Modelling of sleep behaviors of patients with mood disorders

Abdullah Gunay

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 30.12.2022

Supervisor

D.Sc. Talayeh Aledavood

Advisor

D.Sc. Talayeh Aledavood

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Abdullah Gunay

Title Modelling of sleep behaviors of patients with mood disorders

Degree programme Computer, Communication and Information Sciences

Major Machine Learning, Data Science and Artificial Intelligence **Code of major** SCI3044

Supervisor D.Sc. Talayeh Aledavood

Advisor D.Sc. Talayeh Aledavood

Date 30.12.2022 **Number of pages** 53+1 **Language** English

Abstract

Sleep is an essential function of the human body. It has a restorative effect on both physical and mental health functions. Short and long-term consequences of sleep disruption include changes to stress response, anxiety, and depression, as well as deficiencies in memory, cognition, and performance.

Several methods have been developed to assess sleep. While polysomnography is considered the golden standard of sleep assessment, researchers have focused on alternate ways of tracking sleep using non-intrusive and less costly methods such as actigraphy. Some studies suggested that screen activity from smartphones can be an indicator of the sleep and wake states of an individual as smartphone usage increased drastically in the last decade.

Mood disorders are mental health conditions that disrupt the emotional state of individuals. Sudden and extreme mood changes interfere with the patients' daily rhythm in many ways, including their sleep behavior. Timely diagnosis of the severity of mood disorders plays a critical role in their treatment process. Previous research shows strong links between decreased sleep quality in patients suffering from mood disorders.

This thesis uses the data from a digital phenotyping study, Mobile Monitoring of Mood (MoMo-Mood), to analyze the sleep behaviors of patients with mood disorders using some sleep parameters. In addition, a predictive model is built to investigate the severity of depression using the information tracked via actigraph and bed sensor. Lastly, the perceived sleep quality from questionnaires is compared with the data tracked by these sensors to evaluate the differences in the three different groups of patients: bipolar disorder, borderline personality disorder, and major depressive disorder.

Keywords mood disorders, digital phenotyping, sleep assessment, actigraphy, depression

Acknowledgements

I would like to thank Talayeh Aledavood for her continuous support and encouragement as well as her insightful guidance.

I would like to thank Ferhat, Deniz, and my friends from SUTJEK for always keeping me company and comforting me during the hard times. We are friends for life.

Last but not least, I am forever grateful to have my family's endless love and support at all times. Although we live in different parts of the world now, I always feel your presence.

Otaniemi, 30.12.2022

Abdullah Gunay

Contents

Abbreviations

BD	bipolar disorder
BDP	borderline personality disorder
MDD	major depressive disorder
PSG	polysomnography
WASO	wakefulness after sleep onset
SE	sleep efficiency
TST	total sleep time

1 Introduction

Sleep is undoubtedly an essential process for the health and well-being of humans. Without sufficient sleep, our perception, memory, emotions, and thoughts become unclear and incomplete [1]. The factors affecting sleep quality and the disorders caused by inadequate sleep have been major focus areas of health research. According to the study by Shim et al., depression, tiredness and anxiety are correlated with a lower sleep quality [2]. Another study reveals that individuals with poor sleep quality have higher risks of developing psychiatric disorders and lower quality of life [3].

Sleep quality is usually used as an umbrella term for some metrics that are related to the characteristics of sleep such as total sleep/wake time, sleep efficiency, and sleep onset time [4]. Analyzing sleep behavior requires an understanding of sleep quality measures. Previous research on sleep quality focuses on quantifying these metrics above using a diagnostic tool that is equipped with sensors to monitor the patients and the subjective assessment of sleep. For example, O'Donnell et al.'s investigation of the relationship between subjective evaluations and objective measurement of sleep in healthy adults indicates significant associations and correlations for total sleep time and sleep onset latency [5]. On the contrary, some patients who are experiencing depression and who reported sleep complaints were found to have no sleep problems which might back the idea that subjective and objective sleep measurements of the patients who have mood disorders are inconsistent.

Mood disorders are a group of psychiatric conditions that influence one's emotions and behaviors [6]. Major depressive disorder is known to be the most common mood disorder, affecting 25% of women and half as many men throughout their lives [7]. Another common mood disorder is bipolar disorder which affects approximately 2.4% of the population according to a large cross-sectional survey [8]. Bipolar patients usually experience major depression periods along with manic episodes in which the patient is highly energetic and euphoric [7]. There are also other mental disorders that are more related to the personality than the mood of the person and they are known as personality disorders. Although personality disorders are defined as a different category under mental health conditions, patients suffering from personality disorders usually experience mood swings which can vary in length and frequency compared to mood disorders [9]. There is no single screening tool that diagnoses all personality disorders, clinicians use tests that are specifically designed for each disorder. For example, McLean Screening Instrument for BPD (MSI-BPD) and the Structured Clinical Interview for DSM-IV Axis II are used to diagnose borderline personality disorder (BPD) [10].

Accurate sleep measurement is crucial to assess the parameters affecting sleep quality and evaluate their importance in sleep behavior. Polysomnography (PSG) is widely accepted as the main objective sleep measurement method and it is used to validate other methods thanks to its detailed assessment of sleep [11]. PSG monitors numerous functions of the body: the electrical activity in the brain, eye movements, heart activity, oxygen saturation, muscle activity, and respiratory activity [11, 12]. While PSG offers an accurate tracking of sleep, the cost, time as well as the procedure required to utilize PSG make it an inconvenient method for some sleep studies [13].

Recent advances in wearable technology have allowed researchers to track sleep in a portable, cheaper, and non-intrusive way. Actigraphy is a portable wearable that monitors sleep/wake cycles and motor activities. Wrist actigraphy has been validated by comparing its estimations of total sleep time, wakefulness after sleep onset, and sleep efficiency with ones from polysomnography in many studies [14]. Nevertheless, it is still found to be inconsistent when estimating other parameters and factors such as sleep-onset latency and daytime sleeping [14].

With their increasing capabilities and availability, smartphones have become yet another alternative measurement tool for digital health studies [15]. Although wrist actigraphy has already eased the monitoring of patients, it has a couple of shortcomings when it comes to reproducibility and scalability. Some studies [16] use custom-made actigraph devices equipped with an additional sensor related to the study's objective and the costs of scaling the study to a larger group of participants are still high due to actigraph prices. Phone activity data is also utilized in digital phenotyping studies which investigate the daily rhythm of individuals [17, 18]

Against this backdrop, as the amount of sleep-related data stored in digital platforms increases drastically, machine learning and artificial intelligence-based methods have become the solutions to key challenges in digital health studies as in many domains [19]. Traditional machine learning models rely on tuning a best-performing model for the general population but this does not always yield the best result in health and well-being where the model might need to be optimized at the individual level since individual differences are key characteristics of health-related data. In their study on the prediction of sleep-wake status using actigraphy, Khademi et al. claim that personalized models are especially suitable for sleep data which includes high variability in sleep patterns across individuals [20]. Their results show that personalized models outperform the general models for around 30% of the participants, noting that they might perform even better if the study is extended to analyze activity detection and introduce fully unsupervised models which were implemented in previous studies [21].

2 Background

2.1 Sleep assessment methods

Several methods have been developed for assessing different aspects of sleep. This section covers the most common methods used for measuring sleep parameters and their advantages and disadvantages.

2.1.1 Polysomnography

Polysomnography (PSG) is considered the most detailed method for measuring sleep [11]. PSG as a term was coined by Holland et al. to define the monitoring, analysis, and interpretation of several physiological parameters such as brain activity, muscle activation, eye movements, breathing activity, and oxygen saturation level [22]. There might also be additional parameters depending on the aim of the study. The comprehensive information about many different human body activities obtained by PSG helps researchers to diagnose various sleep disorders such as sleep apnea, insomnia, parasomnia, and REM sleep disorders [12]. Previous studies related to sleep assessment of patients with mood disorders often include PSG as a method to track sleep parameters.

Paul et al.[23] investigated the effects of nightmares on the sleep quality and sleep architecture of BPD patients using sleep recordings from an ambulatory PSG device. The sleep measure included many parameters including total sleep time, sleep efficiency, sleep onset latency, REM latency, and the number of REM periods.

Although PSG is a painless and non-invasive method, it is often considered intrusive due to the subject's unusual sleep environment during the study. It requires control and supervision of the subject under standardized conditions in a laboratory setting. It is also expensive, time-consuming, and labor-intensive [13].

2.1.2 Ballistocardiography

Ballistocardiography is a non-intrusive method to monitor the movement in the body caused by the ejection of the blood at each cardiac cycle [24]. Heartbeat can be estimated from the oscillations that are captured using various types of sensors in BCG such as piezoelectric, pressure, and capacitance sensors [25].

Although most studies focus on the diagnosis of cardiovascular diseases using BCG, it is also offered as an alternative method for sleep assessment in several studies. In Mack et al.'s work [26], a BCG-based monitoring system is compared to actigraphy, using PSG as the gold standard. The study shows that BCG has better performance in detecting sleep onset times than in actigraphy. In [25], wakefulness classification is conducted using the heartbeats measured by a BCG sensor. Another work by Guerrero et al. investigates the detection of sleep-related breathing disorders using a pressure bed sensor that measures ballistocardiographic signals that are later used to analyze respiratory movements [27].

2.1.3 Actigraphy

Actigraphy is a non-invasive method for tracking motor activity [28]. An actigraph is usually in the form of a wrist-worn device that includes an accelerometer to measure movement, a clock to specify timestamps, and a non-volatile memory unit to store information [29]. They can also be customized by integrating additional sensors.

Previous studies have focused on quantifying the agreement between sleep parameters measured by actigraphy and PSG. In Sivertsen et al.'s work on elderly adults treated for chronic primary insomnia, actigraphy is found to have high sensitivity for detecting sleep while having a poor performance in detecting wakefulness [30]. Also, actigraphy underestimated total wake time and sleep-onset latency and consequently overestimated total sleep time and sleep efficiency. Compared with polysomnography, actigraphy captured only part of the treatment effects on total wake time and sleep-onset latency and failed to detect significant changes in sleep efficiency.

In their review of sleep assessment studies including subjects with chronic conditions, Conley et al. claim that actigraphy may overestimate sleep and underestimate wake, and the agreement may be lower in people with chronic conditions who often have poor sleep and low activity levels [31].

2.1.4 Smartphone screen activity

Smartphones have been widely used in various digital phenotyping and sleep assessment studies as their capabilities have improved greatly in recent years [15]. Unlike PSG or actigraphy, they are accessible to the majority of the population. Smartphones are equipped with many different sensors such as accelerometer, pedometer, gyroscope, GPS, and ambient light sensor which allow tracking of different types of data [32]. Users can utilize smartphones to track various personal health information thanks to the increasing number of health-related applications [33]. Common uses of these applications include counting calorie intake, tracking physical activities, and monitoring sleep patterns and quality.

A naive approach to estimate sleep intervals from smartphone data would be done by tracking screen activity. The largest period of inactivity can be defined as a sleep event. This approach assumes the last screen activity would occur right before sleep and the screen would be active immediately after the user wakes up. Also, this approach is not sensitive to any sleep disruptions which might occur during the night.

In their work on identifying chronotypes from smartphone activity, Aledavood et al. described the subjects' active periods as the aggregated intervals of screen-on events [34]. As chronotypes do not change often, they utilized the aggregated data over longer periods of time.

2.2 Mood disorders and sleep

Mood is described as a persistent and prolonged feeling that is experienced inwardly and affects all elements of a person's conduct [6]. Mood disorders emerge in the existence of disruptions in emotions. For example, extreme lows are defined as

depression, and extreme highs are called mania. Sleep disturbances are common for patients with all kinds of mood disorders. Likewise, mood disorders are diagnosed in around 40% of patients with chronic sleep problems [7].

Major depressive disorder (MDD) is a mental disorder that is diagnosed in the presence of pervasive depressed mood and lack of interest in pleasurable activities [35]. Often, MDD has other implications such as low self-esteem, feelings of guilt, inability to concentrate, and disrupted sleep [36]. A review of 63 studies shows that the lifetime prevalence of MDD varies from 2% to 21% for different countries in the world [37]. According to the World Health Organization (WHO), MDD will have the largest share in the global burden of disease by 2030 [35]. Sleep disturbance is considered a symptom of MDD and most patients have irregular sleep patterns. Sleep regulation can act as a precursor for depressive episodes in some individuals with MDD [7].

Bipolar disorder (BD) is a severe mood disorder that is characterized by fluctuations in the mood states of an individual. The global lifetime prevalence of bipolar disorder is estimated to be more than 1% [38]. BD is also a significant cause of disability for the young population as it causes cognitive and functional impairment as well as an increased suicidal tendency. In BD patients, decreased need for sleep is observed during the mania periods and total sleep time is shorter than usual [39]. The study by Barbini et al. (1996) shows significant correlations between shorter sleep duration and manic symptoms [40]. In Perlman et al.'s work [41], depressive symptoms were predicted by the shorter sleep duration in patients with bipolar I disorder.

