that the Night Train algorithm generally surpasses Philips in recall for wake stages, with Philips sometimes failing to identify any wake periods accurately, as indicated by a recall score of 0%. In contrast, Night Train achieves considerably higher recall rates, approximating 70% in certain instances. This discrepancy can be due to the fact that Philips only uses IBI values, and Night Train also take accelerometer values into consideration.

In comparison to our own model, which was shown in Figure 4.11, it could be seen that for the participants 24 and 25 the CB model came close to the performance metrics, and even scored highest in accuracy for participant 27. Both Night Train and Philips are trained on a bigger dataset, which holds promise for our models if our dataset will increase.

The performance of these commercially available algorithms underscores the challenges inherent in sleep stage classification, showing that even commercial algorithms struggle in performing consistently well on the same performance metrics, Although our validation set predominantly consists of participants with suspected sleep disorders, these findings still accentuate the difficulties of classifying sleep stages using accelerometry and IBIs from PPG.

5.5. Feature Influence

Via a one-vs-all classification it was explored which features were of significant influence on the classification task. For wake-classification, all three models showed importance for the accelerometer-based features. This is in line with literature, as little movement mostly means that a person is in rest or asleep. Also IBI-features such as rr_range and the 10th or 90th percentiles showed up, which corresponds to the rr_ range fluctuating more during wakeful periods and the lower and higher percentiles being at a higher number compared to sleep periods. The exploratory research of using the 'cole' feature for sleep-wake classification, which yielded a noteworthy accuracy of 90% and a specificity of 59%, also underscores the accelerometer's efficacy in capturing physical inactivity patterns characteristic for wakefulness.

For REM sleep classification, the prominence of frequency features such as 'lf-hf-ratio', 'hfnu', and 'lfnu' aligns with existing research indicating heightened sympathetic activity during REM sleep. This increased sympathetic drive is manifested through elevated low-frequency values and heightened heart rate variability, distinguishing REM from NREM sleep stages. Additionally, the representation of heart rate among significant features during REM sleep further substantiates the link between autonomic nervous system dynamics and REM sleep phenomena. Also accelerometer features play a role in REM classification. This could be explained because during dreaming you could experience little movements.

The observed variability in feature importance between light and deep sleep stages across different models underscores the complexity involved in distinguishing these stages based on unique physiological and movement patterns. This variability emphasizes the subtle differences between the stages, which may not always manifest as clearly defined physiological or movement patterns. Theoretically, one might expect features such as the standard deviation or the range between IBIs to be prominent, reflecting the transition from wakefulness to deep sleep, where heart rate and respiration tend to become more periodic and slow.

5.6. HRV and accelerometer as Sleep Signals

The utilization of HRV and accelerometer data as signals for sleep stage classification hinges on their ability to reflect underlying autonomic nervous system dynamics. While HRV offers a non-invasive glimpse into the heart's behavior as influenced by sleep, its alignment with the AASM criteria, which heavily relies on brain activity patterns, is less straightforward. This disparity arises because the onset and conclusion of sleep stages, as defined by brain activity, might not manifest distinctly within HRV metrics or accelerometer-derived movements, challenging direct comparisons between these modalities.

Also, the extraction of HRV features rely on the IBI-calculation. If the identification of the peaks are not accurate enough or if one of the peaks is mis-identified or not identified at all, then the extracted features fail to capture the variations of the heart. However, the reduced likelihood of movement-induced artifacts during sleep mitigates this risk, enhancing the reliability of HRV data in sleep studies.

6

Conclusion and Future Recommendations

This thesis ambitiously aimed at developing a four-stage sleep classification model using accelerometry and IBI values derived from PPG for the Corsano CardioWatch 287-2. Through the exploration of various sub-research questions, the study aimed to advance our understanding and methodology of sleep stage classification.

It started of with planning a clinical study, which involved obtaining approval from the METC and executing the study, which involved secure planning and communication with the sleep centres. The investigation into SPT window extraction methods revealed the $angle_z$ method as notably effective, capturing the entirety of the PSG recording with a minor discrepancy in start and end times. This precision in capturing sleep windows underscores the method's potential for refining sleep analysis.

Comparative analysis among three optimized models—SVM, RF, and CB—highlighted their competencies and unique feature dependencies in sleep stage classification. The SVM, RF and CB models hold similar performance metrics, although all have different features that are important, whereas one model, CB using all features, scored often the highest for most metrics.

When evaluating against commercial alternatives, there were instances where these models surpassed the capabilities of these commercial alternatives for specific participants or stages. This variability in performance across different patients highlights the inherent challenges in sleep stage classification and underscores the potential for further refinement and enhancement of these models. The observation that no single model consistently excels across all conditions suggests a significant opportunity for advancing the field through targeted improvements and personalized approaches.

Our last research objective focused on the exploration of feature influence on classification. For the sleep stages wake and REM these features aligned with literature, although for light sleep and deep sleep there was no direct link or agreement between models.

6.1. Recommendations for Future Research

6.1.1. Training Set

The variability observed in model predictions across different folds highlights the significant impact of the training set's composition on performance metrics. To address these challenges and improve future models, several strategies are recommended:

- The current balancing technique made sure that most of the data was used, its downside was that one participant could be more represented than others. Equal representation of each participant could be explored, reducing the risk of model overfitting and enhancing its capacity to generalize across a broader spectrum.

- Implementing stricter criteria for epoch selection could enhance the representativeness of included samples. The current methodology involves random sampling from each sleep stage, which may not consistently capture the true essence of the sleep stage. By refining the selection

process to prioritize epochs that are more representative of their respective sleep stages, the overall quality and reliability of the dataset could be improved.

