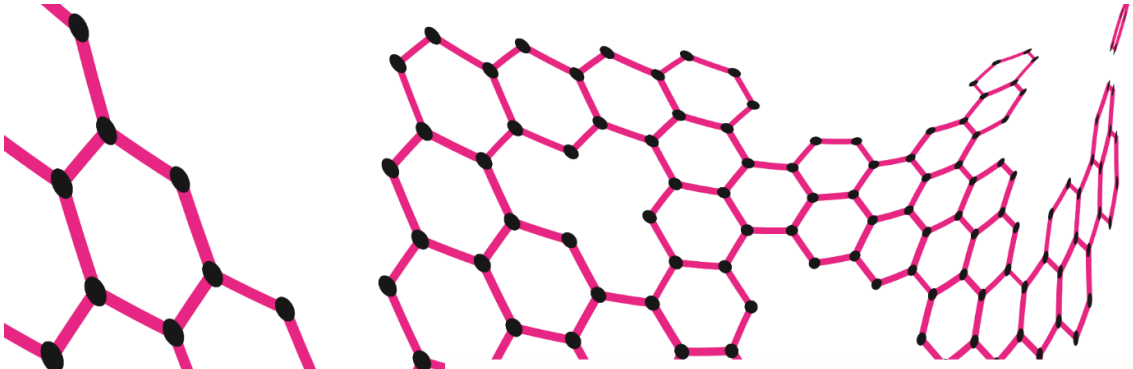


BSc Thesis - Biomedical Technology

Predicting Depression and Sleep Disturbances from Circadian Rhythm Biomarkers with Machine Learning

Nikki Overmars

July 7th, 2025

Committee members:
dr. A. John
F. Oumar, MSc
dr. J. Piano Simoes

Biomedical Signals and Systems
Faculty EEMCS

UNIVERSITY OF TWENTE.

Abstract

The circadian rhythm, the 24-hour biological clock, plays a role in various mental health disorders, for example, depression and anxiety disorders. Disruptions within this rhythm increase the risk of developing one of these disorders. This study investigated the association between circadian rhythm disruptions and depression among young adults (20-45 years old) in the United States, while also validating those disruptions with sleep disturbances, using data from the 2011-2012 NHANES cycle. Objective circadian rhythm features were derived from 4-7 days of accelerometer data, with a daily wear time of more than 16 hours, collected from 392 participants. These features included three non-parametric metrics: the Relative Amplitude (RA), the Intradaily Variability (IV), and the Interdaily Stability (IS), as well as three parametric cosine features: the MESOR (Midline Estimated Statistic Of Rhythm), the amplitude, and the acrophase. Participants were classified as experiencing heightened depressive symptoms (n = 50) or experiencing sleep disturbances (n = 120) via self-reported questionnaires. Multiple classification models (random forest, logistic regression, and k-nearest neighbors) were used to predict depression and sleep disturbances using several preprocessing options, including different oversampling ratios, covariate in-/exclusion, and different interactions between the predictors. The models were compared on F1-score, accuracy, specificity, precision, recall, and the Area Under the Curve (AUC) value. Model calibration did not enhance the reliability of the predicted probabilities. The best-performing model predicting depression was a covariate-inclusive k-nearest neighbors model (F1-score = 0.364, AUC = 0.81), and for sleep disturbances, it was a covariate-inclusive logistic regression model (F1-score = 0.431, AUC = 0.54). Removing equivocal zones improved the F1-scores for both models (depression = 0.471, sleep disturbances = 0.466). In the sensitivity cohort, the best-performing model predicting sleep disturbances achieved a higher F1-score and AUC, especially after the removal of equivocal zones (F1-score = 0.581, AUC = 0.60). The sensitivity analysis showed no impact on the models predicting depression. Plots of variable importance in both cohorts showed that the circadian rhythm features contributed minimally to both outcome predictions. No significant differences were found in circadian rhythm features between the depression group and the healthy controls; thus, no clear association was found between circadian rhythm disruptions and depression. Significant differences were found in amplitude and IV between the sleep disturbances group and the healthy controls, with the MESOR also showing significance in the sensitivity cohort. However, no other clear association was found between circadian rhythm disruptions and sleep disturbances. Limitations include a small sample size, class imbalances (1:9.6 for depression and 1:1.9 for sleep disturbances), and reliance on self-reported questionnaire data for the outcomes, which may have affected model performance and generalizability. Future studies should use larger longitudinal cohorts with clinical mental health diagnoses, use a continuous classification for depression, include supplementary sleep diaries or salivary melatonin assessments, validate findings on an external dataset, and expand modeling by including more predictors, algorithms, oversampling options, and predictor interactions.

Index Terms

Actigraphy, circadian rhythm, depression, machine learning, NHANES, sleep disturbances

CONTENTS

I. Introduction

A. Context

Depression is a common and increasingly prevalent mental health disorder, with rates among adults in the U.S. rising from 7.3% in 2015 to 9.2% in 2020 [1]. Furthermore, sleep disturbances are also frequently reported among U.S. adults, with 14.5% reporting difficulty falling asleep in 2020 [2]. Since disruptions in the circadian rhythm can negatively affect sleep quality and are associated with mood disorders such as depression, analyzing circadian rhythms may contribute to the early detection and prevention of sleep disturbances and mood disorders [3, 4].

1) Definition of circadian rhythm

The circadian rhythm is a biological rhythm of approximately 24 hours [5]. This internal clock in mammals contributes to physiological homeostasis and a healthy lifespan if properly adjusted [6]. Usually, the circadian rhythm is automatic and regulated by an internal clock in the suprachiasmatic nucleus (SCN), located in the brain's hypothalamus. In the brain, the SCN mainly influences the sleep-wake rhythm, hormones, digestion, immune responses, body temperature, and cognitive and physical performance [7]. The signals from the SCN are transmitted to the rest of the body through nerves, hormones, and metabolic processes, which the body then uses to carry out biological processes with a 24-hour rhythm [8]. The circadian rhythm is established in childhood and changes with age. The changes include a dampening, fragmentation, and phase shifting of the rhythm, resulting in poor sleep quality in older adults [9].

The circadian rhythm is strongly connected to day and night through light [5]. Light enters the brain through the eyes and influences the signals the SCN sends to the body. Coordination between the internal (circadian) clock and the external time is also synchronized by this light-dark cycle [8]. Light affects mood, alertness, and sleep, mainly by suppressing melatonin production in the brain, which causes an increase in both sleep onset latency and alertness during the night. Artificial light has this same effect on the brain [10].

2) Influence of the circadian rhythm on physical and mental health

Disruptions within the circadian rhythm correlate with disease development, including infections and both mental and physical diseases [8]. Diseases influenced by circadian rhythms include diabetes, cardiovascular diseases, and cancer [7]. Bone, muscle, and autoimmune diseases also occur more frequently with a disrupted circadian rhythm. Furthermore, recent studies suggest that circadian rhythm disruptions may increase the risk of neurodegenerative diseases such as Alzheimer's and Parkinson's, and that a chronic disruption is associated with a higher risk of premature mortality. Disruptions of the circadian rhythm have also been found to affect biological aging processes [6, 7]. This acceleration in aging is associated with an increased risk of several diseases, such as cardiovascular disease and cancer.

In addition to the aforementioned effects on physical health, disruptions within the circadian rhythm are also associated with mental health, including depression, bipolar disorder, anxiety, seasonal affective disorder, and the overall quality of life [8]. This is mainly due to changes in rest-activity cycles, food intake time, diet composition, and low daytime exposure. Furthermore, a study that linked circadian misalignment to depression in young adults shows that having long-term circadian rhythm disruptions is associated with mental health problems, including depressive symptoms, emotional disturbance, and suicidal thoughts [11]. These mood disorders are associated with sleep problems, poor sleep quality, and disruptions in the circadian rhythm, which consist of a mismatch between endogenous rhythm and the environment. This circadian phase delay leads to a 20-times increased risk of developing depression. The same study shows that people with extreme variants of circadian phenotypes (morningness/eveningness) often have symptoms of anxiety and/or depression.

Disruptions in sleep and circadian rhythms are characteristic of mood disorders and mental well-being [3]. Depression is associated with reduced daytime activity, and mood disorders are associated with activity during rest periods. Rest-activity rhythms also tend to differ between younger and older adults. Furthermore, nighttime light exposure and the reduction of light intake during the day influence the development of depression, as light affects the circadian rhythm [12]. It has been shown that higher daylight intensity offers protection against depression. Finally, poor circadian rhythmicity is common in depressive disorders, including melancholic depression [13]. Circadian disruptions are likely pathophysiological factors for some mood disorders, highlighting the importance of circadian rhythms on people's physical and mental health.

3) Causes of circadian rhythm disruptions

Irregular exposure to light can affect the circadian rhythm [8]. However, light is not the only factor that can influence it. Stress, diet, physical activity, temperature, sleeping habits, and environmental changes also affect the circadian rhythm [5, 7]. After industrialization, these factors have played an increasingly important role, since people have adopted a new lifestyle that differs from pre-industrial lifestyles [8]. This new lifestyle is characterized by a more fat-rich diet, a sedentary life indoors, an increased exposure to artificial light, and frequent travel that results in jet lag (which negatively affects health). Many people also work shifts, which means they are often awake while it is dark outside, resulting in irregular sleep patterns. This new lifestyle considerably impacts the normal circadian rhythm, especially the rest-activity rhythm [6]. It is essential to measure the circadian rhythm to prevent disruptions from occurring, which are associated with several diseases, as mentioned earlier. Preventing these disruptions also prevents the development and progression of those diseases.

4) Monitoring of the circadian rhythm

There are several ways in which the circadian rhythm can be monitored and tracked. One way is to use sleep logs or sleep diaries [14]. In those logs/diaries, a person writes down when they went to bed, fell asleep, and woke up, which gives a picture of rest-activity rhythms. The International Classification of Sleep Disorders (ICSD) uses this as diagnostic criteria for specific Circadian Rhythm Sleep-Wake Disorders (CRSWD), including delayed sleep-wake phase disorder and irregular sleep-wake disorder. Another way to track circadian rhythms is via circadian chronotypes, which are determined using questionnaires. This chronotype depends on when someone is doing daily activities and when they are sleeping.

For many years, measuring the core body temperature (CBT) was a common way to look at circadian rhythms [14]. The CBT reaches its lowest point during the last sleep period, in most people around 3-4 in the morning. However, this period can occur earlier or later in people with sleep-wake phase disorders. The core body temperature can, for example, be measured on the skin, in the mouth, under the armpit, and in the rectum [15]. However, measuring the CBT has become less common, as it can be affected by, for example, food and activity [14]. Furthermore, the CBT is typically measured rectally or via skin sensors for circadian rhythm assessments. The former is considered invasive, while the latter is affected by ambient temperatures, activity levels, and the placement of the sensors that measure the temperature. Since the discovery of melatonin, measuring melatonin levels in blood plasma has become the gold standard for assessing circadian rhythms [14]. Melatonin production, controlled by the SCN, can be used as a marker for timings within the circadian rhythm, since melatonin production depends on light. However, this measurement requires periodic invasive blood sampling, which involves inserting a catheter during the blood plasma collection. Moreover, the samples must be processed quickly and stored in freezers, and the analysis in a laboratory takes at least 7-10 days. Therefore, this measurement cannot be performed at home. Alternatively, melatonin levels can also be measured in saliva, a non-invasive approach that accurately mirrors blood melatonin levels [16].

The aforementioned methods of tracking circadian rhythms can be inconvenient, unreliable, and sometimes even invasive [17]. This has led researchers to explore alternative approaches to track circadian rhythms more reliably. One of those ways is utilizing actigraphy, the measurement of activity. There are digital devices (wearables) on the market that consumers can use to measure their physical activity, sleep-wake cycles, and activity rhythms at home [7]. From this, the circadian rhythm can be derived and monitored relatively easily and over a long period [3]. Wearables, including smartwatches, are becoming increasingly common. About 21% of people in the United States own a smartwatch [17]. With such wearables, continuously measuring activity in real-world scenarios is possible. This method offers a convenient, non-invasive, and cost-effective approach to data collection, and can be performed outside of clinical settings. These wearables are accurate in measuring accelerations and orientations in adults [18].

One of the variables that can be extracted from accelerometer data and track circadian rhythm disruptions is the Relative Amplitude (RA) [19]. This is the difference between the most active 10-hour period (when someone is awake) and the least active 5-hour period (when someone sleeps), normalized by their sum. In addition, midpoints can be determined for both periods, which indicate at what point during the day someone is most active and at what point someone goes to bed. A low RA means that people are relatively active at night and less active during the day, which is associated with lower happiness, lower health satisfaction, loneliness, and a reduced reaction time [6]. Moreover, a low RA increases the risk of health problems and diseases.

In addition to the RA, the Intradaily Variability (IV) and the Interdaily Stability (IS) can also be determined by using accelerometer data [13]. The former reflects the fragmentation of the rest-activity rhythm, while the latter

indicates how well the 24-hour rest-activity rhythm follows the 24-hour light-dark cycle. Another way to analyse circadian rhythms is by fitting a cosine curve to the accelerometer data [20]. Because accelerometer data reflects cyclic activity patterns, a 24-hour cosine curve can be fitted. From this curve, cosinor variables can be extracted: the amplitude, the phase (acrophase), and the center line (Midline Estimated Statistic Of Rhythm (MESOR)).