Borderline personality disorder (BPD) is a mental disorder with a long-term pattern of unstable mood regulation, impulse control, social relationships, and self-image [42]. It is also associated with a high suicide rate and intensive treatment use, which makes early diagnosis very crucial. The prevalence of BPD is estimated to be around 1.6% of the general population according to the surveys [43]. BDP patients often experience disturbances in sleep continuity and altered REM sleep [44]. In Selby's work, BDP was found significantly associated with some chronic sleep problems such as sleep onset difficulties and difficulty in maintaining sleep [45].

2.2.1 Patient Health Questionnaire (PHQ-9)

The Patient Health Questionnaire (PHQ) is a self-report questionnaire consisting of multiple-choice questions that assess the presence and severity of mental health disorders [46]. PHQ-9 is the 9-item module that includes questions about common depression-related symptoms, such as feeling down or hopeless, having difficulty sleeping, and having difficulty concentrating. The PHQ-9 score can be obtained by summing up the answers to each question that can be scored from 0 (not at all) to 3 (nearly every day) and the total score can range from 0 to 27.

The reliability and validity of the PHQ-9 have been assessed in many previous studies. Initially, Kroenke et al. claimed that a PHQ-9 score of 10 or higher had a high accuracy rate for detecting major depression, with a sensitivity of 88% (the ability to correctly identify those with the condition) and a specificity of 88% (the

ability to correctly identify those without the condition) [46]. In the study by Beard et al. [47], a large set of patients with MDD, BD, and other mental disorders (N=1023) were administered PHQ-9 and other several self-reports of depression and anxiety. The PHQ-9 test showed 83% sensitivity and 72% specificity for scores of 13 and higher. Sun et al. investigated the validity of the questionnaire for patients with MDD [48]. First, patients (N=109) were administered an initial PHQ-9, and half of the subjects (N=54) were evaluated again after 2 weeks. The Cronbach's alpha coefficient was found to be 0.892 and the correlation coefficient of the initial test and retest was 0.737.

By using the PHQ-9, healthcare professionals can assess the severity of the depression of a patient and determine the appropriate course of treatment. This may include referral to a mental health specialist, prescribing medication, or recommending therapy. The PHQ-9 can also be used to monitor a patient's progress over time, as a decrease in the score may indicate improvement in symptoms. The table below shows the corresponding depression severity and treatment action for each score group [49].

PHQ-9 score	Severity level	Action for treatment
<5	Minimal	None
5-9	Mild	Monitoring, PHQ-9 retake at follow-up
10-14	Moderate	Treatment planning, counseling, or pharmacotherapy
15-19	Moderately severe	Pharmacotherapy treatment, psychotherapy
>19	Severe	Immediate pharmacotherapy treatment and referral to a mental health specialist in case of inadequate response to therapy

Table 1: Classification of PHQ-9 and Proposed Actions for Treatment

2.2.2 Digital phenotyping

The phenotype of an organism refers to the collection of its observable traits that stem from the genetics as well as the environment of an organism [50]. When it comes to human beings, behavior is a particularly challenging phenotype to delve into due to its temporal and contextual dependence [51]. While orthodox approaches strived to explore it with questionnaires or interviews, digital devices are now an indispensable part of human life and hence offer new methods to understand human behavior. Indeed, since almost all digital devices that humans interact with, including smartphones, computers, wearables, and the internet in general, collect or generate data that can be informative as to the behavior and health of individuals, the term "digital phenotyping" is regarded as a novel technology for data collection in medicine and health research [52].

While there are definitions of varying scope in the literature, digital phenotyping is described as the moment-by-moment quantification of the individual-level human phenotype, in situ, using data from personal digital devices [53]. For instance, biometric and personal data such as voice, pulse rate, or finger taps can be tracked from an individual's smartphone and then analyzed to measure behavior, physiological states, and cognitive functioning [54]. In this sense, the data that are tracked can be both active and passive. While active data is where an individual is required to perform a task or act in a certain way to capture data, such as the completion of a questionnaire, passive data is captured without an explicitly conscious user engagement, namely that no action other than the daily activity of the individual [55].

Digital phenotyping has been utilized in many studies under health and behavioral science and the number of such studies has been rising as the availability and capabilities of the sensors for data collection increase.

Mobile Monitoring of Mood (MoMo-Mood) pilot study by Triana et al. [56] investigated the use of smartphones and wearables for data collection from subjects with MDD and controls. In the study, the data were collected through five different sources: psychological questionnaires, smartphones, experience sampling, actigraph, and bed sensor.

Aledavood et al. [34] identified the daily rhythms of individuals from the data collected through a smartphone app and explored the relationship between social networks and chronotypes. In the study, screen-on events from the smartphone apps for data collection were used as activity identifiers and the frequency of such events was analyzed to create daily activity patterns which are then classified into two chronotypes: larks (above average morning activity) and owls (above average night activity). After constructing the social networks for each subject on call and text data, owls were found to be more central in the social networks and have larger personal networks.

Zulueta et al. [57] investigated the relationship between the severity of mood disturbance and mobile phone keyboard activity to explore the use of passively collected data to predict changes in mood states. Subjects were given a customized mobile phone with a keystroke tracker and they were also given Hamilton Depression Rating

Scale (HDRS) and Young Mania Rating Scale (YMRS) on a weekly basis. Then, linear mixed models with predictor variables such as average typing delay, backspace and autocorrect rates and the total number of typing sessions were implemented to predict the HDRS and YMRS scores of the subjects. As a result, the keystroke activity predicted both manic and depressive symptoms of participants, which could be helpful to identify mood changes in patients with bipolar disorder.

In their meta-analysis on the use of actigraph for the evaluation of mood disorders, Tazawa et al. analyzed 38 studies that utilized actigraphy to compare patient and control groups as well as pre-treatment and post-treatment data [58]. Patients with depression were found to be less active during the day and they had longer time spent awake after sleep onset. Also, total sleep time and sleep latency were longer in euthymic patients than in healthy controls.

3 Materials and Methods

This chapter presents the dataset description and the methods used in this work.

3.1 Dataset description

In this work, the data collected during Mobile Monitoring of Mood (MoMo-Mood) study is used as the dataset. The MoMo-Mood study was carried out in collaboration with Aalto University, University of Helsinki, and Helsinki University Central Hospital in order to evaluate the effectiveness of wearables for quantifying the behaviors and the states of patients with psychiatric disorders [56].

Previous research on the pilot version of the MoMo-Mood study focused on analyzing the subjects in two groups as there were fewer subjects (N=37) [59, 60]. Alakörkkö estimated the sleep duration estimates using the bed sensor data, calculated the correlation between different questionnaires in MoMo-Mood and showed that the sensors used in the study provided similar information about the daily rhythm of the subjects [59]. Hakala classified the controls and patients based on noise, location, and smartphone screen data using different techniques such as LDA, decision tree, and k-means clustering [60]. Recently, Ziaei Bideh investigated the behavioral patterns of the subjects using call, message, and location data from the same dataset [61].

The current dataset includes 164 participants from the following different groups;

- 31 healthy controls
- 21 patients with bipolar disorder (BD)
- 27 patients with borderline personality disorder (BPD)
- 85 patients with major depressive disorder (MDD)

The study included two different phases;

- an initial active period in which the participants were required to answer daily surveys related to their moods
- a passive period in which the participants filled psychological surveys like the PHQ-9

In addition, the participants were asked to use a ballistocardiographic sensor, Murata SCA11H node, and an actigraph, Philips Actiwatch 2, to track their daily sleep and activity levels. Also, a smartphone data collection tool called AWARE was installed on their smartphones to track smartphone use activity from several built-in sensors. Niima study platform, developed at Aalto University, was used to gather this dataset [62]. The details of each sensor and the features of the collected data are described in 3.1.1.

There are six different data sources in the study which can be categorized into two categories; sensors and questionnaires.

3.1.1 Sensors

The sensors used to monitor the sleep activity of the patients are explained in this subsection. Also, the data format of each sensor is described.

Philips Actiwatch 2 Philips Actiwatch 2 is a commonly used wrist-worn actigraph that can track physical activity, sleep, and the intensity of photopic light [63]. The device logs information at 30 seconds epoch length. The device has been used in various studies related to the detection of sleep duration and validation of sleep monitoring devices [64, 65].

Figure 1: Philips Actiwatch 2 [63]

In their study on the relationship between sleep quality and fat mass in college students, Kahlhöfer et al. assessed the sleep quality and total sleep duration using Actiwatch 2 [66]. Lambiase et al. utilized Actiwatch 2 in a study to assess waking movement behavior in older women [67]. The total movement volume and physical activity collected through the actigraph were compared to another wrist-worn actigraph, namely ActiGraph GT1M, and the physical activity from the participants' self-reports. Actiwatch 2 was found to be useful for assessing activity patterns and ranking total movement volume.

Attribute	Description
user	Anonymized user id
device	Identifier of the device used to capture the data
time	Timestamp
activity	Proprietary measure of activity via built-in accelerometer
white light	Illuminance measure (lux/x^2)
interval status	Active/Rest status based on the proprietary algorithm

Table 2: Features of the data collected by Actiwatch 2

Attribute	Description
user	Anonymized user id
device	Identifier of the device used to capture the data
time	Timestamp
hr	Heart rate (beats per minute)
rr	Respiration rate (breaths per minute)
sv	Heart stroke volume
hrv	Heart rate variability
ss	Signal strength
status	Bed occupancy
bbt0	Beat-to-beat time
bbt1	Beat-to-beat time
bbt2	Beat-to-beat time

Table 3: Features of the data collected by Murata SCA11H bed sensor

Murata SCA11H Murata SCA11H is a non-intrusive, ballistocardiographic bed sensor used to track various measurements related to body functions such as circulation and respiration [68]. The sensor includes an accelerometer to collect acceleration data from the movement of the bed that occurs during cardiac activity. The sensor records each parameter at 1Hz frequency and communicates to a server via WiFi for data transfer.

The calibration of the sensor plays an important role in the accuracy of information. The calibration requires four parameters to be set; background noise level during bed occupancy and minimum, maximum, and typical amplitude of BCG signal from an occupied bed [59]

Figure 2: Murata SCA11H

As the status column indicates bed occupancy status, it is expected that the values should stay stable over short periods since it is likely that the status would last for some time. For example, if the bed is occupied during some time at night, there should not be too frequent changes in status. However, there are some counterexamples as shown in the below figure.

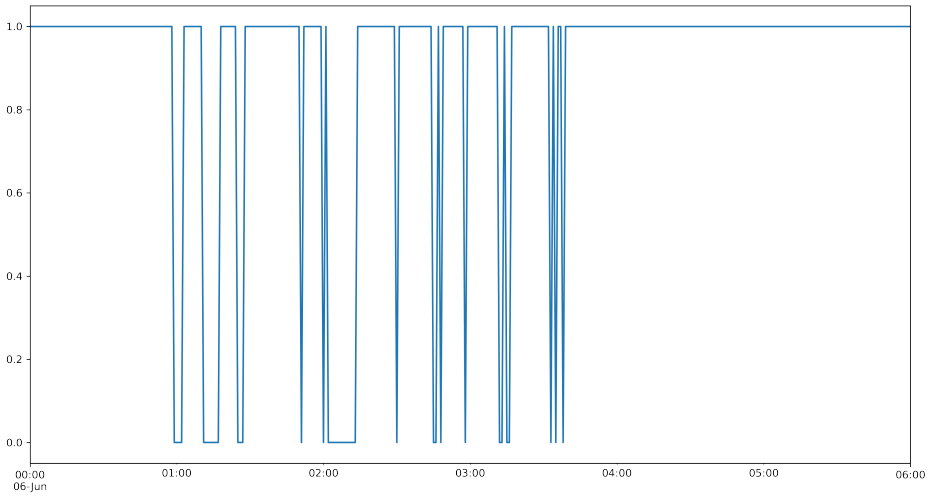

Figure 3: Sleep status changes during a night for the control G96iHU6Pr69i

Given that the bed is occupied when the status value is 1 and empty when the value is 0, Figure 3 indicates that the subject left the bed many times during the night. While there might be a few wake states during the night, the high number of changes in status shows that it is likely that the sensor was not calibrated properly.

AWARE Framework AWARE is an open-source toolkit to collect context on smartphones [69]. It helps researchers to utilize the information that can be captured with various sensors in mobile devices and it is available as an application for both Android and iOS devices. In AWARE, the data is stored in the local device for most small-scale studies but it is also possible to upload the sensor and plugin data to the cloud in a large-scale study. AWARE uses one-way hashing to encrypt data that include personal identifiers.

AWARE tracks data from three different types of sources:

- hardware: accelerometer, photometer, magnetometer, and more
- software: user's call and message activity, calendar
- human-based: questionnaires, voice or gesture input

In this thesis, only the data from the screen sensor is utilized to capture active and passive periods of smartphone use.