- The inclusion of healthy patients, which are currently not clearly represented in the training set, could enhance the generalizability of the training set. Incorporating data from healthy participants is crucial for broadening the model's understanding of various sleep patterns. This inclusion would not only enhance the detection accuracy of standard sleep stages but also improve the model's diagnostic utility by offering a clearer baseline for identifying deviations indicative of sleep disorders.

- The size of the dataset can even be improved by leveraging online databases to supplement the existing dataset. By integrating a larger and more diverse array of participant data, models can benefit from a richer variety of sleep stage patterns, potentially uncovering new insights and achieving greater accuracy. Online databases provide a valuable resource for expanding the training set beyond the initial 19 participants, thereby enabling more comprehensive and robust model training.

- While the AASM criteria for PSG scoring delineates five distinct sleep stages, this analysis merges these stages into four before training. However, there is potential for enhanced results by training on the original five sleep stages and consolidating them post-training. This approach may better capture the nuances and complexities of sleep stage classification, potentially leading to improved accuracy and performance in the model.

6.1.2. Exploring Probabilistic Outputs for Classification

Investigating the probabilistic outcomes from classification models offers a nuanced approach to sleep stage differentiation. Especially when probabilities for stages such as light and deep sleep are closely matched, implementing adjusted thresholds for classification could enhance accuracy. This method acknowledges the inherent uncertainty and overlap in physiological signals between sleep stages, proposing a more flexible classification scheme. Future research could undertake a detailed analysis of how probabilistic thresholds might be optimized to reflect the complex dynamics of sleep stage transitions more accurately.

6.1.3. Prioritizing Certain Performance Metrics

This thesis employs 10 performance metrics without assigning varying degrees of importance to each. However, prioritizing specific metrics could potentially enhance accuracy in particular areas. Additionally, exploring the integration of multiple classification models proficient in distinct sleep stages could offer a pathway to improving overall classification performance.

6.1.4. Aligning Feature Influence with Theoretical Expectations

Future research should prioritize aligning feature selection with established sleep physiology theories more accurately. The analysis of feature importance revealed that not all theoretically significant features were identified as key contributors to specific sleep stages. Investigating the reasons behind their omission, or alternatively, innovating new features that can more precisely encapsulate the expected physiological changes associated with each sleep stage, could enhance model accuracy and interpretability. This focused approach ensures that model development is not only empirically driven but also grounded in the rich theoretical landscape of sleep science.

6.1.5. Reevaluating Sleep Analysis

Traditionally, sleep stages are delineated based on brain wave patterns, a measurement challenging to accurately capture outside of laboratory settings. For diagnosing conditions like OSA, PSG is not common anymore. It now uses mostly PG, which relies on respiratory effort and SPO2 levels. This shift from PSG to PG prompts a critical reevaluation: what are the fundamental aspects of sleep (disorders) we aim to understand, and might there be alternative markers available? For example for deep sleep, the crucial function is its restorative benefits, are there maybe other biomarkers that might relate to the restorative benefits?

6.2. Summary

In sum, the exploration of wearable technologies for sleep analysis is not without its challenges. While offering significant convenience and reduced intrusiveness with only a small training set, the accuracy and depth of data capture compared to traditional methods must be critically assessed. Future research should continue to explore the effectiveness of accelerometry and IBIs, potentially opening new avenues for understanding sleep and diagnosing sleep disorders more effectively. While wrist wearable definetly provide a more user-friendly and accessible means to study sleep, the question remains: Is it all in the wrist?

Appendices

A

SPT Analysis Tables

Table A.1: Subtables A.1.a and A.1.b presents the performance metrics predicted SPT windows derived from the Angle$_z$ and Actigraph counts methods against PSG data. Performance metrics include the Mean Absolute Error (MAE) and Mean Directional Accuracy (MDA) of start and end times, and mean Overlap Coefficient (OCm), and the standard deviation of OC (OCstd).

(a) Performance metrics of the Angle$_z$ method.

	Start Time		End Time			
Threshold	**MAE**	**MDA**	**MAE**	**MDA**	**OCm**	**OCstd**
0.030	00:12:38	73,68%	00:06:38	52,63%	95,22%	9,57%
0.035	00:14:41	73,68%	00:05:52	68,42%	97,53%	5,45%
0.040	00:17:38	84,21%	00:05:58	68,42%	97,62%	5,42%
0.045	00:17:44	84,21%	00:06:08	68,42%	97,63%	5,42%
0.050	00:21:22	84,21%	00:11:52	78,95%	99,33%	1,82%
0.055	00:26:54	89,47%	00:12:33	84,21%	99,61%	1,49%
0.060	00:26:57	89,47%	00:12:39	84,21%	99,61%	1,49%
0.065	00:27:00	89,47%	00:13:24	84,21%	99,61%	1,49%
0.070	00:27:16	94,74%	00:13:46	84,21%	99,63%	1,49%
0.075	00:30:16	94,74%	00:14:16	84,21%	99,87%	0,45%
0.080	00:30:25	94,74%	00:14:16	84,21%	99,89%	0,37%
0.085	00:30:28	94,74%	00:14:16	84,21%	99,90%	0,33%
0.090	00:32:22	94,74%	00:14:19	89,47%	99,90%	0,33%
0.095	00:32:22	94,74%	00:14:28	89,47%	99,90%	0,33%
0.100	00:32:22	94,74%	00:14:35	94,74%	99,92%	0,33%
0.105	00:32:54	100,00%	00:14:38	100,00%	100,00%	0,00%
0.110	00:33:00	100,00%	00:14:41	100,00%	100,00%	0,00%
0.115	00:33:19	100,00%	00:14:44	100,00%	100,00%	0,00%
0.120	00:38:19	100,00%	00:14:54	100,00%	100,00%	0,00%
0.125	00:39:13	100,00%	00:14:57	100,00%	100,00%	0,00%

(b) Performance metrics of the Actigraph counts method.