B. Problem definition

Previous studies have shown that circadian rhythm variables such as the RA, IS, IV, MESOR, amplitude, and acrophase are indicators of circadian rhythm disruptions and that these disruptions are associated with a greater risk of depression [8]. However, it is still unknown whether specific covariates confound the relationship between circadian disruptions and depression. Furthermore, predictive machine learning has rarely been used to investigate the influence of circadian rhythm disruptions on depression. One notable exception is a study by Lim et al., which used this approach. However, it focused solely on 160 South Korean adults with an average age of 64, limiting its relevance to broader, more diverse populations [21]. Additionally, no studies have specifically examined this relationship in the relatively younger age group during which depressive and mood disorders typically onset (21 to 45 years old and 20 to 45 years old, respectively), especially with a machine learning approach [22]. No studies have compared several machine learning models to determine which one has the best predictive performance. Furthermore, the relationship between circadian rhythm disruptions and depression is often examined using case-control studies. Only a few studies have used a population-representative dataset to explore this relationship. Despite this, most circadian rhythm research mainly uses data from the UK Biobank, whereas the Nutrition Examination Survey (NHANES) dataset, which represents the US population, is less frequently used. Still, some research has utilized the NHANES dataset to associate circadian rhythms with depression. For example, a study by Yin et al. explored the relationship between the sleep midpoint and depression using this dataset [23]. However, the NHANES dataset has never been used to examine the influence of circadian rhythm disruptions on depression and sleep disturbances using machine learning.

This study used the NHANES dataset to investigate the association between circadian rhythm disruptions and depression. Using machine learning models, patterns in accelerometer data will be explored, and the influence of circadian regularity measures (IS, IV, RA) and cosinor variables (amplitude, acrophase, and MESOR) on depression will be investigated, while also looking at the influence of covariates such as age, sex, and education. For additional validation, it will also be examined whether circadian rhythm disruptions are directly associated with a diagnosed sleep disorder or self-reported sleep disturbances. These considerations prompt this thesis's two primary research questions: *"To what extent is there an association between circadian rhythm disruptions and the risk of depression?"* with the hypothesis that this association is significant, and *"To what extent are circadian rhythm disruptions directly associated with sleep disturbances?"* to validate circadian rhythm disruptions in relation to sleep disturbances. Additionally, to evaluate the performance of the machine learning models, a secondary question will be explored: *"Which machine learning model can best uncover patterns in the data associated with depression, and how do covariates affect these patterns?"*.

C. Thesis overview

This thesis is structured as follows: Chapter II describes the methods used in this thesis. This consists of defining the study population and the processing of the data before examining the association between circadian rhythm disruptions and both depression and sleep disturbances using machine learning models, while also adjusting for several covariates. Chapter III outlines the results of this thesis. This chapter will also show the characteristics of the study population and the performance of the machine learning models. Finally, chapter IV summarizes the thesis's main findings, compares them to existing research, and gives recommendations for future research.

II. Methods

A. Study design

A cross-sectional observational study was conducted using data from the NHANES. Raw accelerometer data from Physical Activity Monitors (PAM) was analysed to derive circadian rhythm parameters, which were then used to associate circadian rhythm disruptions with depression and sleep disturbances using machine learning models. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for reporting observational studies [24].

1) Setting

This study used data from the 2011-2012 cycle of the NHANES. The NHANES is a large population-level dataset that has monitored people's health in the United States since the 1960s [25]. This cross-sectional survey is conducted annually by the Centers for Disease Control and Prevention (CDC), the national health protection agency of the United States [26]. The random sample used by the NHANES to select participants is representative of the entire U.S. population [27]. Data collection methods for the NHANES include interviews, health and body measurements, and laboratory testing. The NHANES data files are publicly available for download and analysis, with personally identifiable information removed so that all participants remain anonymous. The National Center for Health Statistics Ethics Review Board (NCHS ERB) has approved the NHANES study protocols (NCHS ERB Protocol Number #2011-17 and NCHS ERB Protocol Number #2018-01) [28].

The NHANES used several ways to obtain data, including examination data, which includes PAM data from accelerometer measurements. This monitor was reintroduced in the NHANES in 2011, after first being introduced in 2003 [29]. There are several available PAM data formats: the raw data, where accelerometer values in the x, y, and z direction are sampled at 80 Hz, and summarized accelerometer data aggregated per minute, hour, and day. Each year, around 7.000 participants aged 6 years and older took part in the physical activity measurement, where they wore an ActiGraph GT3X+ (Actigraph, Pensacola, FL) triaxial accelerometer for seven consecutive days, 24 hours per day, on the non-dominant wrist. The other type of data used during this study was questionnaires. This provided information about the participants' mental health conditions. Linking the examination data with the questionnaire data enabled the analysis of circadian rhythm parameters in relation to depression and sleep disturbances. The data is publicly available on the NHANES website: *https://wwwn.cdc.gov/nchs/nhanes/default.aspx*.

2) Study population

In total, 6.917 participants completed the PAM measurements during the 2011-2012 cycle. Participants first had to have valid data to be included in the study, meaning they had to wear the PAM for more than 16 hours on at least 4 days. This information was obtained from the *Physical Activity Monitor - Day* file. By summing the *valid wake wear* and *valid sleep wear* minutes per day per participant, it was determined whether the 16-hour threshold was met for at least 4 days. Additionally, the same file contained data quality flags summed per day, and participants with at least one data quality flag were excluded.

Using the *Demographic Variables & Sample Weights* file, pregnant participants were excluded, as pregnancy often causes insomnia and mood disorders, which could influence the results [30]. Furthermore, the sample was filtered by age. Only participants between 20 and 45 years old were included. Additionally, participants had to have valid data for all covariates and questions related to the outcomes (depressive disorders and sleep disturbances). Those who answered "don't know", refused to answer, or had missing data were excluded. In total, 392 participants met the eligibility criteria.

B. Variables

1) Preprocessing of raw accelerometer data

The SEQNs (the respondent sequence numbers, also referred to as IDs) of the included participants were used to select the correct files for extraction. Because all .tar files started with the same SEQNs, only those with matching SEQNs could be extracted. The extracted files were placed in a temporary directory. Subsequently, all .csv files resulting from one unzipped .tar file were read and merged into one data frame, which was then used for analysis. Fig. 1 shows a plot of the raw triaxial accelerometer data over 24 hours from participant 62189.

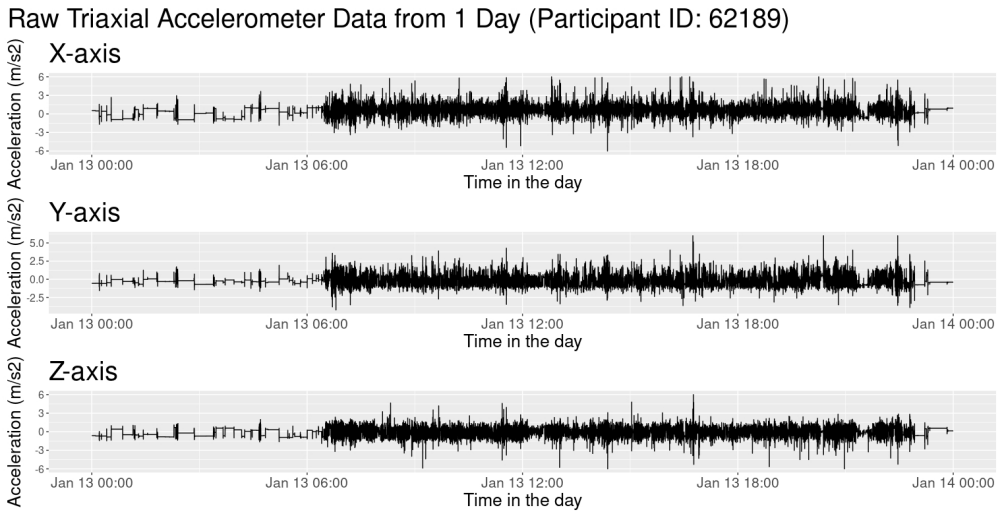


Fig. 1: *Raw triaxial accelerometer data over 24 hours for participant 62189, sampled at 80 Hz. The x-axis is shown in the upper plot, the y-axis in the middle plot, and the z-axis in the lower plot. The plot covers data from January 13th at 00:00 to January 14th at 00:00. The x-axis of the plot shows the time of day, and the y-axis shows the acceleration in m/s^2.*

Subsequently, the 80 Hz tri-axial accelerometer data was summarized per minute by calculating the Euclidean Norm Minus One (ENMO) metric, as depicted in Eq. (1). This calculation computes the magnitude of the vector, which consists of acceleration data in three axes, and then subtracts one gravitational unit (1g) as a correction for gravity [31].

$$ENMO_t = \frac{1}{n_t}\sum_{i=1}^{n_t}(\sqrt{x_i^2 + y_i^2 + z_i^2} - 1)\mathbb{1}(\sqrt{x_i^2 + y_i^2 + z_i^2} > 1) \tag{1}$$

where i represents each sample and t represents the time point rounded to one second [32].

The raw accelerometer data was summarized into ENMO_t values per minute using the *SummarizedActigraphy* package in R. Fig. 2 shows the summarized ENMO_t values over 24 hours for the same participant shown in Fig. 1. As a final preprocessing step, the output containing the summarized actigraphy data from one participant was written to a new .csv file, after which the extracted raw accelerometer files were removed from the temporary directory.

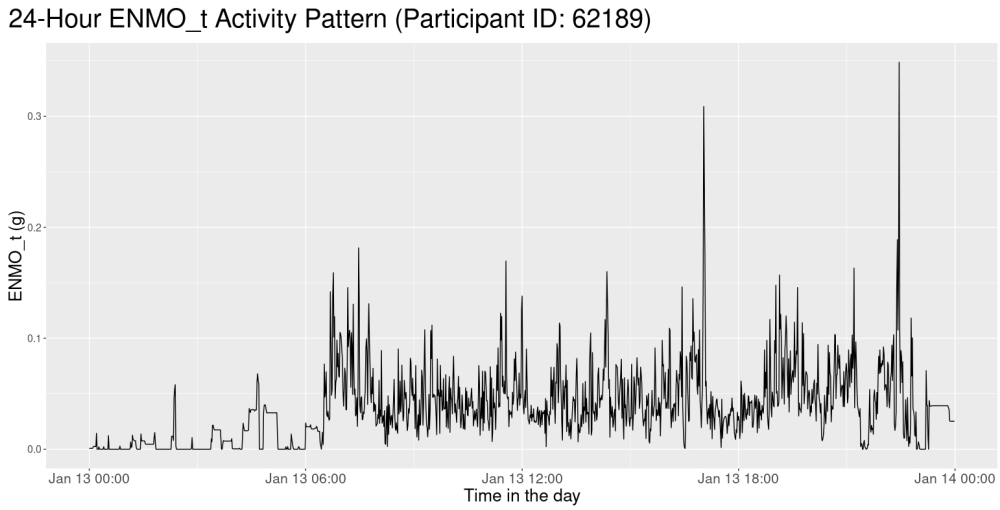


Fig. 2: *Summarized $ENMO_t$ values over a 24-hour period for participant 62189. The plot covers data from January 13th at 00:00 to January 14th at 00:00. The x-axis shows the time of day, and the y-axis shows the $ENMO_t$ values in gravities. The peaks indicate that this participant was active starting around 7:00 in the morning.*

2) Extraction of non-parametric circadian rhythm parameters

To assess circadian regularity, three different non-parametric (since they are not associated with a known function [33]) measures were evaluated: the Interdaily Stability (IS), reflecting day-to-day pattern consistency, the Intradaily Variability (IV), indicating rhythm fragmentation, and the Relative Amplitude (RA), indicating the difference between the most active 10-hour period and the least active 5-hour period [7, 19]. An IV of 0 indicated a perfect sine wave, while an IV of 2 consisted mainly of Gaussian noise. These high IVs were primarily observed in people awake at night and asleep during the day. An IS of 0 reflected primarily Gaussian noise, whereas a value of 1 indicated perfect consistency. The higher the IS was, the more closely the circadian rhythm synchronized with the external light-dark cycle. The RA has a value between 0 and 1. A higher RA indicated a better rest-activity rhythm, which meant a larger difference between the activity while awake and asleep. These metrics were computed per participant (i.e., per summarized actigraphy file) using the *nparACT* package in R, using the time and ENMO_t data as input. Using a for loop, the outcomes of all participants were combined into a single data frame. A column containing each participant's SEQN was added to this data frame for later identification. Finally, this data frame was written to a .csv file.

To ensure that computations for each participant covered full 24-hour periods from 00:00 to 23:59, only days with complete minute-by-minute data were included in the circadian regularity assessment. This effectively meant that the first and last monitoring days were excluded from (nearly) all participants. The IS and IV were calculated using the formulas shown in Eq. (2) and Eq. (3):

$$IS = \frac{n \sum_{h=1}^{p} (\overline{X}_h - \overline{X})^2}{p \sum_{i=1}^{n} (X_i - \overline{X})^2} \tag{2}$$

$$IV = \frac{n \sum_{i=2}^{n} (X_i - X_{i-1})^2}{(n-1) \sum_{i=1}^{n} (X_i - \overline{X})^2} \tag{3}$$

in which n is the total number of sampling points per participant, p the number of sampling points per day, $\overline{X}_\text{h}$ the hourly means, $\overline{X}$ the grand average of all data per participant, and X_i the activity value from each sampling point [34].

The RA was calculated using the formula as depicted in Eq. (4):

$$RA = \frac{M10 - L5}{M10 + L5} \tag{4}$$

in which $M10$ is the averaged activity of the most active 10-hour period and $L5$ is the averaged activity of the least active 5-hour period [34].