Attribute	Description
user	Anonymized user id
device	Identifier of the device used to capture the data
time	Timestamp
datetime	Datetime format of the timestamp
screen_status	Phone screen status (off, on, locked, unlocked)

Table 4: Features of the screen activity data collected through AWARE

3.1.2 Questionnaires

A set of questionnaires were filled out by the subjects during the study period. The timing of the questionnaires was divided into morning and evening. Morning questionnaires were mostly about the mood at wake state and the quality of the previous night's sleep while the evening questionnaires included modified questions from multiple known questionnaires including, The Patient Health Questionnaire (PHQ) [46], Perceived Social Support Scale-Revised (PSSS-R), Morning Evening Questionnaire (MEQ), and Adult ADHD Self-Report Scale (ASRS) [70], Overall Anxiety Severity and Impairment Scale (OASIS), and NEO Five Factor Inventory (NEO-FFI). The scores of each questionnaire were calculated as the sum of the responses to the items. In this work, only the morning questionnaire was analyzed as it contained items related to the subject's sleep [56].

3.2 Methods

This section presents the methods applied to answer the research objectives mentioned earlier.

3.2.1 Defining the features related to sleep behaviour

As the aim of this study is to investigate the differences in sleep behaviors of patients with different mood disorders, some sleep parameters are extracted from the data. The definitions of these parameters and how they are retrieved are explained below.

Total sleep time (TST)

Total sleep time is defined as the time spent sleeping during the night. It is an important sleep parameter as it might indicate abnormal sleeping behaviors such as insomnia, which is characterized by difficulties in staying asleep. TST is also used in calculating sleep efficiency, which is another key parameter in sleep studies [71].

In this work, total sleep time is calculated by checking the information tracked by the sensors. Each sensor used in the study has a different way to indicate the sleep event. The day period is adjusted from 3 p.m. to the next day's 3 p.m. to

avoid miscalculations for sleep events starting after midnight.

In actigraph, *status* column has three possible values;

- **ACTIVE**: high activity, the subject is awake
- **REST**: low activity, the subject is resting
- **REST-S**: low activity, the subject is likely to be sleeping

For actigraph data, the algorithm for extracting the sleep interval uses the largest difference in time for two consecutive records with different status values, defining the first point as bedtime and the last as wake-up time.

Similarly to actigraph, the bed sensor has a status column that indicates bed occupancy. *Status* column in the bed sensor has four possible values;

- **0**: low signal, the subject is not in bed
- **1**: ok signal, the bed is occupied and the subject is likely to be sleeping
- **2**: high signal, the bed is occupied but the subject is likely to be awake
- **3**: signal overload, the measured heart rate is close to the maximum value

For smartphone screen activity, there are four possible values for the activity status of the participant.

- **0**: screen is off
- **1**: screen is on
- **2**: screen is locked
- **3**: screen is unlocked

In screen activity data, the largest inactive period in a day is treated as the sleep interval. It is possible to divide the status codes into two groups; 0 and 3 are possible bedtime indicators while 2 and 4 represent activities that might occur at wake-up time.

Another approach for capturing the sleep event from screen activity could be limiting the bedtime and wake-up time status codes to 2 and 3, respectively.

Wakefulness after sleep onset (WASO)

Wakefulness after sleep onset is defined as the amount of time a person spends time awake during the sleep period [71]. The unit for this parameter is minutes. WASO reflects the fragmentation of sleep which refer to the awakenings during sleep.

In this work, after classifying sleep and wake states for all participants, WASO is calculated by comparing the white light illuminance values during the sleep states with a threshold value. The threshold value refers to the lowest illuminance value that would indicate that the person has turned the lights on. The threshold is decided after an analysis of the luminance values received by the actigraph during the night periods in which the subjects were asleep. The threshold was found as 40 lux which does not contradict the previous research. Dautovich et al. defined dim light exposure as lux values smaller than 100 in their work about the effects of light exposure in healthy adults' sleep [72]. If the luminance value is higher than 40, it is very likely that a light source is activated which could trigger the awakening of the subject.

Sleep efficiency (SE)

Sleep efficiency is defined as the ratio of effective sleep time and total sleep time. It refers to the percentage of total time spent sleeping while being in bed and it is used to determine how well the sleep was [71].

Unlike in previous studies, sleep efficiency is calculated as the percentage of total sleep which was not disturbed by a wake-up event. The wake-up events which are captured to calculate WASO are subtracted from the TST. Effective sleep time is defined as the sleep period under luminance values lower than the threshold, meaning the lights were off.

$$sleep_{effective} = sleep_{total} - waso$$

Then sleep efficiency would be;

$$sleep_{eff} = \frac{sleep_{effective}}{sleep_{total}}$$

Based on the formula above, sleep efficiency can only take values between 0 and 1. A sleep efficiency score of 0 would indicate that the subject was awake during the whole sleep event, which is unlikely to happen. A score of 1 would mean that the subject did not experience any sleep disturbance due to exposure to white light during sleep. Previous research suggests a good sleep efficiency score to be at least 80%, indicating scores less than 80% cause risks to the health and well-being of the individuals. Dew et al. claimed that sleep efficiency below 80% increases the mortality risk in older adults [73]. Another study on anxiety symptoms and sleep efficiency in older women used the same cutoff score for defining good sleep [74]. Åkerstedt et al. suggested a "rather good" sleep would occur at a sleep efficiency of 87% or more while a "rather poor" sleep occurs at an efficiency score of 57% or less [75].

3.2.2 Predicting the depression severity using the information from the sensors

Another research question of this thesis is to investigate whether the severity of depression can be predicted using the data obtained by the sensors for sleep assessment, including actigraph and bed sensor.

Early detection of increasing frequency of depressive episodes from passively collected data can help the patients to receive treatment at the early stages of a mood disorder which makes the treatment more manageable as it reduces the emotional and financial burden [76]. It also allows patients to access support and resources which help them to manage their symptoms and improve their overall quality of life.

Actigraph and bed sensors are non-invasive tools that are used to collect data on a person's activity levels and sleep patterns over time. This data can provide valuable insights into an individual's overall well-being, including their mood and level of distress. A machine learning model that can effectively analyze this data may be able to identify patterns or trends that are associated with different levels of depression severity, which could be useful for identifying early warning signs of depression and for monitoring treatment progress.

The following subsections explain the main concepts of the prediction model implemented for this work.

Decision trees

Decision trees are a non-parametric used mainly for supervised classification and regression tasks [77]. A decision tree predicts the value of a target variable by learning decision rules inferred from the features in the data. These features are called predictor variables [78]. In decision trees, the number of parameters and the structure of the model are decided by the given data rather than an assumption of the distribution.

Decision trees are commonly used for [78];

- Feature selection: It refers to the selection of relevant variables in the data
- Assessing the relative importance of features: After identifying the relevant variables, it is important to understand how each affects the model output. The importance of a feature is based on the loss of the model accuracy when the variable is removed from the model. Usually, the more a variable has an effect on the accuracy, the greater the importance of that variable is.
- Handling of missing data: A common way to deal with missing data is to discard the records which contain missing values in any dimension of the data. However, this causes a loss of information in the data and can result in increasing the bias of the model. Decision trees can be used to treat the missing values as a target variable, to predict them based on the present data which contain information.
- Prediction: Most popular usage of decision trees is to predict the values of a target variable based on a set of predictor variables.

- Data manipulation: It is possible to split the complex categories and break them down into more manageable parts using decision trees.

Depending on the use case, the performance of decision trees can be increased by combining the results of multiple trees [79]. There are several ways to combine decision trees including bagging, boosting, and random forests.

Bagging, or bootstrap aggregating, involves training multiple decision trees on different subsets of the data and then combining the results of the individual trees to make a final prediction [80]. Boosting involves training multiple decision trees in a sequence, where each tree is trained on the errors made by the previous tree in the sequence [81]. Random forests combine both bagging and feature bagging, to train multiple decision trees on different subsets of the data and aggregate their predictions by often averaging [82].

A previous comparison of three methods [83] show that in cases with little or no classification noise, random forest is competitive with (and perhaps slightly superior to) bagging but not as accurate as boosting. On the other hand, if there is a significant amount of classification noise, bagging performs much better than boosting, and sometimes better than randomization.

Overall, combining decision trees can improve the performance of a model by reducing overfitting and improving the generalizability of the model. However, it is important to carefully tune the parameters of the combined model in order to achieve the best possible performance.

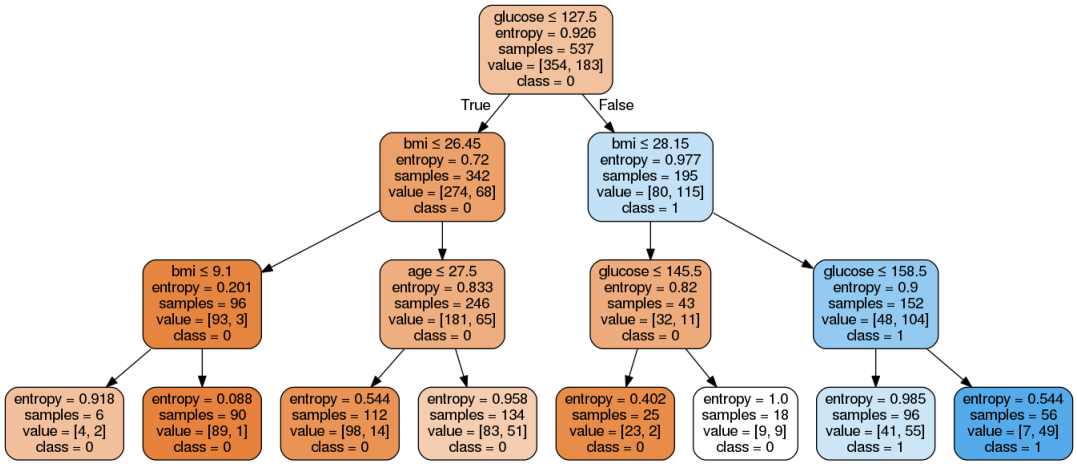

Figure 4: A visualization of a decision tree model for classifying diabetes [84]

XGBoost

XGBoost, also known as Extreme Gradient Boosting, is an ensemble tree method that is implemented using the gradient boosting framework [85]. It is a popular machine learning algorithm used for classification and regression tasks and has been widely applied to various fields such as computer vision, natural language processing, and

finance. XGBoost has gained widespread adoption due to its ability to achieve state-of-the-art results on many different types of data, and its computational efficiency. The algorithm works by iteratively learning a series of weak models, which are then combined to produce a strong, ensemble model [85]. XGBoost uses decision trees as its base learners, and introduces new features such as regularization and sparsity-aware splitting, to improve the predictive performance of the model. The algorithm differs from standard gradient-boosted decision trees as it is a more regularized form of gradient boosting which allows better control on overfitting hence increasing model performance [85].

XGBoost algorithm has also been widely used in medical applications in which a decision is made through the output of a prediction model. Budholiya et al. have utilized a Bayesian-optimised XGBoost classifier to predict heart disease in patients [86]. In their work on improving the diagnosis of depression, Sharma et al. implemented an XGBoost classifier to detect mental disorder cases [87].

Data preprocessing

Data preprocessing is a crucial step in any data analysis or machine learning pipeline, as it involves a range of techniques to clean, transform, and prepare the raw data for further analysis [88]. This process is necessary as real-world data is often noisy, incomplete, and inconsistent, which can hinder the performance of downstream tasks [89]. Data preprocessing techniques aim to address these issues by improving the quality and homogeneity of the data, and by making it more suitable for the intended analysis or modeling. Common data preprocessing techniques include data cleaning, data scaling, data reduction, and data transformation [90]. These techniques are crucial for ensuring the validity and reliability of the downstream analysis, and for enabling the effective use of machine learning algorithms.

In this work, the following preprocessing techniques are involved:

- Data cleaning: the missing information in the data is removed, the noisy data is smoothed, and outliers and other inconsistencies in the data are removed. The data collected in the MoMo-Mood study contains outliers due to inaccurate measurements or incorrect calibration of the sensors.

- Data integration: refers to combining the data from multiple sources to provide a unified view of the whole dataset. The dataset is not collected through a single channel in some real-world scenarios. For example, in the MoMo-Mood study, the subjects are given unique actigraph devices which collect the data separately from each other, then the data from each device is integrated to form a single dataset.

- Data scaling: Each feature in the data collected by the sensors has its own statistical distribution. For example, the data collected by the actigraph have different ranges of values for the light sensor and daily activity. As both features are included in the training data, they are normalized to be interpreted on the same scale.

- Data transformation: The data used in supervised machine learning tasks include a label which is the target variable for the prediction. In this work, labels that reflect the PHQ9 score group are created using the scores of each PHQ9 item as the MoMo-Mood dataset does not contain any labels that provide context for the depression severity.

The classification model

The model implemented in this work is based on a supervised classification task in which the labels 0 and 1 represent moderate and severe depression, respectively. As mentioned in 2.2.1, PHQ-9 is a valid instrument for measuring depression severity therefore the classes of the scores can be used as labels in this task. The classes are divided according to the severity level information in 2.2.1;

- 0: mild and moderate levels combined
- 1: moderately severe and severe levels combined

As there have been multiple instances of PHQ-9 surveys during the study for the participants, some participants have multiple PHQ-9 scores. The scores from all survey instances are utilized to create the dataset for this prediction model. The below table shows the distribution of PHQ-9 scores among different groups.