	Start Time		End Time			
Threshold	**MAE**	**MDA**	**MAE**	**MDA**	**OCm**	**OCstd**
1.0	00:29:51	73,68%	00:01:28	26,32%	98,40%	1,68%
1.5	00:34:51	84,21%	00:04:47	42,11%	99,03%	1,31%
2.0	00:37:19	89,47%	00:05:02	42,11%	99,33%	0,92%
2.5	00:39:16	89,47%	00:08:47	47,37%	99,40%	0,90%
3.0	00:40:13	89,47%	00:08:57	47,37%	99,43%	0,83%
3.5	00:44:54	94,74%	00:09:03	47,37%	99,50%	0,79%
4.0	00:46:00	94,74%	00:09:09	47,37%	99,52%	0,79%
4.5	00:48:32	100,00%	00:09:09	47,37%	99,54%	0,79%
5.0	00:48:32	100,00%	00:09:19	47,37%	99,57%	0,73%
5.5	00:48:35	100,00%	00:09:25	47,37%	99,58%	0,74%
6.0	00:48:35	100,00%	00:09:54	47,37%	99,67%	0,54%
6.5	00:48:47	100,00%	00:10:19	47,37%	99,69%	0,54%
7.0	00:49:06	100,00%	00:11:16	52,63%	99,84%	0,26%
7.5	00:49:41	100,00%	00:11:16	52,63%	99,84%	0,26%
8.0	00:49:41	100,00%	00:11:19	52,63%	99,84%	0,26%
8.5	00:49:47	100,00%	00:11:22	57,89%	99,84%	0,26%
9.0	00:49:51	100,00%	00:11:22	57,89%	99,84%	0,26%
9.5	00:49:51	100,00%	00:11:28	57,89%	99,85%	0,23%
10.0	00:49:51	100,00%	00:11:33	57,89%	99,86%	0,23%
10.5	00:49:54	100,00%	00:11:46	68,42%	99,88%	0,23%
11.0	00:50:22	100,00%	00:11:49	68,42%	99,88%	0,23%
11.5	00:50:22	100,00%	00:11:52	68,42%	99,88%	0,23%

B

Tables Hyperparameter tuning

Performance metrics per hyperparameter set for different feature sets. The dark green color represents the highest score for that metric, while the lighter green represents the second highest score. Each Table represents a machine learning model.