Fig. 3 presents the 24-hour actigraphy plot from participant 62189, averaged across all days:

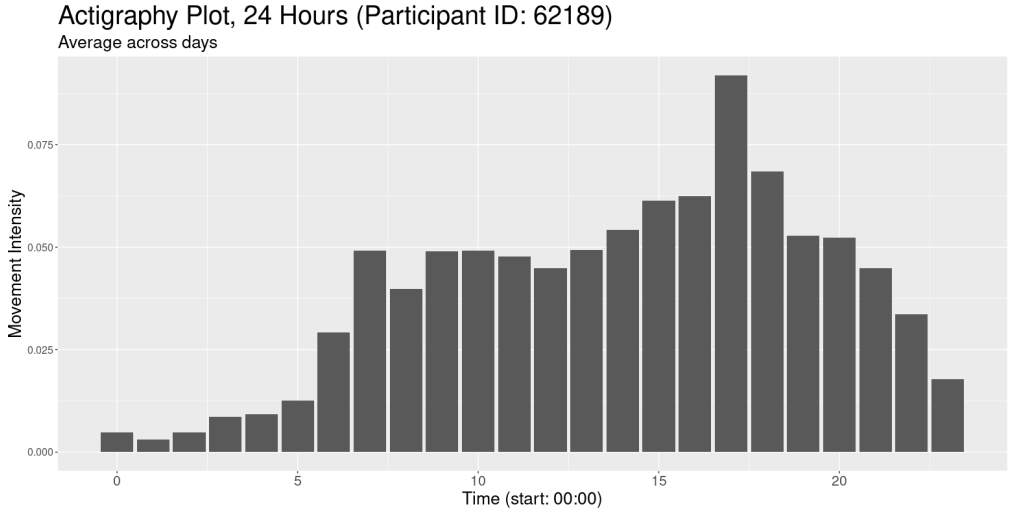


Fig. 3: *24-hour actigraphy plot for participant 62189, averaged across all days. The x-axis shows the time of day from 00:00 to 23:59, in hours, and the y-axis shows the movement intensity (ENMO_t values). The peaks indicate high movement intensity.*

3) Extraction of circadian features using Cosinor analysis

A cosine function was fitted to the 24-hour cycle of $ENMO_t$ values representing the circadian rhythm using cosinor analysis, which determined the best fit between the data and a 24-hour periodic cosine curve [7]. This was done by using the *cosinor* package in R. This package was used to calculate the MESOR, which represented the average activity level, the amplitude, which indicated variation in activity within the cycle, and the acrophase, which marked the time of highest observed activity [35]. A higher MESOR reflected a greater average activity, while a larger amplitude indicated stronger rhythmicity, representing a more pronounced difference between the most active and least active periods [36]. The function for a single-component cosinor model is expressed as shown in Eq. (5):

$$Y(t) = M + A \cdot cos(\frac{2\pi t}{\tau} + \phi) + e(t) \tag{5}$$

in which $Y(t)$ is the accelerometry activity at time t, M the MESOR, A the amplitude, ϕ the acrophase, τ the period (i.e. one cycle, 24 hours) and $e(t)$ the error term [7, 35].

As τ was known, the model was rewritten using the trigonometric angle sum identity, expressed in Eq. (6):

$$Y(t) = M + \beta x + \gamma z + e(t) \tag{6}$$

where $\beta = A \cdot cos(\phi)$; $x = cos(\frac{2\pi t}{\tau})$; $y = -A \cdot sin(\phi)$; $z = sin(\frac{2\pi t}{\tau})$ [35].

Using the least squares method and solving the resulting normal equations, estimates for M, β, and γ were obtained [35], resulting in a fitted cosine curve that captured the 24-hour activity pattern. These computations were done by the *cosinor* package. The *cosinor2* package was used to correct the acrophase, after which its absolute value was taken to ensure it corresponded to the maximum value of the cosine function. Since the calculated acrophase was in radians, the *astroFns* package was used to convert this to clock hours, minutes, and seconds to enhance interpretability. To ensure that each participant's cosinor analysis started at the same time (i.e., 00:00), only days with 24 hours of data were included. This effectively meant that for all participants, the first and last days were removed from the analysis. The time used as input in the *cosinor* function was calculated as the number of hours elapsed since the start of the measurement. The other input to the *cosinor* function was the $ENMO_t$ values. Fig. 4 compares the $ENMO_t$ values with its fitted cosine over two different days from the same participant, with day 1 visualized at the top and day 5 displayed at the bottom. The figures on the left show a cosine fitted over the $ENMO_t$ values from the same day, whereas the figures on the right show a cosine fitted over the $ENMO_t$ values averaged across 7 days.

The cosinor features were calculated per summarized actigraphy file, resulting in one set of cosinor parameters per participant, averaged across the entire week. Using a for loop, the outcomes of all participants were combined into a single data frame. A column containing each participant's SEQN was added to this data frame for later identification. In the end, this data frame was written to a .csv file. Fig. 5 illustrates the fitted cosine, averaged across 7 days. The figure shows how the MESOR, acrophase, and amplitude were determined.

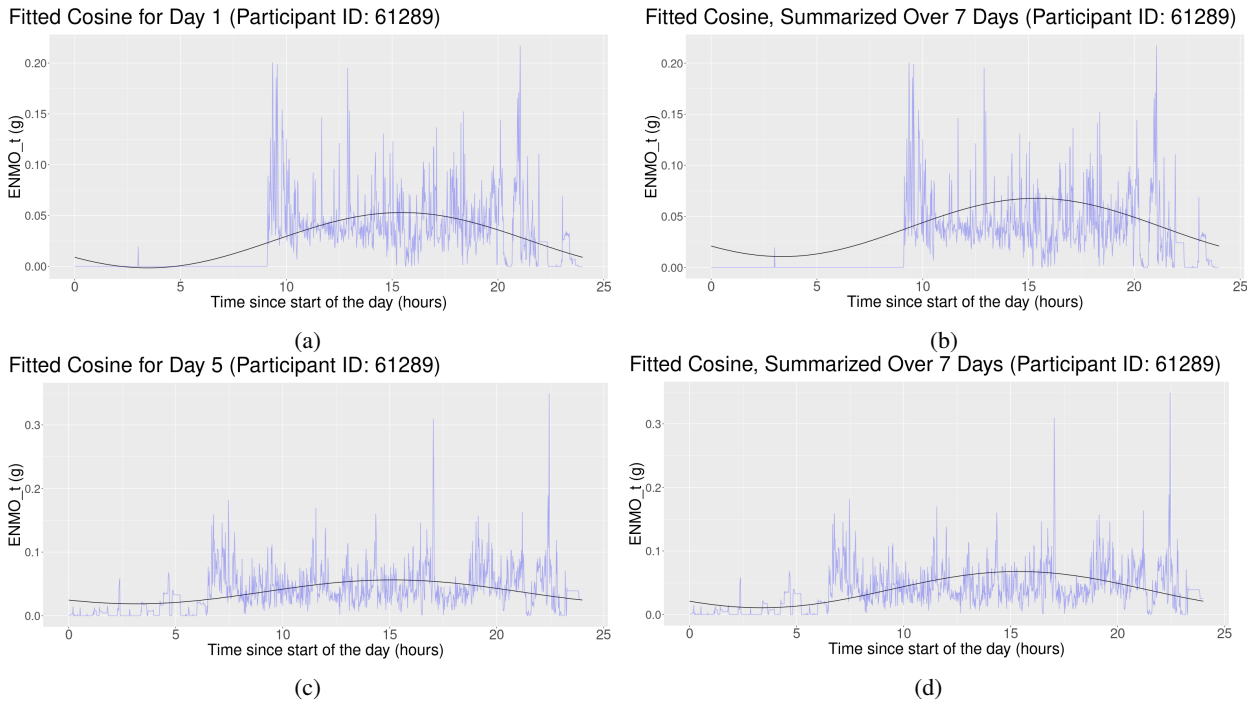


Fig. 4: *$ENMO_t$ values with its fitted cosine over two different days. These figures were partly created by the cosinor package. The x-axis shows the time since the start of the day in hours, and the y-axis shows the $ENMO_t$ values in gravities.* ***(a)*** *$ENMO_t$ values from day 1 with the fitted cosine of that day.* ***(b)*** *$ENMO_t$ values from day 1 with the fitted cosine summarized across 7 days.* ***(c)*** *$ENMO_t$ values from day 5 with the fitted cosine of that day.* ***(d)*** *$ENMO_t$ values from day 5 with the fitted cosine summarized across 7 days.*

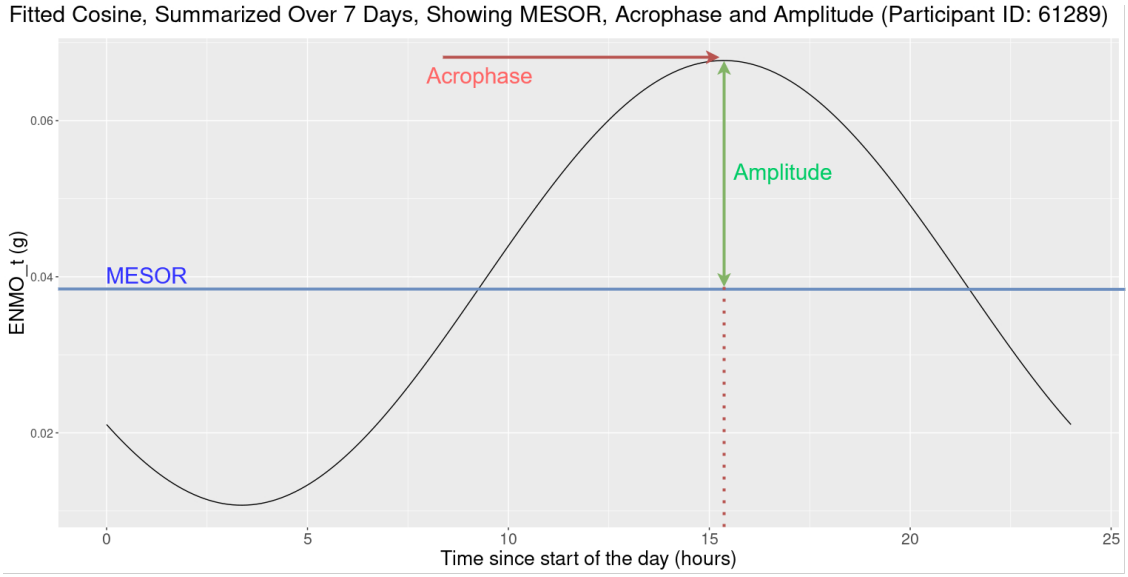


Fig. 5: *Fitted cosine, summarized across 7 days. This figure was created by the cosinor package. The x-axis shows the time since the start of the day in hours, and the y-axis shows the $ENMO_t$ values in gravities. The figure shows the MESOR (the mean), amplitude, and acrophase (the time at which the activity is the highest). For this participant, the MESOR is 0.039 g, the amplitude is 0.028 g, and the acrophase is 15:22.*

4) Depression classification

To determine whether each participant had experienced heightened depressive symptoms during the 2011-2012 cycle, the *Mental Health - Depression Screener* questionnaire was used. This questionnaire provided a useful indication of depressive symptoms. In this questionnaire, participants were asked the following questions:

1) "Do you have little interest in doing things?"
2) "Do you feel down, depressed or hopeless?"
3) "Do you have trouble sleeping or are you sleeping too much?"
4) "Do you feel tired or have little energy?"
5) "Do you have a poor appetite or are you overeating?"
6) "Do you feel bad about yourself?"
7) "Do you have trouble concentrating on things?"
8) "Do you move or speak slowly or too fast?"
9) "Have you thought you would be better off dead?"

The response options for each question were "Not at all", "Several days", "More than half the days", and "Nearly every day". Participants met the criteria for major depressive disorder (MDD) if they answered at least "More than half the days" to more than four questions or answered at least "More than half the days" to question 1) or 2), according to the PHQ-9 criteria [37]. These participants were classified as experiencing heightened depressive symptoms and were subsequently assigned to the "depression" group.

5) Sleep disturbance classification

To determine whether each participant had experienced sleep disturbances during the 2011-2012 cycle, the *Sleep Disorders* questionnaire was used. This questionnaire provided a useful indication of sleep disturbances, since participants were asked the following questions:

1) "Have you ever told a doctor you had trouble sleeping?"
2) "Were you ever told by a doctor that you have a sleep disorder?"

The response options for both questions were "Yes" and "No." Participants who answered "Yes" to at least one of those questions were classified as experiencing sleep disturbances.

6) Covariates

The following covariates were included in the analysis, as they have been associated with circadian rhythms, depressive disorders, and/or sleep disturbances in previous research:

- **Age**: The average age of onset of depressive disorders is between 21 and 45 years [22]. The incidence of depression is also increasing among young adults (aged 18-29 years) [38]. Furthermore, age is positively linked to sleep disturbances such as insomnia, and it also influences circadian rhythms, where it is associated with a changed rhythm phase, a reduced circadian rhythm amplitude, and a more vulnerable circadian-sleep relationship [39, 40].
- **Sex**: Men tend to go to bed later and wake up later, resulting in social jet lag, which means that the circadian rhythm is out of sync with social demands [41]. Additionally, sleep disturbances such as insomnia are more common in women than men, and there is also a higher incidence of depressive disorders in women [39, 42].
- **Race**: Minority groups (Hispanic, African, and Asian Americans) tend to sleep shorter than non-Hispanic white Americans [43]. Furthermore, African Americans show a shorter average circadian rhythm than non-Hispanic white Americans. In addition, minority groups not only have lower rates of depressive disorder symptoms, but they are also more likely to be misdiagnosed and are less likely to seek help for their mental health problems [44]. The dataset included seven different options for races: Mexican American, other Hispanic, non-Hispanic white, non-Hispanic black, non-Hispanic Asian, and others, which included multi-racial participants.
- **Education level**: People with a low education level, especially between the ages of 28 and 50, have a greater risk of depression [45]. On the other hand, people with a higher educational level are more likely to suffer from insomnia and daytime sleepiness [46]. The dataset included five options for educational levels: less than 9th grade, 9-11th grade, high school graduate, a college degree, and anything above a college degree. For the analysis, participants with a high school diploma or less were classified as having low education, while those above that were classified as highly educated.