		PHQ-9 Score Group (%)				
Subject Group	N	1	2	3	4	5
Control	271	94.83	4.80	0.37	0.00	0.00
BD	100	4.00	21.00	24.00	18.00	33.00
BPD	124	8.06	19.35	20.16	18.55	33.87
MDD	577	9.36	22.88	29.64	15.94	22.18

Table 5: Distribution of severity groups of all PHQ-9 tests among different subject groups

Since each PHQ-9 score represents the person's mood within the past two weeks, actigraph and bed sensor data tracked in that interval are combined to capture the relevant data for a specific PHQ-9 score. As the epoch lengths of the actigraph (30 seconds) and bed sensor (1 second) differ, the data from the bed sensor is resampled in 30 seconds bins to accurately match with actigraph data. The datasets are matched by the epoch timestamp, using the subject identifier as the key and a tolerance period of 30 seconds in case of a missing epoch. The PHQ-9 score data is also combined following a similar logic, checking the matched data for timestamps up to 2 weeks prior to the score date.

Then, a decision-tree-based boosting algorithm, XGBoost, is implemented to be trained using the data from the actigraph and bed sensor to predict the depression severity of subjects. Before training the model, the subjects are divided into two cohorts: training and test. The training cohort includes 80% of all subjects (N=40) whose data is used to train the model. The test cohort includes the remaining 20%

(N=11) whose data is used to evaluate the model. Using separate cohorts allows for assessing the generalizability of the model and controlling the bias which might be introduced if a subject's data is used for both training and testing.

The prediction model trained using the data from the actigraph and bed sensor is evaluated using several metrics commonly used in similar classification tasks, including accuracy, precision, recall, and F1 score.

Accuracy is the proportion of correct predictions made by the classifier out of all the predictions made. While accuracy is a useful metric in evaluating classifiers, it can be misleading in cases where the classes are imbalanced, as the classifier may achieve high accuracy by simply predicting the majority class all the time [91].

Precision is calculated as the number of true positive predictions divided by the total number of positive predictions made. It is useful for evaluating the classifier's ability to correctly identify positive instances. In this work, precision refers to the proportion of highly depressed subjects out of all subjects that are classified as highly depressed by the model.

Recall is calculated as the number of true positive predictions divided by the total number of positive instances in the data. It refers to the proportion of highly depressed subjects that were correctly identified by the model.

F1-score is the harmonic mean of precision and recall. It is calculated as the product of precision and recall divided by the sum of precision and recall. The f1-score is a useful metric for evaluating classifiers when the goal is to balance precision and recall.

The choice of evaluation metric depends on the specific goals of the classification task. In some cases, maximizing accuracy may be the most important goal, while in other cases, maximizing precision or recall may be more important. Since the aim of the classifier is to correctly detect the subjects whose PHQ-9 scores fall into moderately severe and severe categories, it is desired to minimize the false negatives as the consequences of identifying a highly depressed subject as mildly depressed would cause more harm to the treatment process. On the other hand, classifying subjects with mild depression as highly depressed could result in unnecessary treatment, increasing the costs and the chances of unwanted side effects.

Another aspect of model evaluation is to understand how the model's predictions are decided and how each feature contributes to these predictions. The need for model interpretability has increased as machine learning models are present in many decision-making systems in a wide range of domains including healthcare [92]. Model interpretability is particularly important in healthcare as the consequences of incorrect or misunderstood predictions can be severe [93]. For example, if a machine learning model is used to predict the likelihood of a patient developing a particular disease, it is important that the model's predictions can be understood and trusted by both the patient and the healthcare provider. This is because the patient may make treatment decisions based on the model's predictions, and the healthcare provider may use the model's predictions to guide their treatment recommendations. Also, model interpretability is needed for regulatory purposes to ensure the predictions are transparent and trustworthy [94].

Elshawi et al. [95] compared five global and two local interpretability techniques

in a machine learning model implemented for predicting hypertension. They found the global techniques had an edge over local techniques as they can explain the entire set of features. In applications in which the risk of a disorder is predicted, it is more important to understand the main risk factors than each factor's role in the outcome. Local explanations of individual cases can be combined to get a global interpretation, but this approach would be computationally expensive.

In this work, model interpretability is explained using Feature Importance and Shapley Values. Feature importance is a global interpretability technique as it evaluates each feature's importance by the increase in the model's prediction error when the feature is absent [96]. Shapley value explanation is a local interpretability technique that is derived from game theory [97]. Each feature in the model is assumed to be a player in a game where the prediction is the payoff. The Shapley value fairly distributes the payoff among the features by evaluating the model's performance when each feature is included or excluded in a set of features.

3.2.3 Analyzing the perceived sleep quality from morning survey responses

During the course of the MoMo-Mood study, the subjects were asked to answer some questions at different stages of the day. In the morning questionnaire during the active phase of the study, they received a couple of questions about the previous night's sleep.

- Q1: Did you sleep well?
- Q2: Do you feel well-rested?
- Q3: How many hours did you sleep last night?

Q1 and Q2 have a 7-point Likert scale for the responses, ranging from 1 (strong disagreement) to 7 (strong agreement). Q3 has an interval scale as the options start from less than 5 hours up to more than 10 hours with hourly intervals in between.

All three questions are used to reflect the perceived sleep quality of the participants. A correlation analysis is conducted between the survey answers and the sleep features from the sensors to reveal possible relationships between perceived sleep quality and actual sleep quality tracked by the sensors.

4 Results

This chapter includes the findings from the sensor analyses using the methods described in the section 3.2.

4.1 Total sleep time

Daily total sleep time is one of the most common features used in studies related to sleep quality. Previous research shows bidirectional links between mood disorders and reduced/increased total sleep time [98] depending on the type of the disorder.

Total sleep time is calculated as the difference between the bedtimes and the wake-up times. It is calculated by the methods explained in 3.2.1 for each sensor. In order to compare the sleep amount tracked by each sensor, each subject group is filtered so that the included participants are tracked by all three sensors on a given day.

Subject group	N	Data length (rows)
control	12	2155
bd	12	1634
bpd	15	1120
mdd	38	6011
all	77	10920

Table 6: Statistics of the filtered data

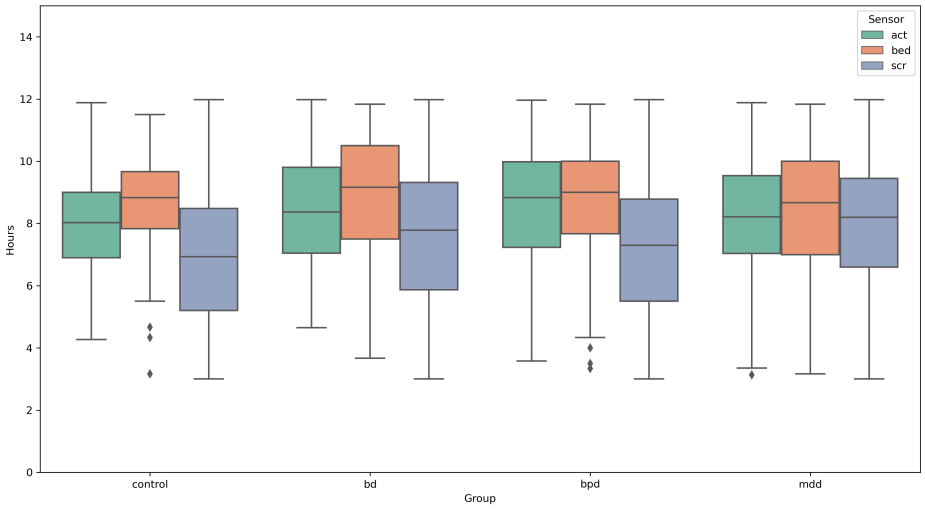

Figure 5: Total sleep amount in each sensor for all groups

Figure 5 shows the distribution of total sleep time in each sensor for all subject groups. Total sleep times tracked by the bed sensor are lower while actigraph has the highest values. The median sleep amount in the bed sensor is also the lowest compared to the other two sensors in all groups. The groups share a similar pattern when it comes to the distribution of sleep amounts in each sensor.

Figure 6 below shows the comparison of two different approaches to detect the sleep event for screen activity mentioned in the section 3.2.1. The median TST seems to be higher in the new approach as the nightly sleep is no longer disturbed as often as in the old approach where the longest inactive period was defined as sleep. In the new approach, periods during the night spent inactive that are longer than 1 hour is also counted in nightly sleep. The dispersion seems to remain the same for almost all subject groups.

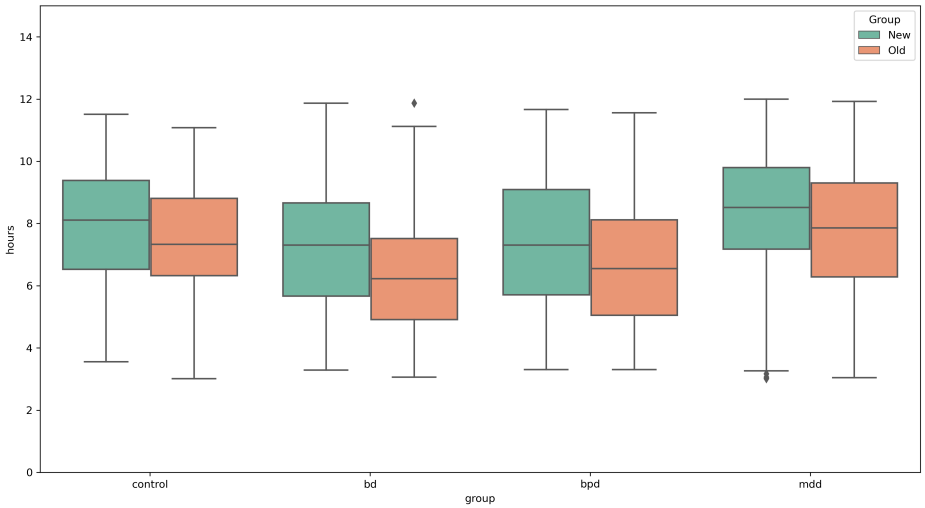

Figure 6: Comparison of two approaches for calculating sleep interval from screen activity

4.2 Difference in bedtime

Actigraph is selected as the ground truth for bedtime comparisons since it provides the most stable sleep status for a user. In the bed sensor, there are abrupt changes in status which make it harder to detect the sleep event accurately.

The regular interval for data collected by the bed sensor is one second. Figure 2 is an example of the sudden status changes in the bed sensor.

Comparing total sleep amounts is not sufficient to understand whether the information from the sensors agrees with each other. Bedtime and wake-up times can be very different although the sleep amount is close for the two sensors. Figure 7 shows the distribution of differences between the actigraph's bedtime and bed sensor's bedtime for all groups. MDD patients have the largest dispersion while the differences in the control group and BD patients vary less. The bedtime differences are shorter than half an hour for the controls and BD group. For BPD and MDD, it is slightly over 30 mins.

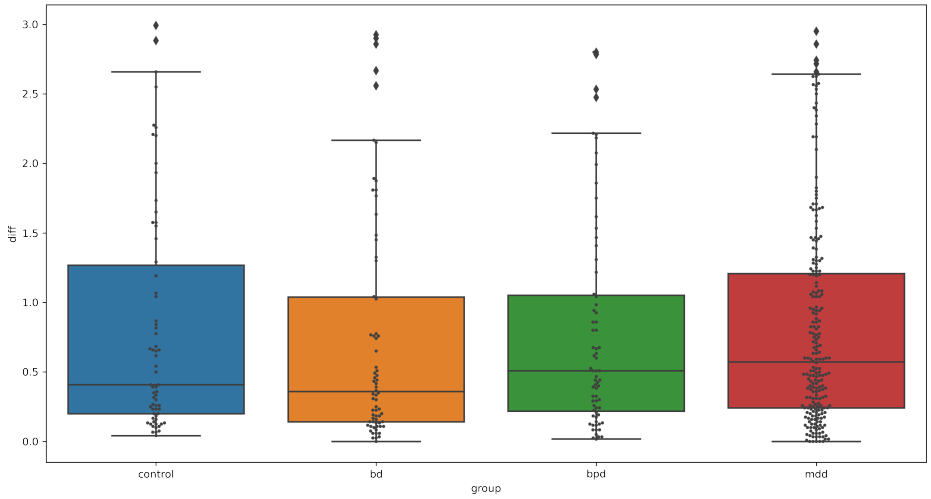

Figure 7: Difference of bedtimes in hours for actigraph and bed sensor

In Figure 8, the difference in bedtimes in screen and actigraph is shown. The median difference is around half an hour for the control, BPD, and MDD groups while it is close to an hour for the BD group. Also, the dispersion is largest for the BD group. Compared to Figure 7, the distributions of the control and MDD groups remain very similar.