Table B.1: Support Vector Machine

Feature Set	kernel	C	gamma	Accuracy	Kappa	Precision W	Precision LS	Precision DS	Precision REM	Recall W	Recall LS	Recall DS	Recall REM
		1		49,67%	0,24	32,84%	54,22%	45,42%	45,94%	62,77%	26,11%	57,60%	52,19%
	linear	10		49,30%	0,23	32,53%	54,13%	44,73%	46,58%	61,88%	23,50%	59,43%	52,40%
		100		50,21%	0,24	32,44%	56,75%	45,27%	46,06%	63,21%	24,12%	59,52%	53,98%
		1	0.001	50,66%	0,23	35,74%	56,07%	39,54%	48,95%	70,04%	19,73%	63,01%	49,87%
			0.0001	47,48%	0,20	30,61%	55,02%	41,42%	58,31%	76,78%	13,72%	58,35%	41,06%
	rbf	10	0.001	51,03%	0,24	34,73%	60,89%	46,78%	45,99%	64,24%	25,05%	60,37%	54,45%
			0.0001	50,41%	0,23	35,13%	57,19%	39,14%	49,50%	70,42%	18,79%	63,56%	48,88%
All		100	0.001	49,04%	0,23	31,48%	57,85%	48,73%	41,82%	57,46%	28,20%	55,86%	54,62%
			0.0001	50,94%	0,24	35,79%	57,28%	44,62%	47,87%	66,32%	22,84%	61,97%	52,65%
		1	0.001	50,36%	0,23	33,69%	57,56%	39,26%	50,60%	72,87%	17,48%	63,65%	47,45%
			0.0001	44,78%	0,17	28,45%	58,79%	38,64%	61,00%	79,90%	14,47%	48,27%	36,47%
	sigmoid	10	0.001	49,83%	0,23	35,68%	55,59%	43,96%	47,61%	64,67%	21,01%	62,17%	51,48%
			0.0001	50,33%	0,23	33,97%	57,01%	38,79%	50,65%	71,99%	17,10%	63,95%	48,29%
		100	0.001	47,55%	0,21	31,69%	58,46%	43,90%	39,75%	61,44%	20,44%	56,01%	52,32%
			0.0001	50,67%	0,24	36,86%	58,27%	43,75%	48,12%	66,41%	23,07%	61,79%	51,41%
Feature set	kernel	C	gamma	Accuracy	Kappa	Precision W	Precision LS	Precision DS	Precision REM	Recall W	Recall LS	Recall DS	Recall REM
		1		48,42%	0,20	28,78%	61,64%	40,60%	47,53%	69,87%	16,75%	53,61%	53,44%
	linear	10		49,03%	0,21	31,54%	60,69%	42,52%	46,92%	71,33%	18,46%	52,50%	53,82%
		100		48,71%	0,21	32,61%	60,07%	41,22%	47,64%	70,67%	17,05%	52,95%	54,18%
		1	0.001	47,57%	0,19	30,68%	61,39%	39,35%	50,94%	78,18%	12,86%	47,08%	52,15%
			0.0001	40,71%	0,12	25,66%	58,27%	23,80%	59,83%	80,14%	8,59%	34,79%	39,32%
	rbf	10	0.001	50,30%	0,22	32,77%	60,04%	44,28%	47,29%	74,37%	18,01%	55,38%	53,45%
			0.0001	46,99%	0,18	28,41%	62,25%	39,48%	51,71%	80,40%	12,47%	43,06%	52,03%
CFS1		100	0.001	50,53%	0,23	34,38%	61,32%	45,80%	43,84%	68,29%	22,54%	56,69%	54,61%
			0.0001	49,38%	0,22	31,11%	61,68%	44,14%	50,11%	73,68%	19,51%	51,65%	52,67%
		1	0.001	45,06%	0,16	28,15%	57,03%	37,99%	52,70%	79,02%	12,47%	38,61%	50,16%
			0.0001	38,19%	0,10	23,20%	50,17%	24,79%	61,56%	83,28%	8,54%	26,52%	34,44%
	sigmoid	10	0.001	49,37%	0,21	30,02%	63,54%	45,49%	49,29%	79,21%	15,98%	49,44%	52,86%
			0.0001	45,27%	0,16	27,58%	59,36%	38,92%	52,98%	79,50%	11,06%	40,47%	50,03%
		100	0.001	48,00%	0,20	30,76%	62,50%	44,00%	46,99%	71,25%	17,26%	51,53%	51,97%
			0.0001	49,27%	0,21	29,84%	63,46%	45,37%	50,75%	78,80%	16,30%	49,60%	52,35%
Feature Set	kernel	C	gamma	Accuracy	Kappa	Precision W	Precision LS	Precision DS	Precision REM	Recall W	Recall LS	Recall DS	Recall REM
		1		47,72%	0,19	27,99%	63,47%	42,13%	49,20%	74,66%	16,38%	47,41%	52,42%
	linear	10		47,90%	0,20	28,31%	62,14%	42,60%	47,97%	75,14%	15,34%	48,32%	52,82%
		100		47,41%	0,19	27,70%	62,65%	42,12%	49,02%	75,39%	13,94%	47,71%	52,62%
		1	0.001	45,34%	0,17	27,63%	62,41%	35,21%	52,43%	79,19%	11,14%	41,63%	49,41%
			0.0001	36,88%	0,09	22,10%	39,05%	18,83%	60,12%	86,76%	4,42%	25,43%	30,92%
	rbf	10	0.001	48,33%	0,20	29,09%	63,53%	40,78%	48,66%	77,44%	15,90%	47,09%	52,88%
			0.0001	45,12%	0,17	26,76%	63,18%	37,33%	52,30%	80,69%	11,96%	39,10%	48,73%
CFS2		100	0.001	48,86%	0,20	29,40%	60,27%	42,85%	46,86%	75,97%	14,27%	50,70%	54,52%
			0.0001	47,48%	0,20	28,53%	62,54%	41,86%	48,82%	76,88%	15,56%	45,33%	52,17%
		1	0.001	43,16%	0,15	26,17%	60,07%	29,60%	53,98%	81,62%	10,67%	34,29%	46,04%
			0.0001	33,92%	0,07	20,02%	23,22%	16,49%	61,90%	94,05%	1,53%	16,52%	23,56%
	sigmoid	10	0.001	47,21%	0,19	28,19%	62,96%	41,59%	48,91%	79,22%	14,26%	44,19%	51,17%
			0.0001	43,06%	0,14	26,26%	59,57%	29,12%	54,28%	81,58%	9,49%	35,61%	45,56%
		100	0.001	47,62%	0,19	28,19%	61,46%	42,22%	48,43%	76,05%	15,21%	46,55%	52,65%
			0.0001	47,03%	0,19	27,63%	65,76%	41,90%	49,28%	79,88%	13,84%	43,18%	51,21%