- **Household size**: The more people live in one household, the greater the chance of insufficient sleep and depression [47, 48]. Participants' households with less than three people were classified as small, households of up to 5 people as medium, and anything above that as large.
- **Smoking status**: Studies have shown that circadian rhythms in rats are disrupted by passive smoking [49]. Cigarette smoking also increases the risk of both sleep and depressive disorders considerably, both of which show a dose-response relationship [50]. In the dataset, smoking status was based on having smoked at least 100 cigarettes in life. Participants who smoked more than 100 cigarettes were considered smokers, and the other participants were considered non-smokers.
- **Alcohol consumption**: Consuming alcoholic beverages changes the phase of circadian rhythms and causes the circadian rhythm to no longer influence body temperature and locomotor activity [51]. In addition, alcohol consumption also has a positive effect on the risk of depressive episodes and sleep disorders, the latter because alcohol causes insomnia and disrupts circadian rhythms [52, 53]. In the dataset, alcohol consumption was based on the average number of alcoholic drinks per day in the last year. Participants consuming 1 or fewer alcoholic drinks per day were classified as non-drinkers, those consuming between 2 and 4 alcoholic beverages as light drinkers, and those consuming more than that as heavy drinkers, based on the definition from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) [54].
- **Employment**: The risk of depressive disorders increases if someone is unemployed [55]. In addition, unemployed people also suffer more from insomnia, disturbed sleep, and they tend to get either more or less sleep compared to people with a job [56, 57]. In this study, participants who answered that they had been working at a job or business in the past week were considered employed, and participants who were looking for a job or did not work at a job or business were considered unemployed.
- **BMI**: Disruptions in the circadian rhythm cause metabolic processes to be disrupted [58]. This increases the risk of (the progression of) obesity. Furthermore, the risk of depression in people with obesity is also increased, and they have a greater chance of long-term sleep disorders [59, 60]. In the dataset, participants had filled in their height (in inches) and weight (in pounds). For the analysis, the BMI was calculated using this information with the formula defined in Eq. (7):

$$703 \cdot \frac{weight\ (pounds)}{height\ (inches)^2} \tag{7}$$

  Participants with a BMI below 18.5 were classified as underweight, those between 18.5 and 24.9 were classified as within the optimal range, those between 25 to 29.9 were classified as overweight, and any participant above that as obese [61].

Covariate data were obtained from the following datasets, all available on the NHANES website:

- *Demographic Variables & Sample Weights* for age, sex, race, education level and household size.
- *Smoking - Cigarette Use* for smoking status.
- *Alcohol Use* for alcohol consumption.
- *Occupation* for employment status.
- *Weight History* for BMI data (derived from height and weight).

C. Merging all features and outcomes into a unified dataset

Each participant had a single set of features summarizing the entire week: one set of cosinor features (i.e., MESOR, amplitude, and acrophase), one set of non-parametric circadian features (i.e., IS, IV, and RA), and one set of covariate features. A script was used to merge all of these features and the depression/sleep disturbance classifications. No data imputation was performed, and all participants with missing data were excluded from the analysis. As the SEQN was included in all questionnaires and appended to each participant's feature data during the feature extraction, all tables could be merged based on the SEQN, mainly by using inner joins. The output of this script was one large data frame with all circadian features, covariates, and outcomes of all participants. NHANES sampling weights and strata were also part of this merged data frame. The final step was writing the table to a .csv file, which was then used to train machine learning models.

D. Implementation of machine learning models

To develop predictive models of depression (yes/no) and sleep disturbances (yes/no), multiple machine learning models were trained using the *tidymodels* package in R. For reproducibility, a seed was first set. Then, the dataset

was split into a ratio of 80/20, stratified by depression or sleep disturbances, depending on the model in question. Based on this split, 80% of the data was used to train the model and the remaining 20% for testing.

A workflow map was set up, consisting of 24 different workflows for depression and 27 for sleep disturbances. The workflows combined specific preprocessing steps (referred to as recipes in *tidymodels*) and machine learning models, using the *workflowsets* package in R. All of the recipes used a distinct set of predictor variables. Four out of eight (five out of nine for sleep disturbances) recipes included the covariate variables, while the others only included circadian rhythm variables. Due to class imbalance in the depression outcome (50-342), the *step_smote()* function was applied to all of the depression recipes. This function generated synthetic samples of the minority class to make the classes more balanced [62]. An over-ratio of 1 was used, meaning the number of depression samples generated matched the number of healthy controls. Since the imbalance in sleep disturbances was less pronounced (120-272), several over-ratios (1 and 0.6) were tested in the recipes for sleep disturbances. Further recipe options for both outcomes included interactions between several variables, using the *step_interact()* function. The model recipes used are shown in Table I.

TABLE I: Model recipes overview

Outcome	Predictors	Over-ratio	Interactions
Both	All	1	None
Both	All	1	MESOR x amplitude; alcohol x smoking x BMI; IS x IV x RA
Both	All	1	alcohol x smoking
Both	All	1	alcohol x BMI
Both	All	1	IS x IV
Both	Circadian only	1	None
Both	Circadian only	1	MESOR x amplitude; IS x IV
Depression	Circadian only	1	MESOR x amplitude
Sleep disturbances	Circadian only	0.6	MESOR x amplitude
Sleep disturbances	All	0.6	IS x IV

MESOR = Midline Estimated Statistic of Rhythm, IS = Interdaily Stability, IV = Intradaily Variability, RA = Relative Amplitude, BMI = Body Mass Index.

Three types of machine learning models were trained: a k-nearest neighbors model, a logistic regression model, and a random forest model with 1000 decision trees, all using classification mode. Hyperparameter tuning was performed on the k-nearest neighbors and random forest models using grid search, evaluating 15 parameter combinations. For the k-nearest neighbors model, the number of neighbors and the weighting function were tuned. In contrast, for the random forest model, the number of predictors randomly sampled at each split and the minimum number of data points in a node were tuned.

The machine learning models were evaluated using 10-fold cross-validation and the key performance metrics visible in Table II [63]:

TABLE II: Key performance metrics overview

Metric	Definition	Formula
Accuracy	Proportion of all participants who were correctly classified	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Proportion of participants correctly identified as having sleep disturbances or depression out of all participants predicted to have sleep disturbances or depression	$\frac{TP}{TP+FP}$
Recall	Proportion of participants with sleep disturbances or depression who were correctly identified	$\frac{TP}{TP+FN}$
Specificity	Proportion of participants without sleep disturbances or depression who were correctly identified	$\frac{TN}{TN+FP}$
F1-score	A balance between precision and recall	$\frac{2 \cdot Pre. \cdot Rec.}{Pre. + Rec.}$
AUC	How well the model can distinguish between participants with and without sleep disturbances or depression [64]	N/A

AUC = Area Under the Curve, TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives, Pre. = Precision, Rec. = Recall.

For each possible workflow option, the best-performing model was selected based on the F1-score, after which the workflow was finalized. The test set was then evaluated using these finalized workflows, and the resulting predictions were used to calculate performance metrics. Finally, using the *vip* package in R, the variables with the most considerable influence on the predictions were analysed, which provided insight into the most important determinants of depression and sleep disturbances.

1) Post-fitting evaluation

To improve the reliability of the machine learning models, calibration was applied to the best-performing models [65]. Model calibration describes how much a model's predicted probability corresponds to real-world outcomes. Using a calibrator, the scores produced by the models can be adjusted to represent true probabilities better [66]. Calibration ensured that the key performance metrics became more reliable for further analysis and conclusions. Logistic regression (Platt scaling) calibration was used for the calibration, a parametric method that performs well on smaller datasets [66].

2) Removal of equivocal zones

Some predicted probabilities fell within an "equivocal" zone, an area around a decision threshold where the model was uncertain about the actual outcome [67]. This zone was defined as the interval $threshold \pm buffer$ and was determined by analysing the calibrated models to identify probabilities with high error rates. All predictions falling within this zone were removed. By excluding those predictions, key performance metrics improved. Various thresholds and buffer zones were tested for all of the best-performing models. For the k-nearest neighbors model predicting depression, the best performance was achieved at a threshold of 0.60 and a buffer of 0.05. For the random forest model predicting depression, a threshold of 0.30 and a buffer of 0.05 were applied. The sleep disturbance model achieved the best performance metrics at a threshold of 0.40 and a buffer of 0.005.

E. Sensitivity analysis

A sensitivity analysis was conducted to examine whether having an extremely fragmented or inconsistent circadian rhythm, defined as having an IV value above the 95th percentile or an IS value below the 5th percentile, respectively, negatively influenced the results. This was done based on the research from Shim et al. [7]. As a result, the sensitivity analysis excluded 36 additional participants, leaving 356 participants in the sensitivity cohort.

All analyses were conducted using R software 4.4.3 and RStudio version 2024.12.1. The code used to conduct all analyses is available in Appendix B.

III. Results

A. Population characteristics

This study studied 392 participants with available PAM data, all with valid measurements and complete information on the covariates. Fig. 6 shows the flowchart for inclusion and exclusion of study participants. In total, two different datasets were studied: the primary dataset, which included all 392 participants, and the sensitivity dataset, which excluded a total of 36 additional participants with an extremely fragmented (IV $>$ 95th percentile) and extremely inconsistent (IS $<$ 5th percentile) circadian rhythm. After applying an 80/20 split on the datasets, the primary dataset used 314 participants to train the models, while 78 participants were used for testing. The sensitivity dataset included 356 participants, of which 285 were used for training and 71 for testing.

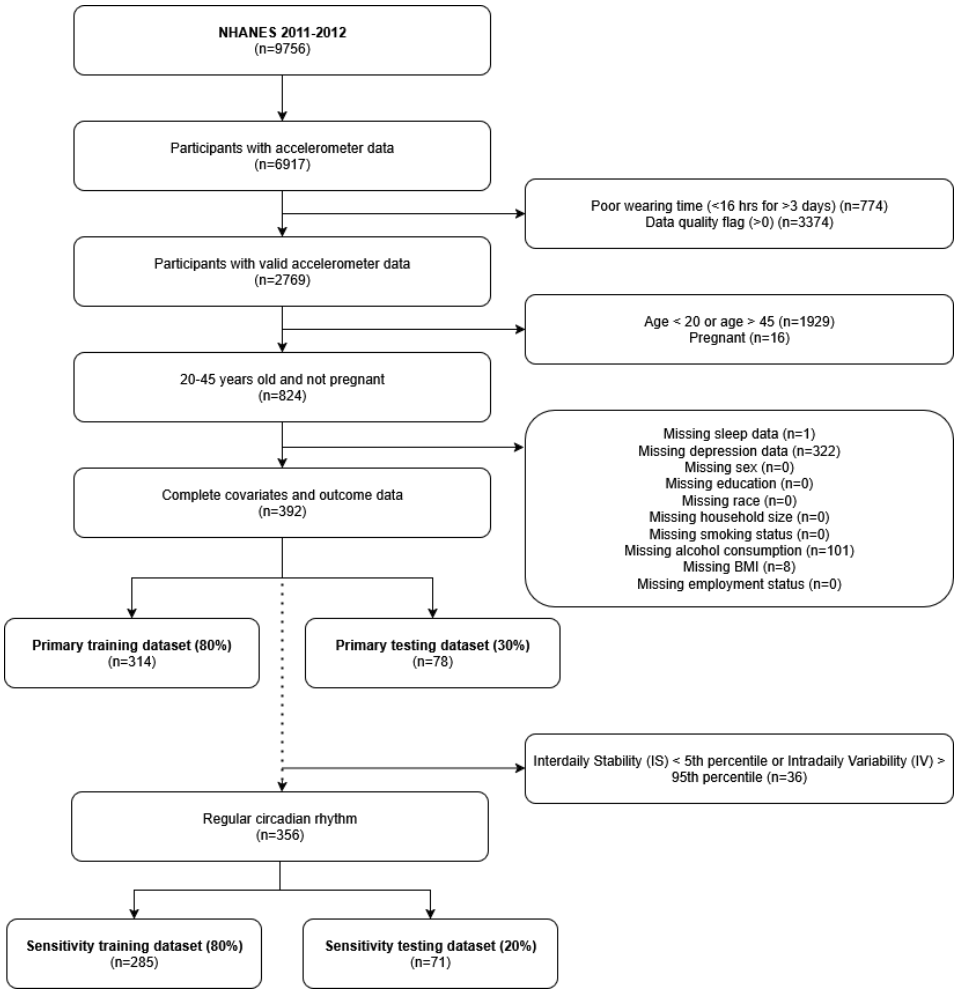


Fig. 6: *Flowchart for inclusion and exclusion of study participants from the NHANES.*

Table III describes the characteristics of the population. Of the primary cohort (392 participants, median age 33 years), 225 participants (59.6%) were female and 167 (40.4%) were male. Most participants were of white origin (66.9%) and highly educated (70.3%). 272 participants (73.4%) were unemployed, and most participants (51.8%) lived in a medium household. Furthermore, most participants were obese (32.2%), had never smoked (55.0%), and were light drinkers (54.7%). In total, there were 26 participants (3.6%) who met the criteria for heightened depressive symptoms, 96 (28.2%) participants who met the criteria for sleep disturbances, and 24 (5.8%) participants who met the criteria for both. Participants with heightened depressive symptoms were referred to as the depression group. The participants' median accelerometer-derived circadian rhythm features were 23.85 mg for the MESOR, 17.94 mg for the amplitude, 15:06 for the acrophase, 0.86 for the IV, 0.41 for the IS, and 0.67 for the RA.