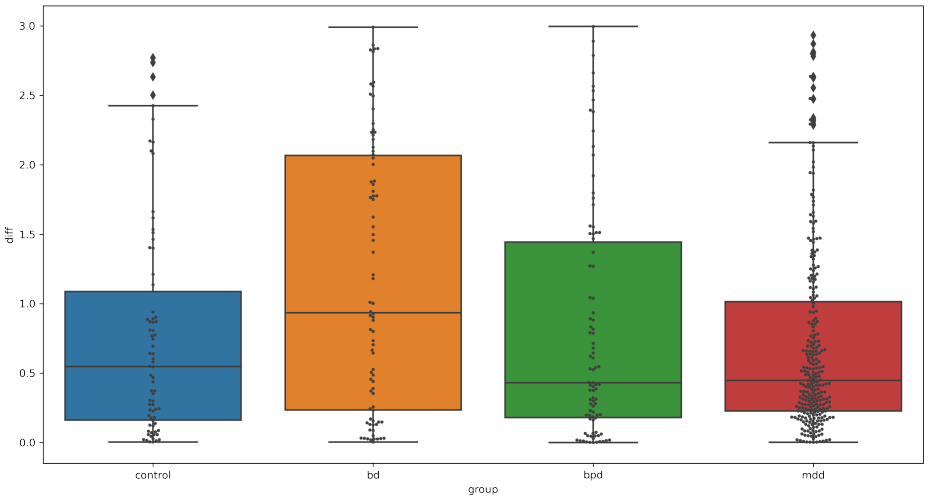

Figure 8: Difference of bedtimes in hours for actigraph and screen activity

4.3 Sleep efficiency and time spent awake after sleep onset

As defined in the section 3.2.1, sleep efficiency (SE) and time spent awake after sleep onset (WASO) are commonly used parameters for determining sleep quality. As sleep disturbances are a well-known symptom of mood disorders, our hypothesis is that the subjects in the control group would have lower WASO and higher sleep efficiency. The figures 9 and 10 show the cumulative distribution of the values for each of the parameters. It is important to note

As expected, both figures reflect a similar pattern as sleep efficiency is calculated using WASO and TST. Controls have the highest proportion of smaller values of WASO as was stated in our hypothesis. However, it does not apply to all the data points as there are also controls with a WASO greater than a patient in another group. MDD patients have the highest proportion of smaller WASO values in the patient groups and the highest values of WASO were observed in BPD patients. At least 70% of the WASO values are shorter than one hour in all subject groups. Extreme values of WASO (higher than 2 hours) are present in around 20% of the BPD and MDD groups. In controls and MDDs, they only occur in around 5% and 10% of the data, respectively (Figure 9).

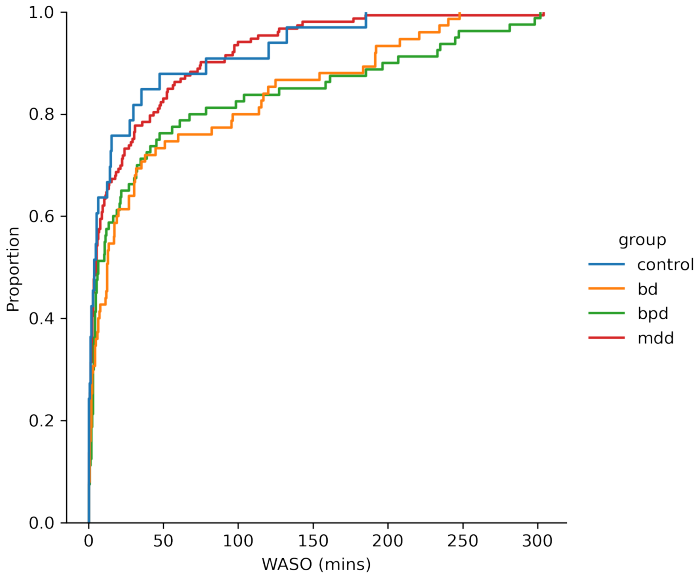

Figure 9: Empirical cumulative distribution function of WASO

In Figure 10, sleep efficiency scores obtained after subtracting WASO from total sleep time are given. Controls are MDD groups have the largest share of a sleep efficiency score of 1 which means there were no wake-up events during the course of the sleep.

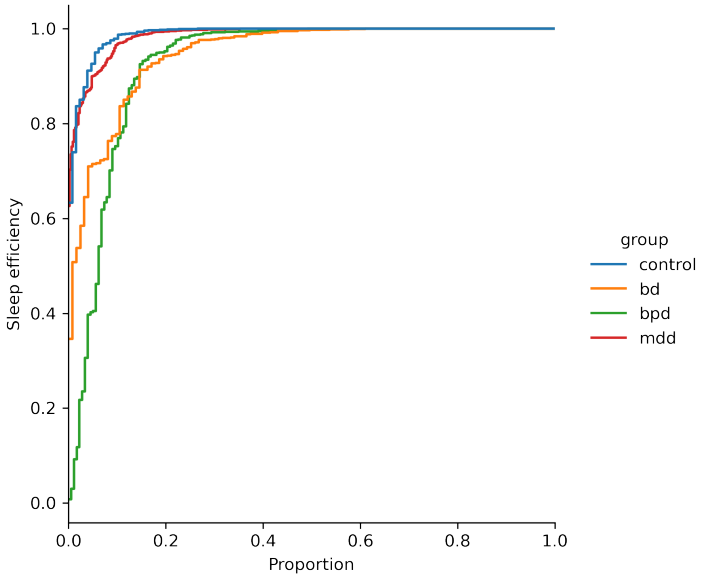

Figure 10: Empirical cumulative distribution function of sleep efficiency

4.4 Predicting the severity of depression

As discussed in 3.2.2, this work aims to show whether the data received by sleep assessment tools could indicate the severity of depression of an individual. This section presents the evaluation of the classification model using the metrics described in the model subsection in 3.2.2.

Figure 11 illustrates the normalized confusion matrix for the test set. Normalization allows for easier comparison of the performance of the classifier across different classes, as it removes the influence of class imbalance. From the axes, it is seen that the areas top-left and bottom-right define the true negatives (TN) and true positives (TP), respectively. Top-right and bottom-left refer to the false positives (FP) and false negatives (FN) in the same order. The model achieved an accuracy of 0.77 for the positive class but only an accuracy of 0.24 for the negative class. It overestimates the severity of depression as the positive class refers to high levels of depression.

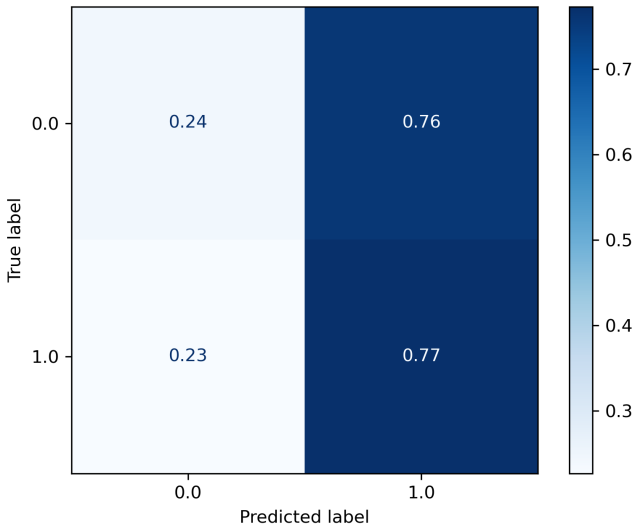

Figure 11: Confusion matrix of the trained model

Figure 12 shows the feature importance scores of each input feature of the model which means the relative contribution of each feature to the overall prediction made by the model. The values are assigned based on the number of times the features are used in the tree. Features coming from the data tracked by the bed sensor seem to have the highest contribution. The weights of these features are very close as they are related. For example, rr (respiration rate) is linked to hr (heart rate). Activity level from actigraph has the least contribution in the model's decision on depression severity.

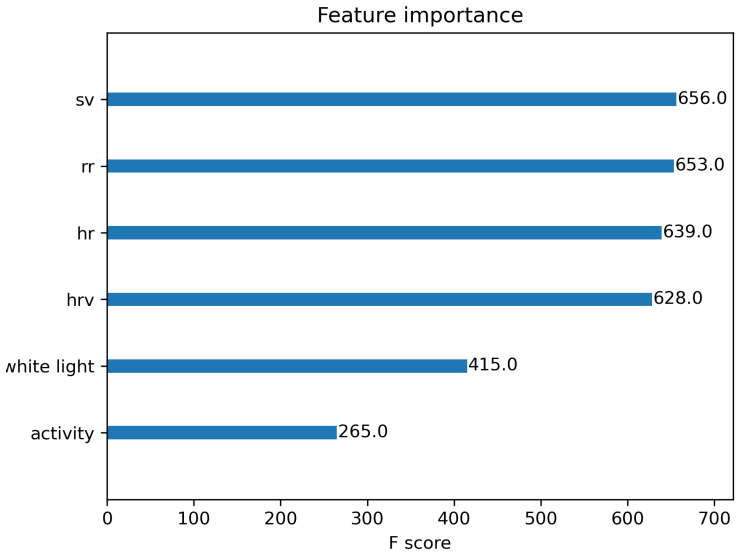

Figure 12: Feature importance scores of the model

Shapley values of the model features for the whole data are given in Figure 13. As defined in 3.2.2, Shap values indicate the impact of features in the model output. Feature value indicates the numerical value the features get. Heart rate (hr) does not seem to affect the model output as much as other values as both low and high values of hr are assigned positive SHAP values. On the other hand, lower values of heart rate variability lead to lower chances of severe depression. Higher values of heart stroke volume (sv) lead to higher chances of classifying high levels of depression. As activity level increases, the model predicted more likelihood of depression. It is worth noting that the separation of Shap values for a feature should be more distinct based on the feature value. This figure shows clearly that the model has done arbitrary predictions for some data points as the effects of the features remain unclear.

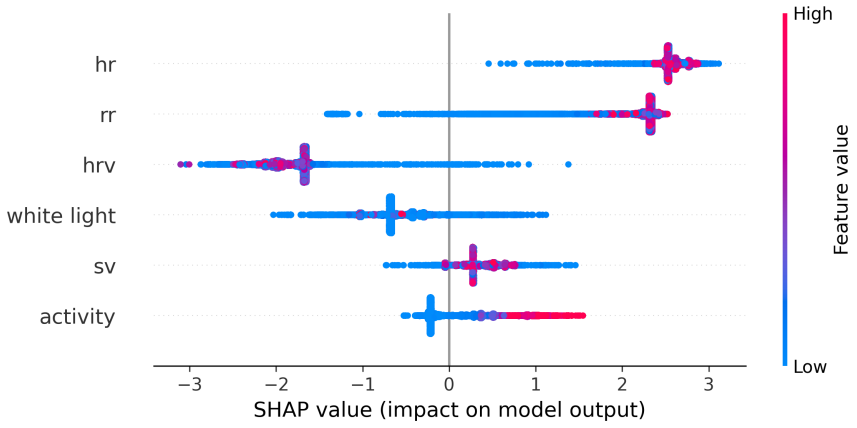

Figure 13: Summary of the Shap values of the features

4.5 Analysis of the perceived sleep quality from the morning questionnaires

During the course of the MoMo-Mood study, the subjects were asked to answer some questions at different stages of the day. In the morning questionnaire during the active phase of the study, they received a couple of questions about the previous night's sleep.

- Q1: Did you sleep well?
- Q2: Do you feel well-rested?
- Q3: How many hours did you sleep last night?

Q1 and Q2 have a 7-point Likert scale for the responses, ranging from 1 (strong disagreement) to 7 (strong agreement). Q3 has an interval scale as the options start from less than 5 hours up to more than 10 hours with hourly intervals in between.

The statistical summary of the responses for Q1 is given in Table 7. Most of the responses come from the MDD group as it contains the highest number of subjects. However, the average number of days the survey was answered is highest in the BPD group followed by the control group. Controls have the best-perceived sleep quality followed by the BD and MDD groups.

Group	N	Total response	Avg. days answered	Mean of the answers
control	30	479	15.97	5.36
bd	20	268	13.4	4.41
bpd	18	305	16.94	3.57
mdd	57	840	14.74	4.07

Table 7: Statistics of the responses for Q1

Participation statistics are the same for both Q1 and Q2 according to Table 8. Also, controls and BPD patients are in the same order as the previous question's answers. MDD patients feel more well-rested than BD patients unlike Q1 answers, but the difference between both groups seems to be lower than in Q1.

Group	N	Total response	Avg. days answered	Mean of the answers
control	30	479	15.97	5.13
bd	20	268	13.4	3.55
bpd	18	305	16.94	3.04
mdd	57	840	14.74	3.69

Table 8: Statistics of the responses for Q2

Figure 14 and Figure 15 show the average TST tracked by actigraph for each answer in both questions. Total sleep time is not directly related to both sleeping well and feeling well-rested as there are other parameters to take into account, including WASO and sleep latency. TST might become a better indicator in case of low scores for both of these questions as the lack of sleep impacts the morning mood.