Table B.2: Random Forest model

Feature Set	n_estimators	min_samp_split	max_depth	Accuracy	Kappa	Precision W	Precision LS	Precision DS	Precision REM	Recall W	Recall LS	Recall DS	Recall REM
			10.0	52,46%	0,265916	39,67%	0,615117	37,51%	52,52%	39,98%	37,61%	57,78%	74,48%
		2	20.0	50,73%	0,261904	39,48%	0,599831	40,73%	52,63%	31,35%	47,86%	50,79%	72,91%
			10.0	52,50%	0,264052	41,57%	0,615945	36,89%	51,28%	41,50%	36,32%	58,64%	73,54%
	100	5	20.0	51,04%	0,269019	39,38%	0,593765	41,29%	51,37%	31,57%	48,67%	50,61%	73,30%
			10.0	52,62%	0,266296	43,84%	0,620668	37,28%	51,10%	42,12%	36,59%	59,14%	72,62%
		10	20.0	50,70%	0,261527	36,78%	0,597232	40,87%	51,11%	30,29%	47,07%	52,00%	73,42%
			10.0	52,35%	0,262711	42,43%	0,62192	37,29%	51,96%	39,88%	36,94%	58,80%	73,76%
		2	20.0	50,70%	0,265786	38,95%	0,605302	40,25%	52,69%	27,43%	50,17%	52,35%	72,86%
			10.0	52,27%	0,261098	41,10%	0,613516	37,58%	51,25%	39,64%	36,92%	58,54%	73,97%
	300	5	20.0	51,31%	0,275243	39,28%	0,611901	41,74%	57,71%	31,19%	49,83%	50,85%	73,36%
			10.0	52,18%	0,264866	41,07%	0,622775	38,19%	52,51%	39,46%	38,46%	57,47%	73,34%
		10	20.0	50,49%	0,258744	36,94%	0,598977	41,12%	51,54%	29,70%	48,07%	50,12%	74,09%
			10.0	52,72%	0,271395	43,55%	0,627913	38,01%	52,00%	38,88%	38,68%	59,11%	74,20%
		2	20.0	50,83%	0,264645	39,28%	0,600507	40,95%	52,68%	30,56%	48,93%	50,78%	73,05%
			10.0	52,33%	0,267245	41,21%	0,631886	36,91%	51,02%	38,33%	37,38%	60,02%	73,57%
	500	5	20.0	50,91%	0,267642	39,45%	0,601337	41,83%	51,70%	27,96%	49,44%	52,37%	73,88%
			10.0	52,40%	0,263143	40,71%	0,621318	38,07%	50,99%	39,86%	36,36%	59,51%	73,86%
		10	20.0	50,91%	0,267525	39,88%	0,611478	39,94%	50,00%	31,11%	46,67%	52,34%	73,52%
Feature set	**n_estimators**	**min_samp_split**	**max_depth**	**Accuracy**	**Kappa**	**Precision W**	**Precision LS**	**Precision DS**	**Precision REM**	**Recall W**	**Recall LS**	**Recall DS**	**Recall REM**
			10.0	45,19%	0,19	32,30%	53,76%	40,03%	39,14%	34,29%	33,77%	50,59%	62,09%
		2	20.0	43,92%	0,19	32,65%	54,80%	42,23%	38,04%	26,14%	44,20%	44,99%	60,33%
			10.0	45,50%	0,19	34,45%	53,79%	40,17%	37,44%	36,69%	31,33%	51,76%	62,21%
	100	5	20.0	43,24%	0,18	32,25%	53,79%	40,88%	38,65%	26,26%	42,32%	43,57%	60,82%
			10.0	46,73%	0,21	35,62%	55,73%	40,34%	40,11%	39,09%	33,15%	51,20%	63,50%
		10	20.0	43,49%	0,18	32,01%	54,37%	40,98%	38,20%	25,62%	41,77%	45,05%	61,51%
			10.0	45,65%	0,20	34,40%	54,73%	41,11%	39,21%	34,58%	34,96%	51,56%	61,50%
		2	20.0	43,63%	0,18	31,63%	53,81%	41,32%	38,57%	26,41%	42,85%	44,59%	60,67%
			10.0	46,20%	0,20	34,71%	55,75%	40,03%	38,55%	36,98%	32,27%	52,93%	62,63%
CFS1	300	5	20.0	43,33%	0,18	31,60%	53,66%	40,00%	39,32%	24,72%	42,97%	45,03%	60,60%
			10.0	46,33%	0,20	35,34%	55,60%	39,90%	39,34%	36,72%	34,27%	51,75%	62,59%
		10	20.0	43,63%	0,18	32,29%	53,61%	41,36%	38,06%	26,23%	42,16%	45,00%	61,13%
			10.0	45,75%	0,20	35,73%	55,22%	41,01%	38,85%	35,88%	34,32%	50,34%	62,46%
		2	20.0	43,14%	0,18	33,14%	53,08%	40,88%	38,23%	26,33%	40,47%	44,88%	60,86%
			10.0	46,24%	0,20	34,25%	55,09%	40,13%	38,83%	37,67%	33,58%	51,40%	62,29%
	500	5	20.0	43,54%	0,18	32,29%	54,28%	41,85%	37,35%	25,44%	41,73%	45,63%	61,35%
			10.0	45,94%	0,20	34,42%	54,60%	41,19%	38,81%	36,99%	32,75%	51,11%	62,91%
		10	20.0	43,80%	0,19	32,50%	54,14%	41,34%	38,58%	25,86%	41,04%	46,86%	61,45%
Feature set	**n_estimators**	**min_samp_split**	**max_depth**	Accuracy	Kappa	Precision W	Precision LS	Precision DS	Precision REM	Recall W	Recall LS	Recall DS	Recall REM
			10.0	47,18%	0,20	37,60%	54,56%	37,35%	38,54%	41,12%	31,28%	52,00%	64,29%
		2	20.0	44,73%	0,19	33,50%	54,25%	39,90%	37,72%	32,16%	40,35%	45,44%	60,99%
			10.0	47,03%	0,20	37,36%	54,52%	36,69%	37,63%	42,02%	28,63%	53,96%	63,51%
	100	5	20.0	45,69%	0,20	33,83%	54,69%	39,63%	38,91%	32,16%	38,63%	49,79%	62,20%
			10.0	46,47%	0,19	35,83%	53,60%	37,31%	37,37%	41,38%	28,41%	53,24%	62,85%
		10	20.0	45,42%	0,19	34,64%	53,97%	38,56%	37,38%	32,75%	37,97%	47,22%	63,73%
			10.0	47,33%	0,21	36,29%	55,64%	36,99%	37,56%	39,43%	29,50%	55,92%	64,49%
		2	20.0	45,04%	0,19	34,40%	54,02%	39,11%	38,83%	30,60%	39,50%	47,02%	63,04%
			10.0	46,73%	0,20	37,02%	53,74%	38,16%	37,94%	38,92%	31,57%	52,68%	63,76%
CFS2	300	5	20.0	44,23%	0,18	32,25%	53,98%	38,73%	37,76%	27,60%	40,42%	45,59%	63,31%
			10.0	46,92%	0,20	37,66%	54,90%	37,68%	37,17%	40,33%	29,44%	53,54%	64,38%
		10	20.0	45,16%	0,19	32,87%	54,25%	38,90%	38,70%	30,85%	38,25%	48,19%	63,33%
			10.0	47,64%	0,21	35,84%	55,75%	37,96%	38,72%	41,67%	30,67%	54,52%	63,69%
		2	20.0	44,51%	0,19	31,96%	54,17%	39,09%	38,57%	28,30%	40,34%	47,34%	62,06%
			10.0	47,44%	0,21	37,69%	55,44%	36,80%	38,15%	41,76%	29,91%	54,28%	63,84%
	500	5	20.0	45,78%	0,20	35,76%	54,67%	40,72%	38,59%	33,11%	40,76%	46,60%	62,63%
			10.0	47,24%	0,20	36,17%	54,54%	36,90%	37,95%	40,97%	28,91%	54,53%	64,55%
		10	20.0	45,10%	0,19	33,24%	53,30%	39,60%	37,42%	31,25%	38,73%	47,99%	62,41%