TABLE III: Population characteristics

Median [IQR] or n (%)	**Primary cohort (n = 392)**	**Sensitivity cohort (n = 356)**
Age (years)	33 [27, 41]	33 [27, 42]
Sex		
Female	225 (59.6)	204 (59.5)
Male	167 (40.4)	152 (40.5)
Race		
White	170 (66.9)	152 (66.2)
Black	86 (10.5)	80 (10.7)
Mexican-American	45 (10.0)	42 (10.2)
Other Hispanic	32 (6.6)	32 (7.3)
Asian	44 (3.6)	36 (3.3)
Other	15 (2.3)	14 (2.3)
Education		
Low educated	136 (29.7)	124 (30.2)
Highly educated	256 (70.3)	232 (69.8)
Employment		
Employed	120 (26.6)	107 (25.2)
Unemployed	272 (73.4)	249 (74.8)
Household size		
Small	129 (39.2)	114 (38.5)
Medium	218 (51.8)	200 (52.2)
Large	45 (9.0)	42 (9.3)
BMI		
Underweight	7 (1.0)	5 (0.9)
Normal	126 (30.9)	110 (30.0)
Overweight	116 (35.9)	108 (36.4)
Obese	143 (32.2)	133 (32.7)
Smoking status		
Non-smoker	224 (55.0)	202 (55.1)
Smoker	168 (45.0)	154 (44.9)
Alcohol consumption		
No drinker	106 (24.9)	95 (25.0)
Light drinker	200 (54.7)	179 (54.3)
Heavy drinker	86 (20.3)	82 (20.8)
Group		
Depression	26 (3.6)	25 (3.7)
Sleep disturbances	96 (28.2)	87 (28.2)
Both	24 (5.8)	22 (5.9)
Healthy controls	246 (62.4)	222 (62.2)
Circadian rhythm features		
MESOR (mg)	23.85 [23.85, 34.55]	28.90 [24.41, 34.55]
Amplitude (mg)	17.94 [11.97, 22.50]	18.21 [12.92, 22.71]
Acrophase (clock hour)	15:06 [14:02, 16:12]	15:08 [14:07, 16:13]
IV	0.86 [0.68, 1.09]	0.84 [0.68, 1.04]
IS	0.41 [0.29, 0.51]	0.43 [0.31, 0.52]
RA	0.67 [0.50, 0.75]	0.78 [0.50, 0.75]

Adjusted for 2-year sample weights.
IQR = Inter-Quartile Range, n = number of participants, BMI = Body Mass Index,
MESOR = Midline Estimating Statistic Of Rhythm, IV = Intradaily Variability,
IS = Interdaily Stability, RA = Relative Amplitude.

B. Outcomes

Visible differences were observed in accelerometer-derived $ENMO_t$ values between participants in the healthy controls group, the depression group, and the sleep disturbances group. Fig. 7 compares this per day over one week between three random participants from each group (for illustration purposes). It can be seen that participant

63033 (from the sleep disturbances group) showed more activity than the other two participants during the sleep periods. It can also be observed that participant 67442 (from the healthy controls group) had a more consistent sleep, with this participant waking up around the same time almost every day. The activity patterns of the same three participants, averaged across all days, are shown in Fig. 8. This figure reveals that participant 67442 (from the healthy controls group) had a higher average movement intensity than the other two participants. However, participant 62189 (from the depression group) and participant 63033 (from the sleep disturbances group) showed the largest movement intensity difference between the day and night periods. Fig. 9 displays a similar figure, now summarizing the activity patterns of all participants per group. Although the activity patterns were comparable across groups, some subtle differences could be observed. For example, the average activity was slightly higher in the healthy controls group compared to the other groups, while the group with both outcomes showed the lowest average activity. In addition, the depression group exhibits the highest activity during the night.

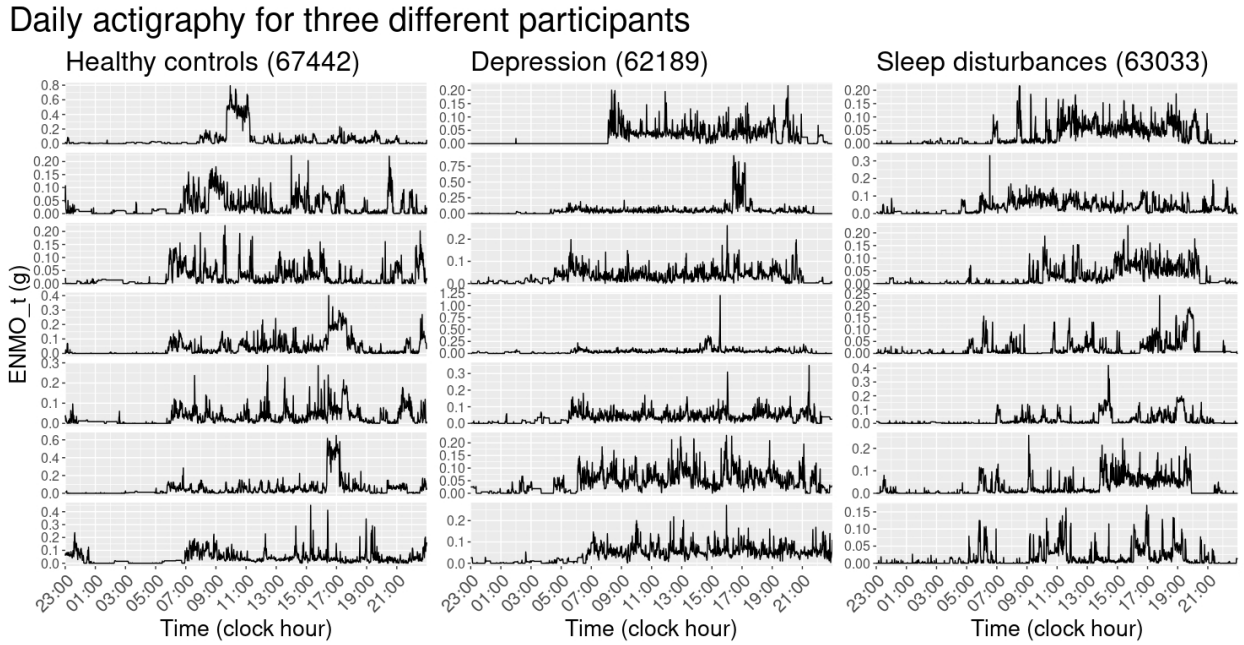


Fig. 7: *$ENMO_t$ values from three different participants from three different groups (healthy controls, depression, and sleep disturbances), summarized per minute, across one week. The x-axis shows the time in clock hours, from 23:00 to 22:59 the next day, while the y-axis shows the $ENMO_t$ values across 7 different days. The y-axis scale is not fixed. The value level indicates the level of physical activity. Higher $ENMO_t$ values indicate that the participant was active during those moments (active periods), and low $ENMO_t$ values indicate little to no movement (rest periods).*

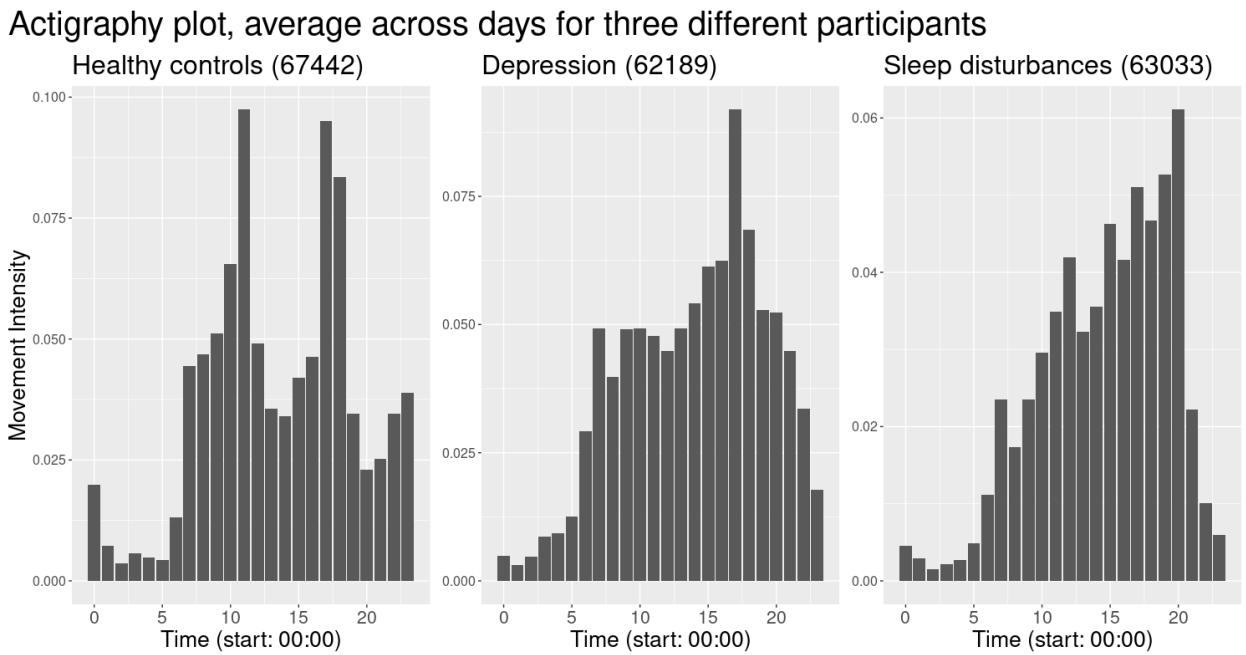


Fig. 8: *Averaged movement intensity across all days from three different participants from three different groups (healthy controls, depression, and sleep disturbances). The x-axis shows the time in hours (from 0:00 to 23:00), while the y-axis shows the movement intensity. The y-axis scale is not fixed. The movement intensity indicates the level of physical activity. Higher movement intensity indicates that the participant was active (active periods), and lower movement intensity indicates little to no movement (rest periods).*

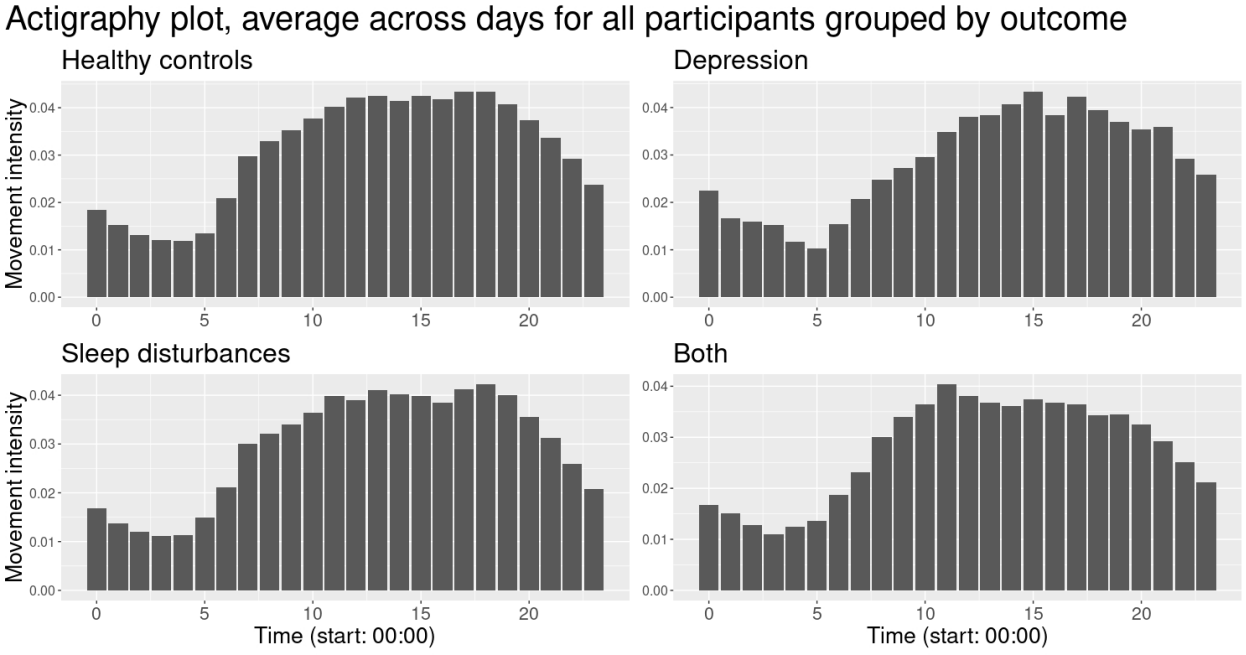


Fig. 9: *Averaged movement intensity across all days from the different groups (healthy controls, depression, sleep disturbances, and both). The x-axis shows the time in hours (from 0:00 to 23:00), while the y-axis shows the movement intensity. The y-axis scale is not fixed. The movement intensity indicates the level of physical activity. Higher movement intensity indicates that the participants from that group were active (active periods), and lower movement intensity indicates little to no movement (rest periods).*

1) Depression

As shown in Table IV, the circadian rhythm parameters differ (although not significantly) between participants who met the criteria for depression and the healthy controls. The p value is also shown, determined using Welch's Two Sample t-test. Significance was set at $p < 0.05$. In addition, Cohen's d is also shown in the table. Depression resulted in a lower average MESOR, amplitude, IS, and RA, a later average acrophase, and a higher average IV. However, looking at the p value, these differences were not significant. Cohen's d also reflected negligible differences across all parameters. A similar pattern is also evident in the density plots illustrated in Fig. 10. Apparent differences in covariates between participants who met the depression criteria and healthy controls were observed, as shown in Table V, the bar plots in Fig. 11, and the density plot in Fig. 12. As highlighted in these figures, the incidence of depression was relatively the highest among older, female participants of black descent who were obese, low educated, unemployed, had large households, were heavy drinkers, and smoked. Statistical significance was found for differences in education, employment, household size, and alcohol consumption, as indicated in the table.