In Figure 14, survey scores of 1 or 2 have the least total sleep time for all groups. Controls have less TST compared to other groups for all responses. The difference between the TSTs of the groups is the smallest around the survey responses close to the average. The Pearson correlation between TST and the survey answers is highest for the BD group (0.34) while it is lower than 0.1 for the rest of the groups. As mentioned earlier, total sleep time is not the only element of good sleep hence a correlation is not expected.

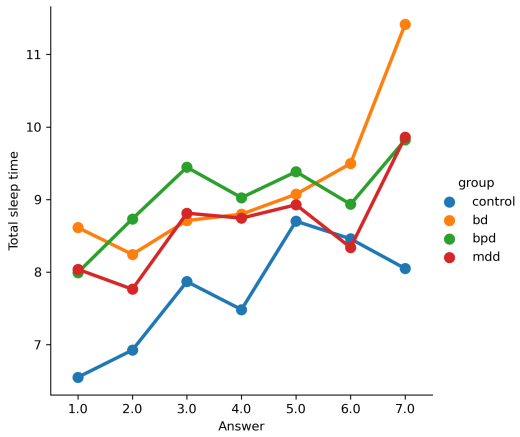

Figure 14: Mean values of TST for each response in Q1

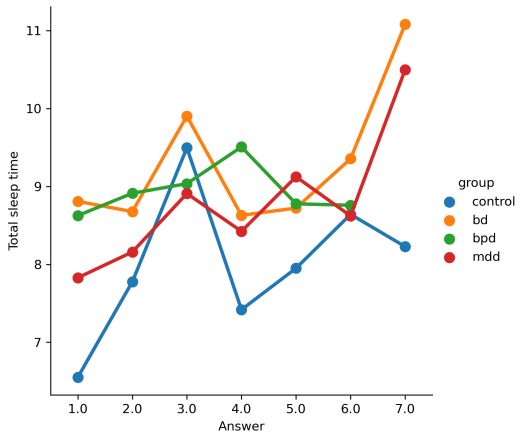

Figure 15: Mean values of TST for each response in Q2

Figure 15 shows a similar trend as controls and MDDs have shorter TSTs compared to BDs and BPDs, especially the lower end of the survey answers (1 to 4). Answers on the higher end (5 to 7) are present when the TST is around 8.5 hours.

Figure 16 shows the histogram of the counts of each answer to the Q3. Total sleep times of controls, BDs, and MDDs seem to have normal distribution as the data is symmetrical and the data around the mean is more frequent. It is not possible to compare the total sleep time from the sensors with the survey responses as the responses are in categorical order and no numerical data points are included.

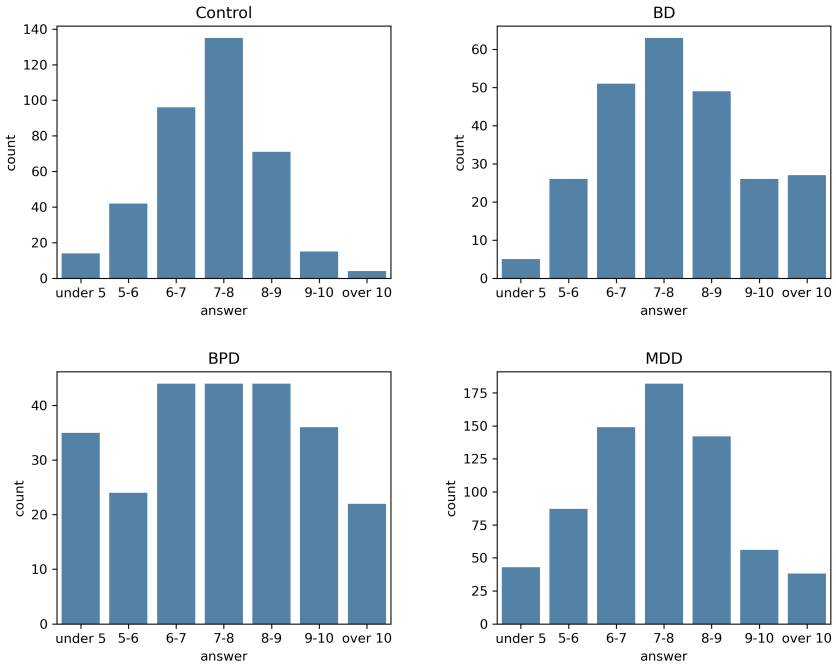

Figure 16: Total sleep time based on survey responses

5 Discussion

5.1 Estimating sleep parameters using the data from the sensors

In the MoMo-Mood study, daily rhythms of healthy controls and patients suffering from different mood disorders were tracked including their sleep behaviors. As the first step in this thesis, some parameters widely used in sleep studies are calculated from the data obtained by the sensors to interpret the sleep-related data. As the sensory data included sleep/wake statuses, it was possible to calculate a set of selected parameters including total sleep time (TST), wakefulness after sleep onset (WASO), and sleep efficiency (SE).

Each sensor had its own method for detecting sleep events. For actigraph, sleep and wake states were calculated by the built-in Actiwatch algorithm using the information tracked, such as physical activity and the amount and duration of the ambient white light illuminance. The states were already included in a status column in the data therefore no specific algorithm was needed. For the bed sensor, the states were determined based on the calibration and the signal strength. Similar to the actigraph, the states were logged in a status column. Lastly, in screen activity data, there was no status column, hence sleep states were detected by calculating the longest inactive period during the day which could refer to nightly sleep. Since this approach did not take the screen status into account, a new method was proposed by defining the sleep period as the period between the latest screen lock event and the next earliest screen unlock event. However, this approach caused some data to be discarded as the lock and unlock events were not available as often as screen-on and screen-off events. The reason for not using screen-on and screen-off to determine the sleep/wake states was that they could be initiated without the user's consent, such as receiving a notification or lack of activity for a short period of time triggering a change in the screen status.

In the analysis, for all subject groups, median TST was the lowest from the screen activity while it was the highest from the data collected by the bed sensor. All sensors combined, the mean TST for the control group was the shortest with 7.03 hours while the MDD group had a mean TST of 8 hours. When compared with the values from the actigraph, the bedtimes from the bed sensor showed more agreement than the ones from the screen activity. Also, the BD group had the highest dispersion, meaning that the differences were more likely to be unique values.

5.2 Predicting the depression severity of patients using the actigraph and bed sensor data

As mentioned earlier in 4.1, previous research claims a bidirectional relationship between sleep disturbance and depression. To explore if such a relationship is present in the MoMo-Mood dataset, an XGBoost classification model is implemented in this work to classify patients with low and high depression levels using their data from the actigraph and bed sensor. The choice of the algorithm was made based on its ability to

produce accurate and interpretable predictions. The model evaluation revealed that the information tracked by actigraph and bed sensor was not sufficient to interpret the effects of the features on understanding the severity of depression given the poor accuracy. However, the model was more accurate in correctly predicting the patients with high levels of depression than the individuals with slighter depression. This information suggests that the model can be used in scenarios in which the depression levels of individuals are checked in a preliminary step to ensure only patients with high levels of depression are likely to be evaluated by the model to form the final decision of receiving treatment.

5.3 Analyzing the perceived quality from the morning questionnaires

As the final analysis in this work, the answers to three sleep-related questions from the morning questionnaire were compared with the sleep parameters calculated earlier from the sensory data. The initial hypothesis was that the perceived sleep quality would be correlated with the sleep quality obtained from the parameters. However, there was no significant correlation found between the answers to all three questions and the parameters TST and SE. The lack of correlation can be associated with the fact that sleep quality does not depend on the sleep parameters individually but rather as a combination of them. WASO was not included in this analysis to avoid multicollinearity since it was already used to calculate SE and was known to have a negative correlation with it.

The perceived TST from the survey answers mostly represented a symmetrical distribution (control, BD, MDD) which is similar to the self-rated sleep durations from Landolt's study [99]. As the survey answers were in categorical order, an exact comparison between TSTs from the sensory data was not conducted.

5.4 Further work

Sleep is a complex process that is affected by various physiological, biochemical, and other internal and external factors. This work only utilizes the sleep-related features from the sensory data in the MoMo-Mood dataset to understand the sleep behaviors of the subjects. The sensory data includes a limited set of factors that might affect an individual's sleep behavior. There are several ideas for further work that could be conducted to build on the findings of this work.

First, the sleep analysis can be extended by gathering more specific information from the participants related to their sleep. A sleep diary including bedtime and wake-up time, the number of naps during the day, caffeine/alcohol consumption, and daily activity level would be a valuable tool to understand the other factors affecting nightly sleep.

Secondly, some information regarding the treatment process of the subjects suffering from mood disorders could help to understand the effect of different types of treatments on sleep quality. This information could be utilized to improve sleep in a particular patient group. It would also be useful to verify the results of this work

as it does not assume any difference in the subjects belonging to the same patient group.

Another direction of future work would focus on the changes in PHQ-9 scores of the subjects to investigate whether they affect sleep behaviors. While this could provide additional information about the relationship between the severity of depression and sleep, it would also lower the number of subjects in the analysis as not all subjects were administered the test multiple times.

Finally, the analysis modules related to sleep that were part of this thesis will be incorporated in Niimpy, an open-source behavioral data analysis toolbox developed by Ikäheimonen et al. [100], to utilize them in studies using other sets of similar data.

6 Summary

This thesis aimed to investigate three main research objectives using the data collected in the MoMo-Mood study. The dataset included information tracked by three different sources; actigraphy, ballistocardiography, and smartphone screen activity. The first objective was to find out if the sleep behaviors of each subject group differs. Sleep parameters of healthy controls and patients suffering from mood disorders (BD, BPD, MDD) are calculated and compared at group levels. These parameters included total sleep time, wakefulness after sleep onset, and sleep efficiency. The second objective was to predict the severity of depression for the subjects in patient groups. Firstly, the scores from the PHQ-9 tests administered during the study were collected and matched with the sensor data from the appropriate time period- up to 2 weeks prior to the test date. After the data preprocessing, the data was split into training and test cohorts using an 80:20 ratio. Then, an XGBoost classifier was trained using the data of the subjects in the training cohort. The model was evaluated using accuracy, recall, and precision. Furthermore, the feature importance of the model was investigated to have a better understanding of the predictions. The poor accuracy of the model suggested that the data from the actigraph and bed sensor was not sufficient for the model to understand the classes. The last objective of this thesis was to compare the perceived sleep quality obtained from the morning questionnaires with the sleep parameters calculated from the actigraph data. There was no significant correlation found between the survey answers and the sleep parameters. The distributions of total sleep time from the survey answers of the three groups were found to be normally distributed.

References

[1] James M. Krueger, Ferenc Obál, and Jidong Fang. "Why we sleep: a theoretical view of sleep function". en. In: *Sleep Medicine Reviews* 3.2 (June 1999), pp. 119–129. ISSN: 10870792. DOI: 10.1016/S1087-0792(99)90019-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S1087079299900199.

[2] Joohee Shim and Seung Wan Kang. "Behavioral Factors Related to Sleep Quality and Duration in Adults". en. In: *Journal of Lifestyle Medicine* 7.1 (Jan. 2017), pp. 18–26. ISSN: 2234-8549, 2288-1557. DOI: 10.15280/jlm.2017.7.1.18. URL: http://www.jlifestylemed.org/journal/DOIx.php?id=10.15280/jlm.2017.7.1.18.

[3] Daniel J. Buysse et al. "The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research". In: *Psychiatry Research* 28.2 (1989), pp. 193–213. ISSN: 0165-1781. DOI: https://doi.org/10.1016/0165-1781(89)90047-4. URL: https://www.sciencedirect.com/science/article/pii/0165178189900474.

[4] Andrew D. Krystal and Jack D. Edinger. "Measuring sleep quality". en. In: *Sleep Medicine* 9 (Sept. 2008), S10–S17. ISSN: 13899457. DOI: 10.1016/S1389-9457(08)70011-X. URL: https://linkinghub.elsevier.com/retrieve/pii/S138994570870011X.

[5] Deirdre O'Donnell et al. "Comparison of subjective and objective assessments of sleep in healthy older subjects without sleep complaints". en. In: *Journal of Sleep Research* 18.2 (June 2009), pp. 254–263. ISSN: 09621105, 13652869. DOI: 10.1111/j.1365-2869.2008.00719.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2869.2008.00719.x.

[6] Sandeep Sekhon and Vikas Gupta. "Mood Disorder". eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2022. URL: http://www.ncbi.nlm.nih.gov/books/NBK558911/.

[7] R. M. Benca et al. "Sleep and mood disorders". eng. In: *Sleep Medicine Reviews* 1.1 (Nov. 1997), pp. 45–56. ISSN: 1087-0792. DOI: 10.1016/s1087-0792(97)90005-8.

[8] Kathleen R. Merikangas et al. "Prevalence and Correlates of Bipolar Spectrum Disorder in the World Mental Health Survey Initiative". In: *Archives of general psychiatry* 68.3 (Mar. 2011), pp. 241–251. ISSN: 0003-990X. DOI: 10.1001/archgenpsychiatry.2011.12. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3486639/.