Table B.3: CatBoost)

Feature set	iterations	learning rate	depth	Accuracy	Kappa	Precision W	Precision LS	Precision DS	Precision REM	Recall W	Recall LS	Recall DS	Recall REM
			3	55,11%	0,248543	39,75%	60,75%	36,21%	52,95%	65,88%	16,86%	63,28%	74,42%
		0.01	6	56,14%	0,269069	43,60%	61,74%	35,90%	56,94%	63,98%	24,21%	63,60%	72,77%
	200		3	54,44%	0,287347	41,27%	64,09%	37,58%	58,84%	52,03%	37,15%	58,95%	69,66%
		0.1	6	54,22%	0,293526	41,63%	63,19%	39,71%	59,60%	49,74%	43,49%	54,58%	69,07%
All			3	55,15%	0,246268	38,79%	59,77%	35,49%	55,41%	66,64%	14,26%	66,12%	73,57%
		0.01	6	55,84%	0,260534	41,65%	60,70%	38,20%	56,22%	65,68%	20,00%	64,78%	72,91%
	100		3	55,58%	0,282575	43,31%	63,02%	37,11%	57,26%	57,75%	30,27%	62,14%	72,15%
		0.1	6	54,81%	0,291853	39,34%	62,70%	39,45%	58,59%	50,44%	40,38%	58,22%	70,19%
Feature set	**iterations**	**learning rate**	**depth**	**Accuracy**	**Kappa**	**Precision W**	**Precision LS**	**Precision DS**	**Precision REM**	**Recall W**	**Recall LS**	**Recall DS**	**Recall REM**
			3	48,97%	0,190079	37,86%	55,25%	32,74%	34,85%	58,98%	6,53%	60,99%	69,39%
		0.01	6	48,98%	0,204109	36,80%	56,11%	35,23%	37,83%	55,30%	17,69%	54,88%	68,06%
	200		3	48,34%	0,211959	33,99%	56,67%	39,69%	39,47%	45,95%	26,64%	54,49%	66,30%
		0.1	6	45,66%	0,196564	31,07%	54,77%	42,07%	40,47%	33,20%	36,21%	49,56%	63,65%
CFS1			3	48,00%	0,177793	39,09%	58,65%	30,76%	34,05%	56,16%	2,69%	63,25%	69,90%
		0.01	6	49,54%	0,201685	36,14%	58,57%	34,05%	38,11%	58,52%	13,63%	57,75%	68,23%
	100		3	48,37%	0,20873	36,38%	56,46%	38,38%	38,02%	47,66%	23,31%	55,83%	66,68%
		0.1	6	46,68%	0,195245	31,92%	54,23%	40,12%	39,64%	40,19%	29,03%	52,58%	64,94%
Feature set	**iterations**	**learning rate**	**depth**	**Accuracy**	**Kappa**	**Precision W**	**Precision LS**	**Precision DS**	**Precision REM**	**Recall W**	**Recall LS**	**Recall DS**	**Recall REM**
			3	48,93%	0,187081	38,33%	53,69%	32,26%	34,88%	58,56%	6,35%	61,85%	68,95%
		0.01	6	49,94%	0,207236	36,87%	57,51%	35,12%	38,17%	58,38%	16,80%	56,31%	68,26%
	200		3	48,82%	0,212892	34,92%	55,58%	39,33%	38,77%	46,35%	25,22%	57,28%	66,45%
		0.1	6	47,15%	0,207838	34,20%	55,53%	40,61%	40,17%	39,66%	34,40%	52,53%	62,03%
CFS2			3	48,08%	0,178799	38,05%	58,45%	31,31%	34,37%	56,21%	3,25%	62,61%	70,26%
		0.01	6	49,57%	0,197101	35,86%	57,81%	33,35%	36,66%	58,67%	11,81%	58,12%	69,67%
	100		3	48,63%	0,20039	36,93%	53,80%	36,95%	37,36%	50,46%	20,52%	56,08%	67,47%
		0.1	6	48,46%	0,213773	34,58%	56,44%	39,91%	39,97%	45,96%	29,13%	53,21%	65,52%

C

Tables Performance Metrics Validation Set

Model	Feature set	accuracy	Kappa	Precision W	Precision LS	Precision DS	Precision REM	Recall W	Recall LS	Recall DS	Recall REM
SVM	**All**	41,43%	0,25	59,85%	52,41%	28,10%	54,40%	42,20%	26,13%	75,93%	59,23%
	CFS1	40,64%	0,21	43,11%	40,25%	26,24%	49,76%	55,86%	20,77%	47,12%	57,03%
	CFS2	39,01%	0,21	37,32%	51,07%	25,03%	46,57%	58,98%	17,34%	41,60%	61,68%
RF	**All**	46,26%	0,25	66,76%	47,77%	29,68%	58,54%	26,16%	44,89%	48,76%	73,78%
	CFS1	41,28%	0,20	48,07%	50,38%	23,66%	50,92%	26,47%	33,65%	43,48%	67,57%
	CFS2	44,57%	0,26	51,08%	52,65%	25,92%	53,56%	42,15%	27,97%	56,40%	74,26%
CB	**All**	50,72%	0,31	66,03%	51,57%	33,82%	63,20%	36,39%	48,13%	62,40%	71,97%
	CFS1	42,70%	0,25	50,19%	52,12%	23,21%	47,94%	45,16%	16,38%	57,52%	82,10%
	CFS2	42,62%	0,25	52,05%	53,94%	22,77%	45,38%	51,54%	18,03%	52,65%	77,81%

D

Table Performance Metrics Philips and Night Train

Table D.1: Performance metrics of the commercially available algorithms of Philips and Night Train. The total accuracy and cohen's kappa are given, with the precision and recall per class. The Night Train data is shifted with 34 minutes.