TABLE IV: Circadian rhythm parameters of participants with depression and healthy controls

Variable	**Depression**		**Healthy controls**		p	**Cohen's** d
	Mean	SD	Mean	SD		
MESOR (g)	0.0297	0.0105	0.0305	0.00953	0.647	0.07
Acrophase (clock hour)	15:16:43	03:17:33	15:05:16	02:23:42	0.695	-0.08
Amplitude (g)	0.0177	0.00803	0.0179	0.00781	0.839	0.03
IS	0.401	0.139	0.403	0.148	0.928	0.01
IV	0.882	0.265	0.855	0.302	0.942	0.01
RA	0.615	0.228	0.641	0.221	0.445	0.12

Statistical significance (*) = $p < 0.05$, according to Welch's Two Sample t-test.
SD = Standard Deviation, MESOR = Midline Estimating Statistic Of Rhythm,
IV = Intradaily Variability, IS = Interdaily Stability, RA = Relative Amplitude.

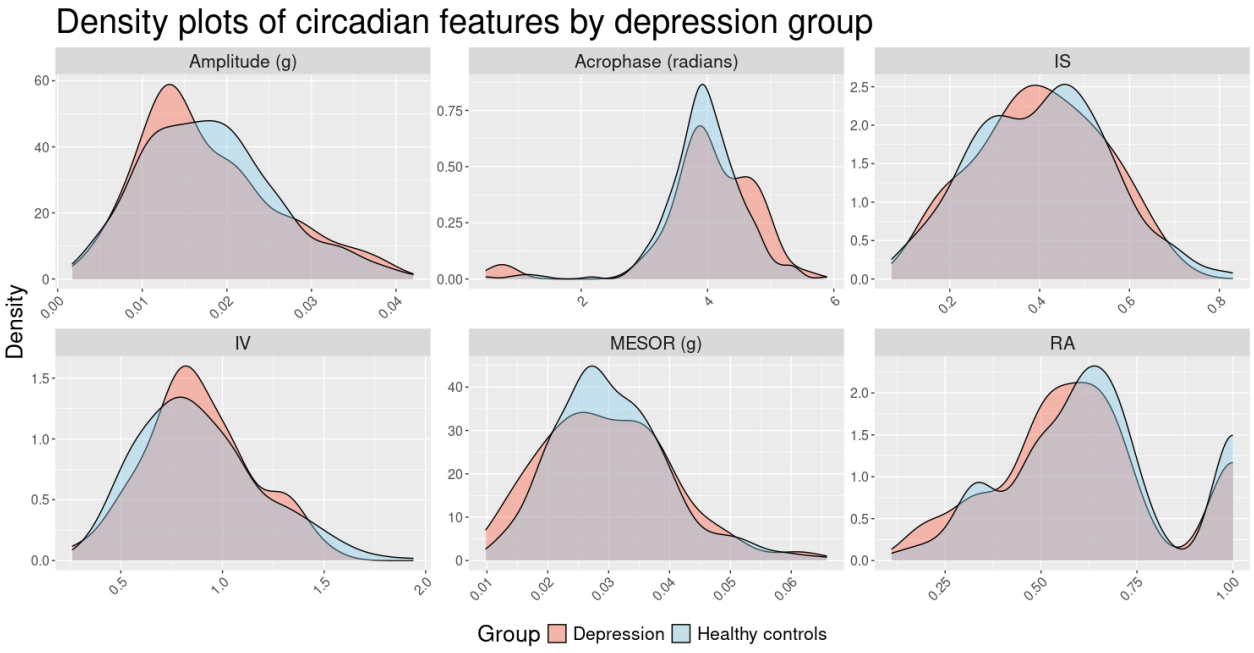


Fig. 10: *Density plots of the circadian features by depression group. Variables include the amplitude, the acrophase (in radians), the IS (Interdaily Stability), the IV (Intradaily Variability), the MESOR (Midline Estimated Statistic Of Rhythm), and the RA (Relative Amplitude). The depression group is shown in red, and the healthy controls are shown in blue. The y-axis scale is not fixed.*

TABLE V: Significance test of covariate parameters between participants with depression and healthy controls

Variable	**Test**	**Test statistic**	p
BMI	Chi-squared	3.96	0.27
Education	Chi-squared	10.43	0.001 (*)
Employment	Chi-squared	13.52	0.0002 (*)
Sex	Chi-squared	0.30	0.58
Smoking behaviour	Chi-squared	0.88	0.35
Household size	Chi-squared	6.93	0.03 (*)
Alcohol consumption	Chi-squared	21.83	0.00002 (*)
Race	Chi-squared	10.77	0.06
Age	t-test	-0.21	0.84

Statistical significance (*) = $p < 0.05$.

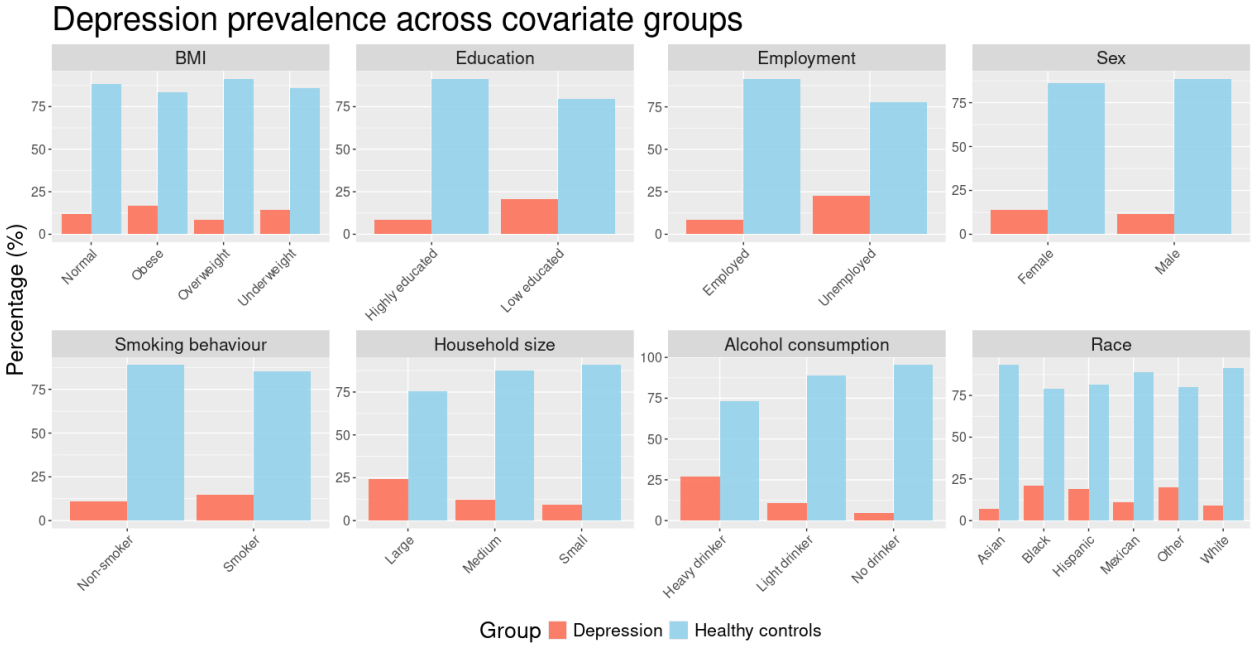


Fig. 11: *Bar plots of covariates by depression group. The depression group is shown in red, and the healthy controls are shown in blue. The y-axis scale is not fixed and shows the percentage of participants within each covariate group with and without depression.*

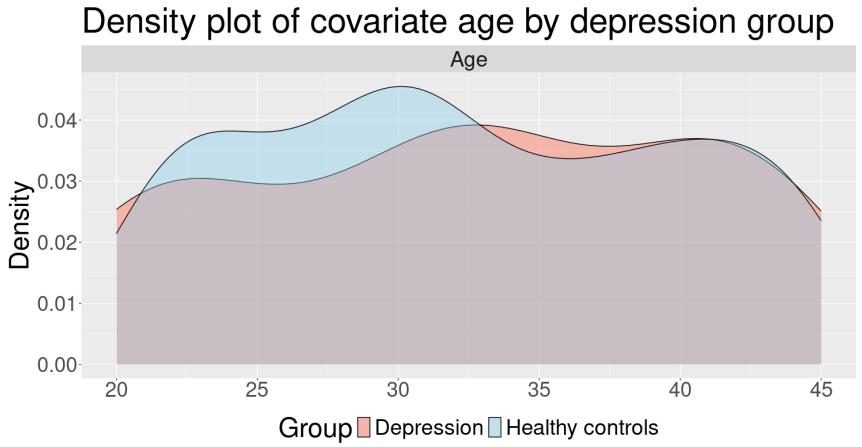


Fig. 12: *Density plot of the covariate age by depression group. The depression group is shown in red, and the healthy controls are shown in blue.*

2) Sleep disturbances

Table VI shows the circadian rhythm parameters of participants with sleep disturbances and healthy controls. Having sleep disturbances resulted in a lower MESOR, amplitude, IS, and RA, an earlier acrophase, and a higher IV. Only the amplitude and the IV were significantly different, looking at the p value. Cohen's d reflected a small effect size for the amplitude and the IV and negligible effect sizes for the other parameters. Fig. 13 visualizes this difference, with the amplitude being lower in the sleep disturbances group than in the healthy controls. At the same time, the IV was higher in the sleep disturbances group. As shown in Table VII, the bar plots in Fig. 14, and the density plot in Fig. 15, the distribution of the covariates varied considerably between participants with and without sleep disturbances. The incidence of sleep disturbances was relatively highest among older female participants of other racial backgrounds who were obese, unemployed, highly educated, had smoked, abstained from alcohol, and lived in smaller households. According to the table, the differences in household size, race, and age were significant.

TABLE VI: Circadian rhythm parameters of participants with sleep disturbance and healthy controls

Variable	**Sleep disturbances**		**Healthy controls**		p	**Cohen's** d
	Mean	SD	Mean	SD		
MESOR (g)	0.0291	0.00913	0.0309	0.00983	0.083	0.19
Acrophase (clock hour)	15:02:59	02:04:41	15:09:51	02:41:48	0.681	0.04
Amplitude (g)	0.0164	0.00688	0.0186	0.00813	0.006 (*)	0.29
IS	0.389	0.148	0.409	0.146	0.209	0.14
IV	0.929	0.279	0.865	0.304	0.046 (*)	-0.21
RA	0.631	0.233	0.641	0.218	0.697	0.04

Statistical significance (*) = $p < 0.05$, according to Welch's Two Sample t-test.
SD = Standard Deviation, MESOR = Midline Estimating Statistic Of Rhythm,
IV = Intradaily Variability, IS = Interdaily Stability, RA = Relative Amplitude.