[9] Gordon Parker. "Is borderline personality disorder a mood disorder?" en. In: *British Journal of Psychiatry* 204.4 (Apr. 2014), pp. 252–253. ISSN: 0007-1250, 1472-1465. DOI: 10.1192/bjp.bp.113.136580. URL: https://www.cambridge.org/core/product/identifier/S0007125000275983/type/journal_article.

[10] Mary C. Zanarini et al. "A Screening Measure for BPD: The McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD)". en. In: *Journal of Personality Disorders* 17.6 (Dec. 2003), pp. 568–573. ISSN: 0885-579X. DOI: 10.1521/pedi.17.6.568.25355. URL: http://guilfordjournals.com/doi/10.1521/pedi.17.6.568.25355.

[11] Miguel Marino et al. "Measuring Sleep: Accuracy, Sensitivity, and Specificity of Wrist Actigraphy Compared to Polysomnography". en. In: *Sleep* 36.11 (Nov. 2013), pp. 1747–1755. ISSN: 0161-8105, 1550-9109. DOI: 10.5665/sleep.3142. URL: https://academic.oup.com/sleep/article/36/11/1747/2558963.

[12] Jessica Vensel Rundo and Ralph Downey. "Polysomnography". eng. In: *Handbook of Clinical Neurology* 160 (2019), pp. 381–392. ISSN: 0072-9752. DOI: 10.1016/B978-0-444-64032-1.00025-4.

[13] Alexandru Corlateanu et al. "To sleep, or not to sleep – that is the question, for polysomnography". In: *Breathe* 13.2 (June 2017), pp. 137–140. ISSN: 1810-6838. DOI: 10.1183/20734735.007717. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5467660/.

[14] Jennifer L. Martin and Alex D. Hakim. "Wrist actigraphy". eng. In: *Chest* 139.6 (June 2011), pp. 1514–1527. ISSN: 1931-3543. DOI: 10.1378/chest.10-1872.

[15] Talayeh Aledavood et al. "Smartphone-Based Tracking of Sleep in Depression, Anxiety, and Psychotic Disorders". en. In: *Current Psychiatry Reports* 21.7 (June 2019), p. 49. ISSN: 1535-1645. DOI: 10.1007/s11920-019-1043-y. URL: https://doi.org/10.1007/s11920-019-1043-y.

[16] Lee D. Mulligan et al. "High resolution examination of the role of sleep disturbance in predicting functioning and psychotic symptoms in schizophrenia: A novel experience sampling study." In: *Journal of Abnormal Psychology* 125.6 (2016), pp. 788–797. DOI: 10.1037/abn0000180.

[17] Talayeh Aledavood, Sune Lehmann, and Jari Saramäki. "Digital daily cycles of individuals". In: *Frontiers in Physics* 3 (Oct. 2015). ISSN: 2296-424X. DOI: 10.3389/fphy.2015.00073. URL: http://journal.frontiersin.org/Article/10.3389/fphy.2015.00073/abstract.

[18] Talayeh Aledavood et al. "Quantifying daily rhythms with non-negative matrix factorization applied to mobile phone data". en. In: *Scientific Reports* 12.1 (Apr. 2022), p. 5544. ISSN: 2045-2322. DOI: 10.1038/s41598-022-09273-y. URL: https://www.nature.com/articles/s41598-022-09273-y.

[19] Iqbal H. Sarker. "Machine Learning: Algorithms, Real-World Applications and Research Directions". en. In: *SN Computer Science* 2.3 (May 2021), p. 160. ISSN: 2662-995X, 2661-8907. DOI: 10.1007/s42979-021-00592-x. URL: https://link.springer.com/10.1007/s42979-021-00592-x.

[20] Aria Khademi et al. “personalized sleep parameters estimation from actigraphy: A machine learning approach”. In: *Nature and Science of Sleep* Volume 11 (2019), pp. 387–399. DOI: 10.2147/nss.s220716.

[21] Yasser El-Manzalawy, Orfeu Buxton, and Vasant Honavar. “Sleep/wake state prediction and sleep parameter estimation using unsupervised classification via clustering”. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Kansas City, MO: IEEE, Nov. 2017, pp. 718–723. ISBN: 9781509030507. DOI: 10.1109/BIBM.2017.8217742. URL: http://ieeexplore.ieee.org/document/8217742/.

[22] Sharon A. Keenan. “Chapter 3 An overview of polysomnography”. en. In: *Handbook of Clinical Neurophysiology*. Ed. by Christian Guilleminault. Vol. 6. Handbook of Clinical Neurophysiology. Elsevier, Jan. 2005, pp. 33–50. DOI: 10.1016/S1567-4231(09)70028-0. URL: https://www.sciencedirect.com/science/article/pii/S1567423109700280.

[23] Franc Paul, Michael Schredl, and Georg W Alpers. “Nightmares affect the experience of sleep quality but not sleep architecture: an ambulatory polysomnographic study”. en. In: *Borderline Personality Disorder and Emotion Dysregulation* 2.1 (Dec. 2015), p. 3. ISSN: 2051-6673. DOI: 10.1186/s40479-014-0023-4. URL: http://bpded.biomedcentral.com/articles/10.1186/s40479-014-0023-4.

[24] L. Giovangrandi et al. “Ballistocardiography – A Method Worth Revisiting”. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2011). DOI: 10.1109/iembs.2011.6091062.

[25] Gih Sung Chung et al. “Wakefulness estimation only using ballistocardiogram: Nonintrusive method for sleep monitoring”. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. 2010, pp. 2459–2462. DOI: 10.1109/IEMBS.2010.5626544.

[26] David C. Mack et al. “Sleep assessment using a passive ballistocardiography-based system: Preliminary validation”. In: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2009, pp. 4319–4322. DOI: 10.1109/IEMBS.2009.5333805.

[27] Guillermina Guerrero et al. “Detection of sleep-disordered breating with Pressure Bed Sensor”. eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2013 (2013), pp. 1342–1345. ISSN: 2694-0604. DOI: 10.1109/EMBC.2013.6609757.

[28] Avi Sadeh and Christine Acebo. “The role of actigraphy in sleep medicine”. en. In: *Sleep Medicine Reviews* 6.2 (May 2002), pp. 113–124. ISSN: 10870792. DOI: 10.1053/smrv.2001.0182. URL: https://linkinghub.elsevier.com/retrieve/pii/S1087079201901820.

[29] J. Krishna and S. Mashaqi. "Actigraphy". en. In: *Encyclopedia of the Neurological Sciences (Second Edition)*. Ed. by Michael J. Aminoff and Robert B. Daroff. Oxford: Academic Press, Jan. 2014, pp. 36–40. ISBN: 9780123851581. DOI: 10.1016/B978-0-12-385157-4.00547-9. URL: https://www.sciencedirect.com/science/article/pii/B9780123851574005479.

[30] Børge Sivertsen et al. "A Comparison of Actigraphy and Polysomnography in Older Adults Treated for Chronic Primary Insomnia". en. In: *Sleep* 29.10 (Oct. 2006), pp. 1353–1358. ISSN: 1550-9109, 0161-8105. DOI: 10.1093/sleep/29.10.1353. URL: https://academic.oup.com/sleep/article-lookup/doi/10.1093/sleep/29.10.1353.

[31] Samantha Conley et al. "Agreement between actigraphic and polysomnographic measures of sleep in adults with and without chronic conditions: A systematic review and meta-analysis". eng. In: *Sleep Medicine Reviews* 46 (Aug. 2019), pp. 151–160. ISSN: 1532-2955. DOI: 10.1016/j.smrv.2019.05.001.

[32] Sumit Majumder and M. Jamal Deen. "Smartphone Sensors for Health Monitoring and Diagnosis". In: *Sensors (Basel, Switzerland)* 19.9 (May 2019), p. 2164. ISSN: 1424-8220. DOI: 10.3390/s19092164. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6539461/.

[33] Clarence Baxter et al. "Assessment of Mobile Health Apps Using Built-In Smartphone Sensors for Diagnosis and Treatment: Systematic Survey of Apps Listed in International Curated Health App Libraries". In: *JMIR mHealth and uHealth* 8.2 (Feb. 2020), e16741. ISSN: 2291-5222. DOI: 10.2196/16741. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7055743/.

[34] Talayeh Aledavood, Sune Lehmann, and Jari Saramäki. "Social network differences of chronotypes identified from mobile phone data". en. In: *EPJ Data Science* 7.1 (Dec. 2018), p. 46. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-018-0174-4. URL: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0174-4.

[35] Navneet Bains and Sara Abdijadid. *Major depressive disorder*. 2022. URL: https://pubmed.ncbi.nlm.nih.gov/32644504/.

[36] Maurizio Fava and Kenneth S. Kendler. "Major Depressive Disorder". English. In: *Neuron* 28.2 (Nov. 2000), pp. 335–341. ISSN: 0896-6273. DOI: 10.1016/S0896-6273(00)00112-4. URL: https://www.cell.com/neuron/abstract/S0896-6273(00)00112-4.

[37] Luis Gutiérrez-Rojas et al. "Prevalence and correlates of major depressive disorder: a systematic review". eng. In: *Revista Brasileira De Psiquiatria (Sao Paulo, Brazil: 1999)* 42.6 (2020), pp. 657–672. ISSN: 1809-452X. DOI: 10.1590/1516-4446-2020-0650.

[38] Iria Grande et al. "Bipolar disorder". eng. In: *Lancet (London, England)* 387.10027 (Apr. 2016), pp. 1561–1572. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(15)00241-X.

[39] David T. Plante and John W. Winkelman. "Sleep Disturbance in Bipolar Disorder: Therapeutic Implications". In: *American Journal of Psychiatry* 165.7 (July 2008), pp. 830–843. ISSN: 0002-953X. DOI: 10.1176/appi.ajp.2008.08010077. URL: https://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.2008.08010077.

[40] B. Barbini et al. "Sleep loss, a possible factor in augmenting manic episode". eng. In: *Psychiatry Research* 65.2 (Nov. 1996), pp. 121–125. ISSN: 0165-1781. DOI: 10.1016/s0165-1781(96)02909-5.

[41] Carol A. Perlman, Sheri L. Johnson, and Thomas A. Mellman. "The prospective impact of sleep duration on depression and mania". eng. In: *Bipolar Disorders* 8.3 (June 2006), pp. 271–274. ISSN: 1398-5647. DOI: 10.1111/j.1399-5618.2006.00330.x.

[42] Klaus Lieb et al. "Borderline personality disorder". eng. In: *Lancet (London, England)* 364.9432 (Aug. 2004), pp. 453–461. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(04)16770-6.

[43] Jennifer Chapman, Radia T. Jamil, and Carl Fleisher. "Borderline Personality Disorder". eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2022. URL: http://www.ncbi.nlm.nih.gov/books/NBK430883/.

[44] Sina Hafizi. "Sleep and borderline personality disorder: A review". en. In: *Asian Journal of Psychiatry*. This issue includes a special section on Spirituality and Psychiatry 6.6 (Dec. 2013), pp. 452–459. ISSN: 1876-2018. DOI: 10.1016/j.ajp.2013.06.016. URL: https://www.sciencedirect.com/science/article/pii/S1876201813001883.

[45] Edward A. Selby. "Chronic Sleep Disturbances and Borderline Personality Disorder Symptoms". In: *Journal of consulting and clinical psychology* 81.5 (Oct. 2013), pp. 941–947. ISSN: 0022-006X. DOI: 10.1037/a0033201. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4129646/.

[46] K Kroenke, R L Spitzer, and J B Williams. "The PHQ-9: validity of a brief depression severity measure." In: *Journal of general internal medicine* 16.9 (Sept. 2001), pp. 606–13.

[47] C. Beard et al. "Validation of the PHQ-9 in a psychiatric sample". In: *Journal of Affective Disorders* 193 (2016), pp. 267–273. ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2015.12.075. URL: https://www.sciencedirect.com/science/article/pii/S0165032715310272.

[48] Yue Sun et al. "The reliability and validity of PHQ-9 in patients with major depressive disorder in psychiatric hospital". In: *BMC Psychiatry* 20.1 (2020). DOI: 10.1186/s12888-020-02885-6.

[49] Kurt Kroenke and Robert L Spitzer. "The PHQ-9: a new depression diagnostic and severity measure". In: *Psychiatric Annals* 32.9 (2002), pp. 509–515.

[50] *Genotype versus phenotype - understanding evolution*. Oct. 2021. URL: https://evolution.berkeley.edu/genotype-versus-phenotype/.

[51] *Research areas*. Nov. 2021. URL: https://www.hsph.harvard.edu/onnela-lab/research/.

[52] *Digital Phenotyping: Big Data Society: SAGE journals*. URL: https://journals.sagepub.com/page/bds/collections/digitalphenotyping.

[53] John Torous et al. "New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research". eng. In: *JMIR mental health* 3.2 (May 2016), e16. ISSN: 2368-7959. DOI: 10.2196/mental.5165.