					Philips					
Participant	**Accuracy**	**Kappa**	**Precision W**	**Precision LS**	**Precision DS**	**Precision REM**	**Recall W**	**Recall LS**	**Recall DS**	**Recall REM**
1	55,07%	0,28	60,47%	71,11%	40,00%	58,33%	68,42%	31,07%	53,33%	27,57%
2	61,12%	0,32	100,00%	66,83%	51,16%	50,84%	2,35%	72,94%	42,51%	81,25%
5	54,59%	0,33	52,50%	52,00%	80,00%	41,67%	35,00%	66,95%	42,70%	59,52%
8	49,82%	0,23	28,89%	51,25%	36,67%	72,00%	22,81%	61,65%	91,67%	43,37%
10	73,19%	0,57	0,00%	73,06%	49,33%	88,28%	0,00%	81,74%	54,41%	78,53%
11	65,00%	0,48	87,50%	60,00%	62,11%	73,33%	62,50%	78,20%	52,21%	55,00%
12	64,12%	0,47	42,22%	77,73%	38,82%	65,00%	23,75%	65,27%	58,93%	92,86%
13	62,04%	0,40	60,00%	74,42%	29,41%	62,94%	60,00%	55,75%	38,46%	90,68%
14	70,59%	0,50	55,71%	85,36%	74,29%	33,33%	69,64%	70,29%	100,00%	48,39%
15	62,29%	0,37	60,00%	64,29%	61,11%	58,10%	21,43%	70,31%	58,51%	52,59%
16	63,55%	0,40	0,00%	72,06%	68,15%	46,88%	0,00%	63,94%	98,92%	48,08%
17	58,22%	0,42	70,00%	57,48%	76,67%	41,38%	57,73%	54,67%	57,50%	68,97%
18	67,16%	0,45	40,00%	68,77%	70,67%	64,00%	25,00%	76,26%	64,63%	57,14%
20	64,36%	0,46	80,00%	57,14%	72,00%	71,85%	69,57%	74,07%	36,36%	67,36%
21	70,19%	0,46	42,86%	84,92%	42,00%	55,00%	65,22%	71,94%	39,62%	81,91%
22	62,99%	0,37	25,00%	76,90%	48,24%	52,50%	38,46%	66,97%	85,42%	49,22%
24	60,67%	0,40	0,00%	60,77%	63,64%	66,67%	0,00%	85,41%	33,33%	65,04%
25	57,31%	0,29	60,00%	70,81%	28,00%	51,67%	42,86%	65,50%	58,33%	39,74%
27	45,28%	0,15	40,00%	46,47%	46,67%	0,00%	12,77%	80,61%	35,00%	0,00%
Mean	61,81%	0,39	46,93%	66,68%	55,50%	55,30%	33,84%	70,14%	58,25%	59,98%
STD	6,81%	0,10	28,14%	11,01%	16,06%	18,80%	24,78%	8,26%	21,36%	21,38%
					Night Train					
Participant	**Accuracy**	**Kappa**	**Precision W**	**Precision LS**	**Precision DS**	**Precision REM**	**Recall W**	**Recall LS**	**Recall DS**	**Recall REM**
1	36,56%	-0,04	16,28%	44,55%	20,00%	40,91%	30,43%	56,63%	10,78%	22,50%
8	48,37%	0,26	43,01%	50,00%	0,00%	86,21%	74,07%	34,29%	0,00%	41,67%
10	59,17%	0,36	52,78%	53,33%	60,00%	85,33%	63,33%	87,91%	22,73%	40,51%
11	66,28%	0,51	71,43%	68,39%	67,89%	60,50%	76,92%	59,80%	66,07%	77,42%
13	68,31%	0,47	23,53%	73,40%	55,17%	75,31%	57,14%	70,77%	64,00%	66,30%
14	57,77%	0,37	19,18%	85,06%	78,95%	51,85%	62,22%	51,93%	88,24%	56,00%
15	64,98%	0,43	36,84%	67,40%	57,33%	79,41%	38,89%	76,35%	46,24%	64,29%
16	56,88%	0,25	1,56%	69,79%	70,77%	39,13%	8,33%	68,55%	65,71%	27,07%
17	63,82%	0,46	26,32%	66,26%	71,11%	71,43%	41,67%	68,79%	71,11%	50,72%
18	65,51%	0,43	35,48%	70,04%	67,01%	57,89%	84,62%	74,11%	86,67%	24,18%
20	57,77%	0,38	25,37%	53,90%	87,04%	82,09%	56,67%	76,32%	47,47%	39,86%
21	50,60%	0,26	3,39%	80,99%	73,17%	60,00%	80,00%	49,36%	88,24%	32,81%
22	50,49%	0,16	15,97%	68,25%	20,51%	41,18%	46,34%	70,96%	17,02%	12,07%
24	66,85%	0,52	23,64%	70,06%	84,51%	74,67%	65,00%	71,78%	67,42%	58,33%
25	50,42%	0,27	17,97%	68,79%	69,57%	66,67%	88,46%	60,10%	23,19%	34,38%
27	32,27%	0,04	26,92%	36,84%	14,71%	81,82%	45,90%	48,51%	5,00%	45,00%
Mean	56,00%	0,32	27,48%	64,19%	56,11%	65,90%	57,50%	64,13%	48,12%	43,32%
STD	10,35%	0,16	17,08%	12,58%	26,11%	15,74%	20,87%	13,10%	29,89%	17,37%