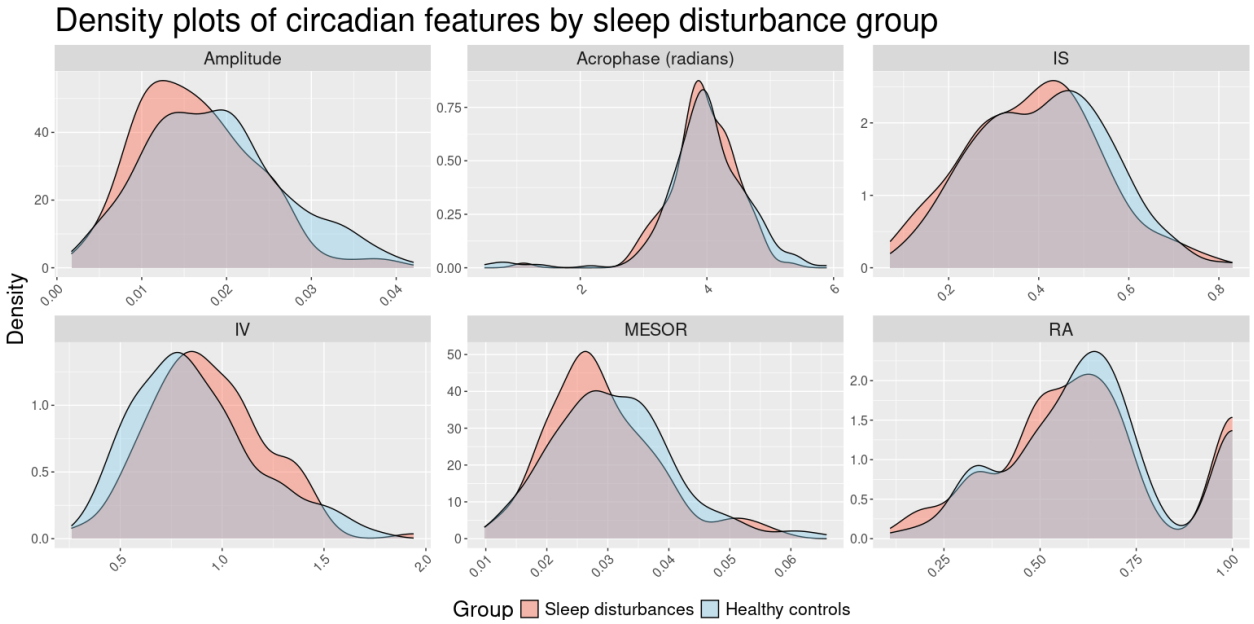


Fig. 13: *Density plots of the circadian features, by sleep disturbance group. Variables include the amplitude, the acrophase (in radians), the IS (Interdaily Stability), the IV (Intradaily Variability), the MESOR (Midline Estimated Statistic Of Rhythm), and the RA (Relative Amplitude). The sleep disturbances group is shown in red, and the healthy controls are shown in blue. The y-axis scale is not fixed.*

TABLE VII: Significance test of covariate parameters between participants with sleep disturbances and healthy controls

Variable	**Test**	**Test statistic**	p
BMI	Chi-squared	3.84	0.27
Education	Chi-squared	0.07	0.79
Employment	Chi-squared	3.41	0.06
Sex	Chi-squared	0.65	0.42
Smoking behaviour	Chi-squared	3.19	0.07
Household size	Chi-squared	10.0	0.007 (*)
Alcohol consumption	Chi-squared	0.81	0.67
Race	Chi-squared	23.0	0.0003 (*)
Age	t-test	-3.17	0.002 (*)

Statistical significance (*) = $p < 0.05$.

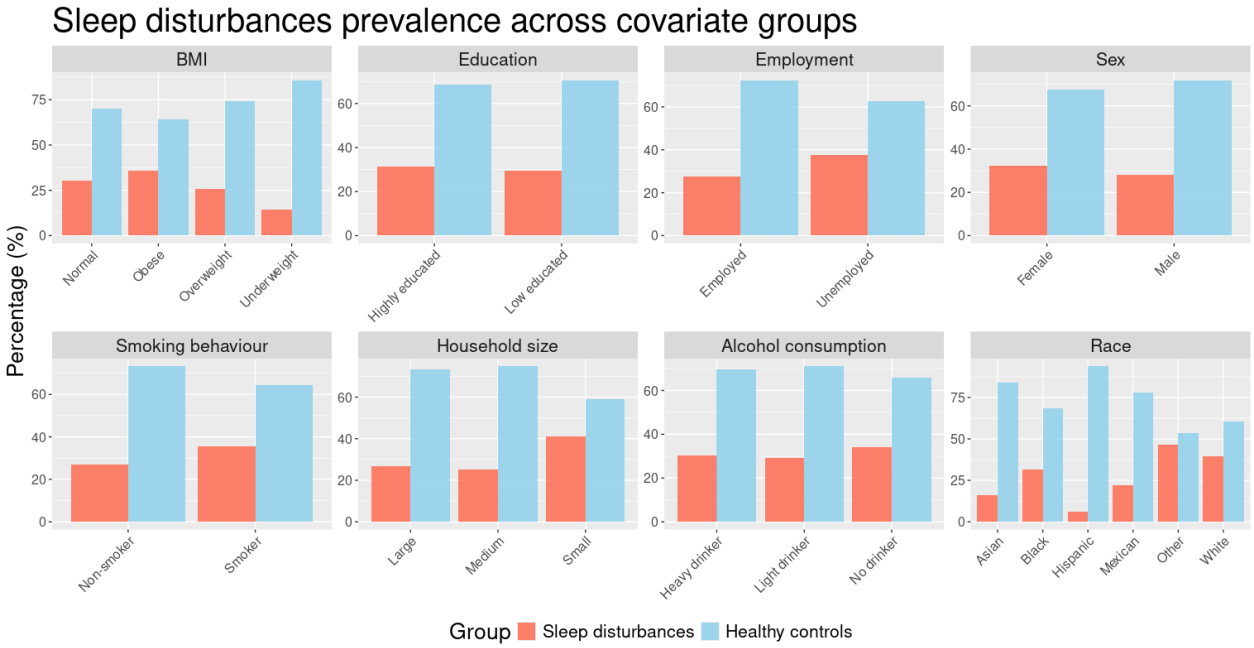


Fig. 14: *Bar plots of covariates by sleep disturbance group. The sleep disturbance group is shown in red, and the healthy controls are shown in blue. The y-axis scale is not fixed and shows the percentage of participants within each covariate group with and without sleep disturbances.*

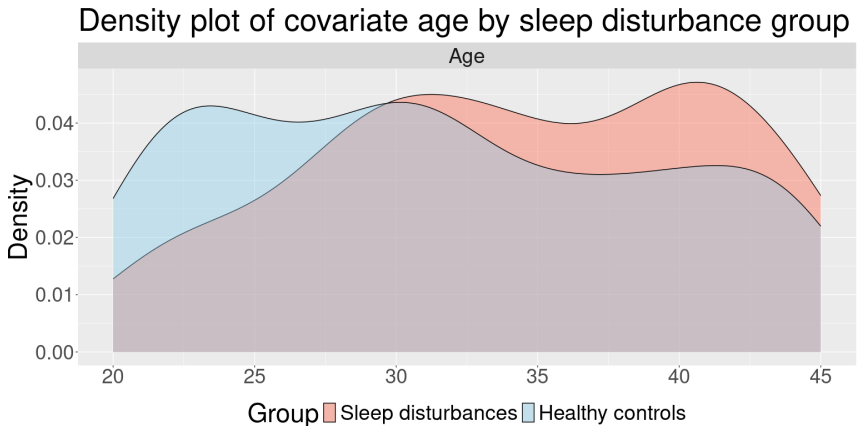


Fig. 15: *Density plot of the covariate age by sleep disturbance group. The sleep disturbance group is shown in red, and the healthy controls are shown in blue.*

C. Performance of predictive machine learning models

Fig. 16 displays the cross-validated performance of the trained machine learning models predicting depression, based on the metrics accuracy, F1-score, precision, recall, and specificity. From this figure, it is clear that the k-nearest neighbors models had the best recall, while the random forest models had the best accuracy, F1-score, precision, and specificity. The cross-validated performance of the trained machine learning models predicting sleep disturbances is highlighted in Fig. 17, based on the same metrics. As shown in the figure, the k-nearest neighbors models had the best F1-score and recall, the random forest models had the best accuracy and specificity, and the precision was similar across the models.

For each outcome and model type, the two best-performing models were chosen for further analysis. One of the best-performing models included the covariates in the preprocessing steps, and the other only included the circadian rhythm features. The best-performing models were chosen based on the F1-score and secondarily on the Area Under the Curve (AUC) value. The F1-score, accuracy, specificity, precision, recall, true positives, true negatives, false positives, false negatives, and AUC values for the best-performing models on the test datasets are presented in Table VIII.

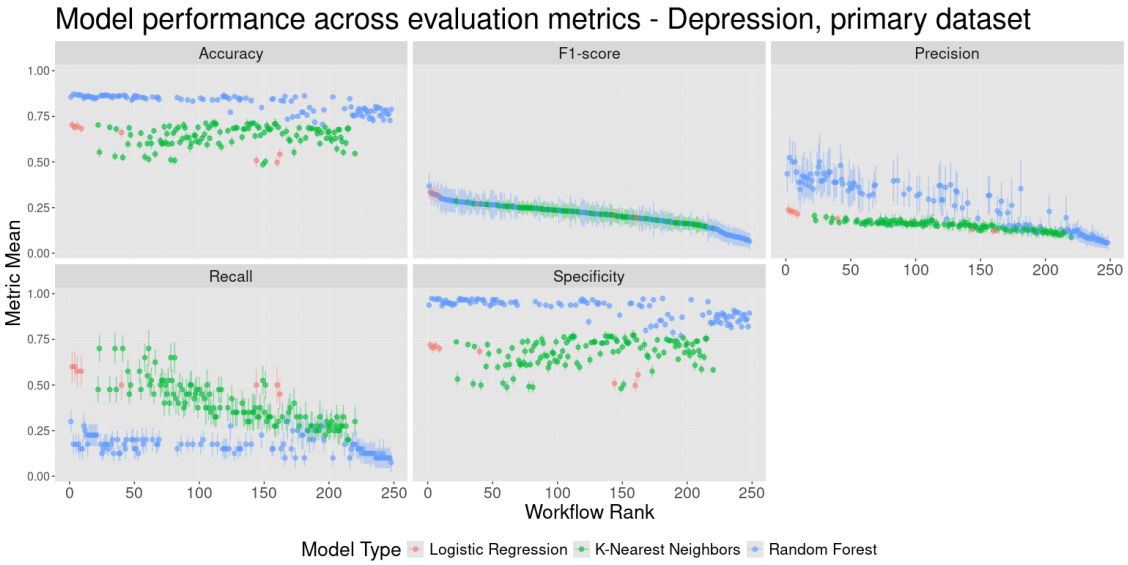


Fig. 16: *Cross-validated performance of the trained machine learning models predicting depression based on five evaluation metrics: accuracy, F1-score, precision, recall, and specificity. Shown in blue are the random forest models, in green the k-nearest neighbors models, and in red the logistic regression models. The x-axis shows the rank of the model, based on the F1-score. The y-axis shows the metric score, which ranges from 0.00 to 1.00.*

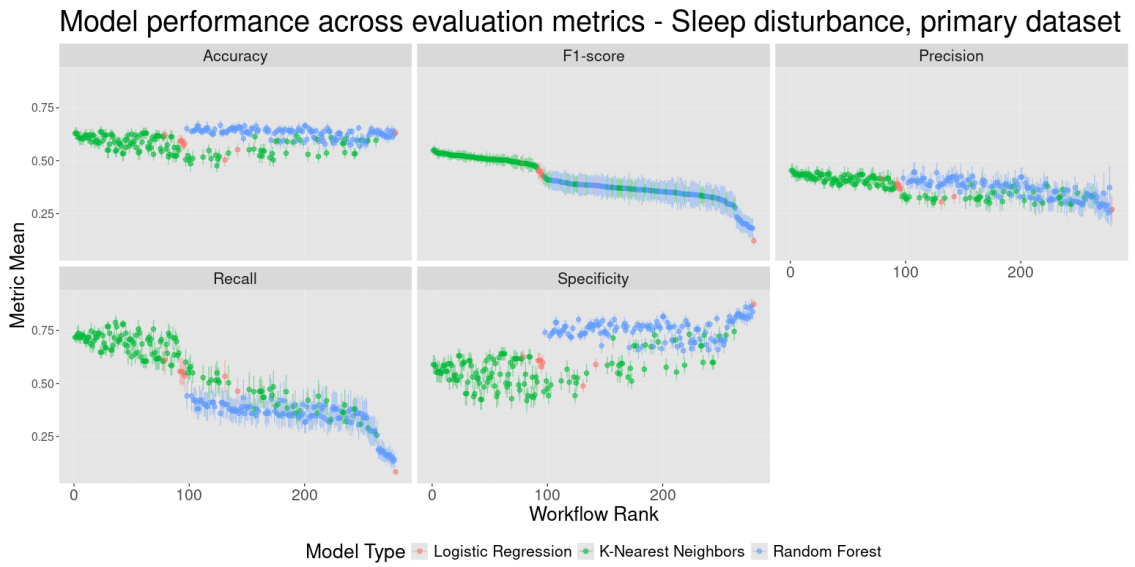


Fig. 17: *Cross-validated performance of the trained machine learning models predicting sleep disturbances based on five evaluation metrics: accuracy, F1-score, precision, recall, and specificity. Shown in blue are the random forest models, in green the k-nearest neighbors models, and in red the logistic regression models. The x-axis shows the rank of the model, based on the F1-score. The y-axis shows the metric score, which normally ranges from 0.00 to 1.00.*

TABLE VIII: Performance of the best-performing models on the test datasets for depression and sleep disturbances outcomes, based on key evaluation metrics, per model type and the inclusion/exclusion of covariates

Key performance metrics	**Random forest**		**K-nearest neighbors**		**Logistic regression**	
	Included	Excluded	Included	Excluded	Included	Excluded
Depression						
F1-score	0.267	0.308	**0.364**	0.190	0.278	0.205
Accuracy	**0.861**	0.772	0.646	0.570	0.671	0.608
Specificity	**0.957**	0.826	0.623	0.594	0.696	0.638
Precision	**0.4**	0.25	0.235	0.125	0.192	0.138
Recall	0.2	0.4	**0.8**	0.4	0.5	0.4
True positives	2	4	**8**	4	5	4
True negatives	**66**	58	43	41	48	44
False positives	3	11	26	28	21	25
False negatives	8	6	2	6	5	6
AUC	0.72	0.50	**0.81**	0.47	0.68	0.41
Sleep disturbances						
F1-score	0.327	0.227	0.375	0.269	**0.431**	0.366
Accuracy	0.532	**0.570**	0.494	0.519	0.532	0.430
Specificity	0.6	**0.727**	0.491	0.618	0.509	0.382
Precision	0.290	0.25	0.3	0.25	**0.341**	0.277
Recall	0.375	0.208	0.5	0.292	**0.583**	0.542
True positives	9	9	14	13	14	**15**
True negatives	34	**35**	28	24	28	26
False positives	21	15	27	31	27	29
False negatives	15	20	10	11	10	9
AUC	0.47	0.45	0.53	0.45	**0.54**	0.50

AUC = Area Under the Curve

Values in **bold** represent the best performance for each metric across models.