[54] J. Torous, J.-P. Onnela, and M. Keshavan. "New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices". en. In: *Translational Psychiatry* 7.3 (Mar. 2017), e1053–e1053. ISSN: 2158-3188. DOI: 10.1038/tp.2017.25. URL: https://www.nature.com/articles/tp201725.

[55] Jennifer Melcher, Ryan Hays, and John Torous. "Digital phenotyping for mental health of college students: a clinical review". In: *Evidence-Based Mental Health* 23.4 (2020), pp. 161–166. ISSN: 1362-0347. DOI: 10.1136/ebmental-2020-300180. eprint: https://ebmh.bmj.com/content/23/4/161.full.pdf. URL: https://ebmh.bmj.com/content/23/4/161.

[56] Ana María Triana et al. "Mobile Monitoring of Mood (MoMo-Mood) Pilot: A Longitudinal, Multi-Sensor Digital Phenotyping Study of Patients with Major Depressive Disorder and Healthy Controls". en. In: (Dec. 2020). DOI: 10.1101/2020.11.02.20222919. URL: https://www.medrxiv.org/content/10.1101/2020.11.02.20222919v2.

[57] John Zulueta et al. "Predicting Mood Disturbance Severity with Mobile Phone Keystroke Metadata: A BiAffect Digital Phenotyping Study". EN. In: *Journal of Medical Internet Research* 20.7 (July 2018), e9775. DOI: 10.2196/jmir.9775. URL: https://www.jmir.org/2018/7/e241.

[58] Yuuki Tazawa et al. "Actigraphy for evaluation of mood disorders: A systematic review and meta-analysis". eng. In: *Journal of Affective Disorders* 253 (June 2019), pp. 257–269. ISSN: 1573-2517. DOI: 10.1016/j.jad.2019.04.087.

[59] Tuomas Alakörkkö. "Monitoring daily behavioral patterns using mobile phone sensors and ballistocardiography for detecting mental health problems". English. Master's thesis. Aalto University. School of Science, 2016, pp. 69+7. URL: http://urn.fi/URN:NBN:fi:aalto-201612085875.

[60] Anna Hakala. "Classification of patients with depression and healthy controls based on behavioural patterns acquired from smartphone sensor data". English. Master's thesis. Aalto University. School of Science, 2021, pp. 84+7. URL: http://urn.fi/URN:NBN:fi:aalto-202108298570.

[61] Amirmohammad Ziaei Bideh. "Exploring Behavioral Patterns of Patients with Mental Disorders Using the MoMo-Mood Dataset". English. Master's thesis. Aalto University. School of Science, 2022, pp. 56+1. URL: http://urn.fi/URN:NBN:fi:aalto-202208285082.

[62] Talayeh Aledavood et al. "Data collection for mental health studies through digital platforms: requirements and design of a prototype". In: *JMIR research protocols* 6.6 (2017), e6919.

[63] *Actiwatch 2*. URL: https://bmedical.com.au/product/actiwatch-2-minimitter-philips/.

[64] Xuan Kai Lee et al. "Validation of a Consumer Sleep Wearable Device With Actigraphy and Polysomnography in Adolescents Across Sleep Opportunity Manipulations". In: *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine* 15.9 (Sept. 2019), pp. 1337–1346. ISSN: 1550-9389. DOI: 10.5664/jcsm.7932. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6760396/.

[65] Wilfred R. Pigeon et al. "Validation of the Sleep-Wake Scoring of a New Wrist-Worn Sleep Monitoring Device". In: *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine* 14.6 (June 2018), pp. 1057–1062. ISSN: 1550-9389. DOI: 10.5664/jcsm.7180. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5991960/.

[66] Julia Kahlhöfer et al. "Relationship between actigraphy-assessed sleep quality and fat mass in college students". eng. In: *Obesity (Silver Spring, Md.)* 24.2 (Feb. 2016), pp. 335–341. ISSN: 1930-739X. DOI: 10.1002/oby.21326.

[67] Maya J Lambiase et al. "Utility of Actiwatch sleep monitor to assess waking movement behavior in older women". In: *Medicine and science in sports and exercise* 46.12 (Dec. 2014), pp. 2301–2307. ISSN: 0195-9131. DOI: 10.1249/MSS.0000000000000361. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4211988/.

[68] Bernardo Silva and Rui Neto Marinheiro. "Non-invasive monitoring with Ballistocardiographic sensors for sleep management". In: *2021 Telecoms Conference (ConfTELE)*. 2021, pp. 1–6. DOI: 10.1109/ConfTELE50222.2021.9435481.

[69] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. "AWARE: Mobile Context Instrumentation Framework". In: *Frontiers in ICT* 2 (2015). ISSN: 2297-198X. URL: https://www.frontiersin.org/articles/10.3389/fict.2015.00006.

[70] Ronald C. Kessler et al. "The World Health Organization adult ADHD self-report scale (ASRS): a short screening scale for use in the general population". en. In: *Psychological Medicine* 35.2 (Feb. 2005), pp. 245–256. ISSN: 0033-2917, 1469-8978. DOI: 10.1017/S0033291704002892. URL: https://www.cambridge.org/core/product/identifier/S0033291704002892/type/journal_article.

[71] Deepak Shrivastava et al. “How to interpret the results of a sleep study”. en. In: *Journal of Community Hospital Internal Medicine Perspectives* 4.5 (Jan. 2014), p. 24983. ISSN: 2000-9666. DOI: 10.3402/jchimp.v4.24983. URL: https://www.tandfonline.com/doi/full/10.3402/jchimp.v4.24983.

[72] Natalie D. Dautovich et al. “A systematic review of the amount and timing of light in association with objective and subjective sleep outcomes in community-dwelling adults”. In: *Sleep Health* 5.1 (2019), pp. 31–48. DOI: 10.1016/j.sleh.2018.09.006.

[73] Mary Amanda Dew et al. “Healthy Older Adults’ Sleep Predicts All-Cause Mortality at 4 to 19 Years of Follow-Up:” en. In: *Psychosomatic Medicine* 65.1 (Jan. 2003), pp. 63–73. ISSN: 0033-3174. DOI: 10.1097/01.PSY.0000039756.23250.7C. URL: http://journals.lww.com/00006842-200301000-00008.

[74] Adam P. Spira et al. “Anxiety Symptoms and Objectively Measured Sleep Quality in Older Women”. en. In: *The American Journal of Geriatric Psychiatry* 17.2 (Feb. 2009), pp. 136–143. ISSN: 10647481. DOI: 10.1097/JGP.0b013e3181871345. URL: https://linkinghub.elsevier.com/retrieve/pii/S1064748112607337.

[75] Torbjörn Åkerstedt et al. “The meaning of good sleep: a longitudinal study of polysomnography and subjective sleep quality”. en. In: *Journal of Sleep Research* 3.3 (Sept. 1994), pp. 152–158. ISSN: 09621105, 13652869. DOI: 10.1111/j.1365-2869.1994.tb00122.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2869.1994.tb00122.x.

[76] Aron Halfin. “Depression: the benefits of early and appropriate treatment.” In: *The American journal of managed care* 13.4 Suppl (Nov. 2007), S92–7.

[77] Anthony J. Myles et al. “An introduction to decision tree modeling”. en. In: *Journal of Chemometrics* 18.6 (June 2004), pp. 275–285. ISSN: 0886-9383, 1099-128X. DOI: 10.1002/cem.873. URL: https://onlinelibrary.wiley.com/doi/10.1002/cem.873.

[78] Y. Y. Song and Y. Lu. “Decision tree methods: applications for classification and prediction”. In: *Shanghai Arch Psychiatry* 27.2 (Apr. 2015), pp. 130–135.

[79] J. R. Quinlan. “Learning Decision Tree Classifiers”. In: *ACM Comput. Surv.* 28.1 (Mar. 1996), pp. 71–72. ISSN: 0360-0300. DOI: 10.1145/234313.234346. URL: https://doi.org/10.1145/234313.234346.

[80] Leo Breiman. “Bagging predictors”. en. In: *Machine Learning* 24.2 (Aug. 1996), pp. 123–140. ISSN: 1573-0565. DOI: 10.1007/BF00058655. URL: https://doi.org/10.1007/BF00058655.

[81] Yann Coadou. “Boosted decision trees”. In: arXiv:2206.09645 [hep-ex, physics:physics]. Mar. 2022, pp. 9–58. URL: http://arxiv.org/abs/2206.09645.

[82] Gérard Biau and Erwan Scornet. "A random forest guided tour". en. In: *TEST* 25.2 (June 2016), pp. 197–227. ISSN: 1133-0686, 1863-8260. DOI: 10.1007/s11749-016-0481-7. URL: http://link.springer.com/10.1007/s11749-016-0481-7.

[83] Thomas G. Dietterich. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization". In: *Mach. Learn.* 40.2 (Aug. 2000), pp. 139–157. ISSN: 0885-6125. DOI: 10.1023/A:1007607513941. URL: https://doi.org/10.1023/A:1007607513941.

[84] *Visualization of a decision tree*. Datacamp, Dec. 2018. URL: https://www.datacamp.com/tutorial/decision-tree-classification-python.

[85] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". en. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, Aug. 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://dl.acm.org/doi/10.1145/2939672.2939785.

[86] Kartik Budholiya, Shailendra Kumar Shrivastava, and Vivek Sharma. "An optimized XGBoost based diagnostic system for effective prediction of heart disease". en. In: *Journal of King Saud University - Computer and Information Sciences* 34.7 (July 2022), pp. 4514–4523. ISSN: 1319-1578. DOI: 10.1016/j.jksuci.2020.10.013. URL: https://www.sciencedirect.com/science/article/pii/S1319157820304936.

[87] Amita Sharma and Willem J. M. I. Verbeke. "Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081)". In: *Frontiers in Big Data* 3 (2020). ISSN: 2624-909X. URL: https://www.frontiersin.org/articles/10.3389/fdata.2020.00015.

[88] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. "A review: Data pre-processing and data augmentation techniques". In: *Global Transitions Proceedings* 3.1 (2022). International Conference on Intelligent Engineering Approach(ICIEA-2022), pp. 91–99. ISSN: 2666-285X. DOI: https://doi.org/10.1016/j.gltp.2022.04.020. URL: https://www.sciencedirect.com/science/article/pii/S2666285X22000565.

[89] Salvador García, Julián Luengo, and Francisco Herrera. *Data Preprocessing in Data Mining*. Vol. 72. Intelligent Systems Reference Library. Cham: Springer International Publishing, 2015. ISBN: 9783319102467. DOI: 10.1007/978-3-319-10247-4. URL: http://link.springer.com/10.1007/978-3-319-10247-4.

[90] Cheng Fan et al. "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data". In: *Frontiers in Energy Research* 9 (2021). ISSN: 2296-598X. DOI: 10.3389/fenrg.2021.652801. URL: https://www.frontiersin.org/articles/10.3389/fenrg.2021.652801.

[91] Haibo He and Edwardo A Garcia. "Learning from imbalanced data". In: *Knowledge and Data Engineering, IEEE Transactions on* 21.9 (2009), pp. 1263–1284.

[92] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. "Machine Learning Interpretability: A Survey on Methods and Metrics". In: *Electronics* 8.8 (2019). ISSN: 2079-9292. DOI: 10.3390/electronics8080832. URL: https://www.mdpi.com/2079-9292/8/8/832.

[93] Inna Kolyshkina and Simeon Simoff. "Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach". In: *Frontiers in Big Data* 4 (2021). ISSN: 2624-909X. URL: https://www.frontiersin.org/articles/10.3389/fdata.2021.660206.

[94] Bryce Goodman and Seth Flaxman. "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"". In: *AI Magazine* 38.3 (Oct. 2017), pp. 50–57. DOI: 10.1609/aimag.v38i3.2741. URL: https://doi.org/10.1609%2Faimag.v38i3.2741.

[95] Radwa Elshawi, Mouaz H. Al-Mallah, and Sherif Sakr. "On the interpretability of machine learning-based model for predicting hypertension". en. In: *BMC Medical Informatics and Decision Making* 19.1 (Dec. 2019), p. 146. ISSN: 1472-6947. DOI: 10.1186/s12911-019-0874-0. URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0874-0.

[96] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously". In: (2018). DOI: 10.48550/ARXIV.1801.01489. URL: https://arxiv.org/abs/1801.01489.

[97] *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge [Cambridgeshire]; New York: Cambridge University Press, 1988. ISBN: 9780521361774.

[98] Andrew Winokur. "The Relationship Between Sleep Disturbances and Psychiatric Disorders". en. In: *Psychiatric Clinics of North America* 38.4 (Dec. 2015), pp. 603–614. ISSN: 0193953X. DOI: 10.1016/j.psc.2015.07.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0193953X15000763.

[99] Hans-Peter Landolt. "Genotype-Dependent Differences in Sleep, Vigilance, and Response to Stimulants". en. In: *Current Pharmaceutical Design* 14.32 (), pp. 3396–3407. URL: https://www.eurekaselect.com/article/13023.

[100] A Ikäheimonen et al. "Niimpy: a toolbox for behavioral data analysis". In: *arXiv preprint arXiv:2212.02192* (2022).