E

Performance metrics CB vs Commercial Alternatives

Participant	Model	Accuracy	Kappa	Precision W	Precision LS	Precision DS	Precision REM	Recall W	Recall LS	Recall DS	Recall REM
	CB	58,30%	0,39	62,86%	48,82%	68,95%	81,71%	10,43%	81,36%	53,69%	71,28%
24	**Night Train**	66,85%	0,52	23,64%	70,06%	84,51%	74,67%	65,00%	71,78%	67,42%	58,33%
	Philips	60,67%	0,40	0,00%	60,77%	63,64%	66,67%	0,00%	85,41%	33,33%	65,04%
	CB	54,47%	0,30	79,10%	67,73%	37,74%	33,51%	37,32%	56,02%	50,63%	84,00%
25	**Night Train**	50,42%	0,27	17,97%	68,79%	69,57%	66,67%	88,46%	60,10%	23,19%	34,38%
	Philips	57,31%	0,29	60,00%	70,81%	28,00%	51,67%	42,86%	65,50%	58,33%	39,74%
	CB	53,22%	0,35	50,00%	49,68%	16,67%	78,62%	36,02%	49,51%	71,79%	64,08%
27	**Night Train**	32,27%	0,04	26,92%	36,84%	14,71%	81,82%	45,90%	48,51%	5,00%	45,00%
	Philips	45,28%	0,15	40,00%	46,47%	46,67%	0,00%	12,77%	80,61%	35,00%	0,00%

Bibliography

[1] T. Song et al. “AI-Driven Sleep Staging from Actigraphy and Heart Rate”. In: *PLOS ONE* 18.5 (2023).

[2] F. Versace et al. “Heart rate variability during sleep as a function of the sleep cycle”. In: *Biological Psychology* 63.2 (2003), pp. 149–162.

[3] A. Malhotra et al. “Performance of an automated polysomnography scoring system versus computer-assisted manual scoring”. In: *Sleep* 36.4 (2013), pp. 573–582.

[4] E. Suni, and A. Sing. *Sleep Foundation: Stages of Sleep*. https://www.sleepfoundation.org/stages-of-sleep. Accessed: 2023-04-07. 2023.

[5] A.K. Patel et al. “Physiology, sleep stages [Updated 2024 Jan 26].” In: *StatPearls [Internet]*. 2022.

[6] V.M. Crabtree and N.A. Williams. “Normal sleep in Children and Adolescents”. In: *Child and Adolescent Psychiatric Clinics of North America* 18.4 (2009). Pediatric Sleep Disorders, pp. 799–811.

[7] G. Medic, M. Wille, and M.E.H. Hemels. “Short-and long-term health consequences of sleep disruption”. In: *Nature and science of sleep* (2017), pp. 151–161.

[8] American Academy of Sleep Medicine. *AASM Manual for the Scoring of Sleep and Associated Events*. 2.2. Accessed: 2023-04-07. 2014. URL: https://aasm.org/clinical-resources/scoring-manual/.

[9] H. Danker-Hopfe et al. “Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard”. In: *Journal of Sleep Research* 18.1 (2009), pp. 74–84.

[10] Y. Chou et al. “Comparison between heart rate variability and pulse rate variability for bradycardia and tachycardia subjects”. In: *2018 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE. 2018, pp. 1–6.

[11] L.J. Roelofs. “Sleep Stage Classification for Wrist Wearables: A Comprehensive Review”. In: *Unpublished* (2024).

[12] Medisch Ethische Toetsingscommissie. *Checklist Onderzoekers - Versie 2, D.d. 24-11-2023*. https://www.metc-ldd.nl/content/files/checklist_onderzoekers_-_versie_2_d-d-_24-11-2023.pdf. [Online; accessed 17 April 2024]. 2023.

[13] American Association of Sleep Technologists. *AAST. Technical Guideline Standard Polysomnography*. Accessed: 2024-04-8. 2021. URL: https://www.aastweb.org/Portals/0/Docs/Resources/Guidelines/AAST%20PSG%20Guideline%20Final.pdf.

[14] V.T. van Hees et al. “Estimating sleep parameters using an accelerometer without sleep diary”. In: *Scientific reports* 8.1 (2018), p. 12975.

[15] A. Neishabouri et al. “Quantification of acceleration as activity counts in ActiGraph wearable”. In: *Scientific Reports* 12.1 (2022), p. 11958.

[16] M. Radha et al. “Sleep stage classification from heart-rate variability using long short-term memory neural networks”. In: *Scientific reports* 9.1 (2019), p. 14149.

[17] R Yan et al. “Multi-modality of polysomnography signals’ fusion for automatic sleep scoring”. In: *Biomedical Signal Processing and Control* 49 (2019), pp. 14–23.

[18] G.G. Berntson et al. “Heart rate variability: origins, methods, and interpretive caveats”. In: *Psychophysiology* 34.6 (1997), pp. 623–648.

[19] J. Tindle and P. Tadi. *Neuroanatomy, Parasympathetic Nervous System*. Accessed: 2023-04-07. Treasure Island (FL), 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK553141/.

[20] J.R. Landis and G.G. Koch. “The measurement of observer agreement for categorical data”. In: *biometrics* (1977), pp. 159–174.

[21] P. Fonseca et al. “Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults”. In: *Sleep* 40.7 (2017).