1) Depression

All analysed machine learning models that predicted depression used a preprocessing step that created example samples of depression on a 1:1 ratio. The following models had additional preprocessing steps:

- **Random forest, covariates:** Interaction between IS and IV.
- **Random forest, no covariates:** Interaction between MESOR and amplitude.
- **Logistic regression, no covariates:** Interaction between IS and IV & interaction between MESOR and amplitude.
- **K-nearest neighbors, covariates:** Interaction between alcohol consumption and smoking.
- **K-nearest neighbors, no covariates:** Interaction between MESOR and amplitude.

Among the evaluated models, the random forest model that included covariates showed the highest accuracy (0.861), specificity (0.957), precision (0.4), and true negatives (66-3) at the expense of the recall (0.2) and true positives (2-8). On the other hand, the covariate-inclusive k-nearest neighbors model showed the highest F1-score (0.364), recall (0.8), true positives (8-2), and AUC (0.81). Since the covariate-inclusive k-nearest neighbors model had the highest F1-score and AUC, this model was seen as the best-performing model and was included in the final analysis for predicting depression. However, since k-nearest neighbors models treat all variables equally in distance calculations and cannot indicate which features are more important, the random forest model was also included in the final analysis to look at variable importance, as it had the second-highest AUC (0.72) and the highest accuracy [68].

2) Sleep disturbances

The analysed machine learning models that predicted sleep disturbances had the following preprocessing steps:

- **Random forest, covariates:** Creating example samples of sleep disturbances on a 0.6:1 ratio & interaction between IS and IV.
- **Random forest, no covariates:** Creating example samples of sleep disturbances on a 1:1 ratio & interaction between IS and IV & interaction between MESOR and amplitude.
- **Logistic regression, covariates:** Creating example samples of sleep disturbances on a 1:1 ratio.
- **Logistic regression, no covariates:** Creating example samples of sleep disturbances on a 1:1 ratio.
- **K-nearest neighbors, covariates:** Creating example samples of sleep disturbances on a 1:1 ratio & interaction between IS and IV.
- **K-nearest neighbors, no covariates:** Creating example samples of sleep disturbances on a 1:1 ratio & interaction between IS and IV & interaction between MESOR and amplitude.

Among the evaluated models, the random forest model that excluded covariates showed the highest accuracy (0.570), specificity (0.727), and true negatives (35-15), at the expense of the recall (0.208) and AUC (0.45), which means this model performed worse than chance. On the other hand, the covariate-inclusive logistic regression model showed the highest F1-score (0.431), precision (0.341), recall (0.583), and AUC (0.54). Since the covariate-inclusive logistic regression model had the highest F1-score and AUC, this model was seen as the best-performing model and included in the final analysis for predicting sleep disturbances.

D. Variable importance

For both depression and sleep disturbances, it was determined which predictor(s) had the most influence on the final prediction of the outcome (i.e., depression or sleep disturbances) in the test datasets. This was visualised by using variable importance plots. Only plots of the best-performing machine learning models were visualised for both depression and sleep disturbances, as can be seen in Fig. 18. For the best-performing model predicting depression, the random forest model was used instead, since, as previously explained, these plots could not be made for k-nearest neighbors models.

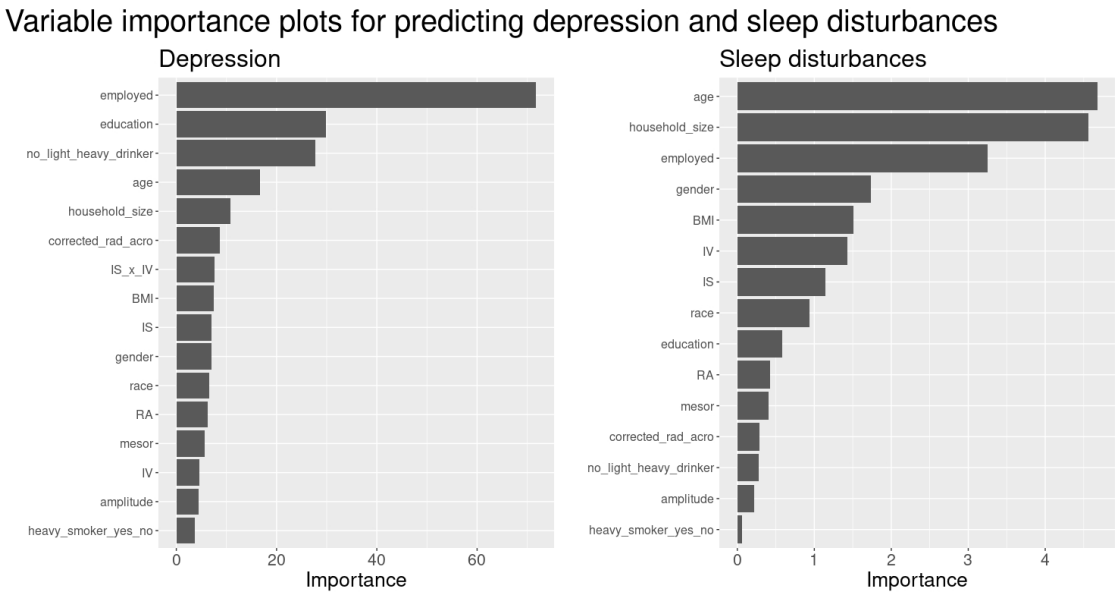


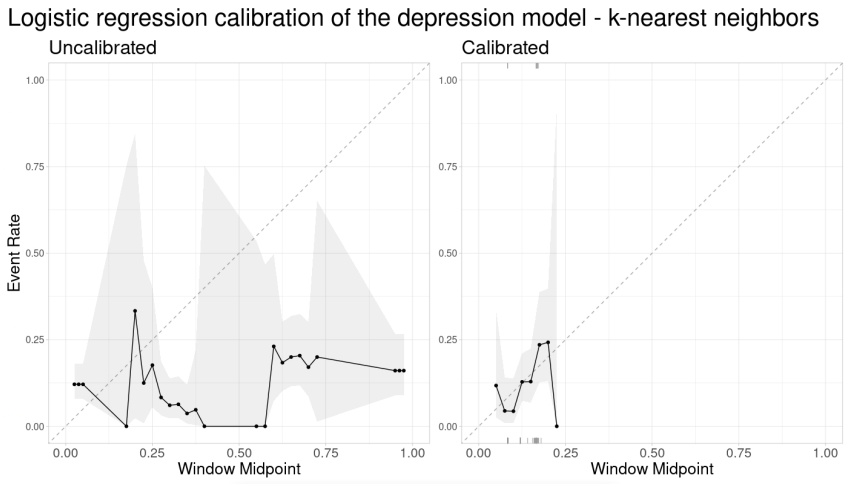
Fig. 18: *Variable importance plots of 2 different machine learning models, one predicting depression and the other sleep disturbances, in the test datasets. The figure on the left shows the variable importance plot of the machine learning model predicting depression. The model used is the covariate-inclusive random forest model. The figure on the right shows the variable importance plot of the machine learning model predicting sleep disturbances. The model used is the covariate-inclusive logistic regression model. The predictors (i.e., variables) are ranked from most to least important, with the most important variables at the top. IS = Interdaily Stability, IV = Intradaily Variability, RA, = Relative Amplitude, BMI = Body Mass Index, MESOR = Midline Estimated Statistic Of Rhythm.*

The variable importance plot of the covariate-inclusive random forest model predicting depression shows that the prediction relied mainly on the covariate predictors, especially employment. The most important circadian predictors were the acrophase, the interaction term between IS and IV, and the IS on its own. Since the interaction term IS x IV does not have more influence than the IS and IV individually combined, there is no significant interaction effect beyond their sum. From the predictors, smoking and amplitude had the least impact.

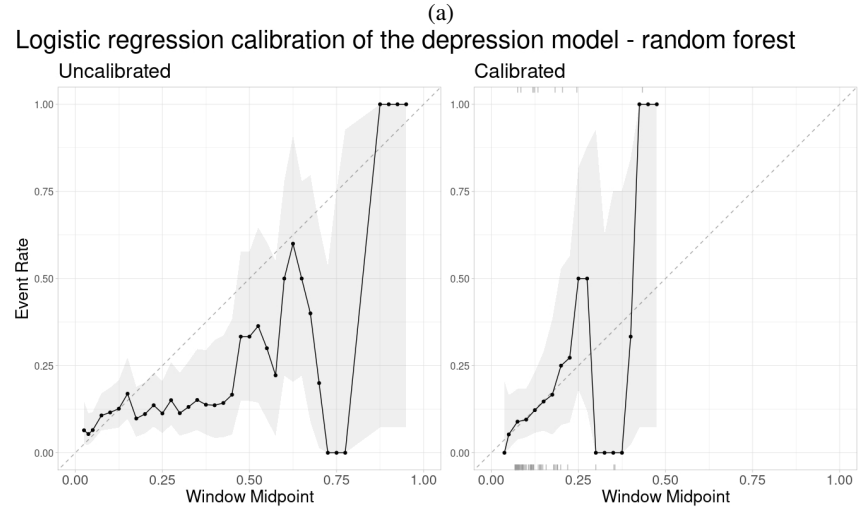
The variable importance plot of the covariate-inclusive logistic forest model predicting sleep disturbance reveals that the final prediction was mainly based on age, household size, and employment. The IV had the most influence from the circadian predictors, followed closely by the IS. Smoking had almost no impact, followed by the circadian predictor amplitude.

E. Calibration of the best performing models

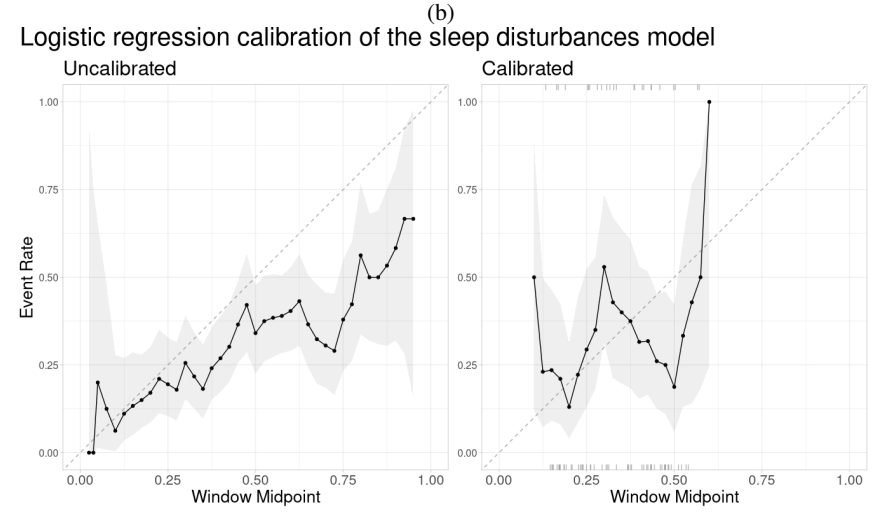
The models were calibrated to improve the best-performing models predicting depression and sleep disturbances. Fig. 19a compares the uncalibrated and calibrated plots of the covariate-including k-nearest neighbors model predicting depression. The figure on the left illustrates that the predicted probabilities did not show good accuracy, as the event rate fell below the diagonal line, indicating that the model was over-predicting probabilities. Furthermore, the curve does not reach an event rate of 1.00. The figure on the right shows improved accuracy. However, the model does not predict probabilities higher than 0.25, and the event rate remains below 0.25. Fig. 19b compares the uncalibrated and calibrated plots of the covariate-inclusive random forest model that predicts depression. The figure on the left shows that the model was over-predicting the probabilities. At the same time, it is visible in the figure on the right that the event rate was on the diagonal line until a window midpoint of around 0.20. The calibrated model did not predict probabilities above 0.50. Lastly, Fig. 19c compares the uncalibrated and calibrated plots of the covariate-inclusive logistic regression model predicting sleep disturbances. Although the figure on the left depicts a small over-prediction of the probabilities and the curve does not reach an event rate of 1.00, the model does seem to follow the diagonal line. However, the calibration curve (which does reach an event rate of 1.00) did not follow the diagonal line and swapped between over-predicting and under-predicting the probabilities. These suggest that logistic regression calibration was insufficient in improving the best-performing models. The performance metrics of the uncalibrated and calibrated models were the same as the ones shown in Table VIII.



(a)



(b)



(c)

Fig. 19: *Uncalibrated and calibrated plots of the best performing models. The figures on the left show the uncalibrated plots, and the figures on the right show the logistic regression-calibrated plots. The x-axis shows the window midpoint (model's predicted probability), and the y-axis shows the event rate (fraction of positive cases).* ***(a)*** *Depression (k-nearest neighbors model).* ***(b)*** *Depression (random forest).* ***(c)*** *Sleep disturbances.*

F. Improvement of best performing models by removing equivocal values

Removing equivocal values improved the performance of the best-performing models predicting depression or sleep disturbances, as shown in Table IX. For the k-nearest neighbors model predicting depression, all key performance metrics improved except recall, with the reportable rate decreasing to 0.873. Removing equivocal zones in the random forest model improved the F1-score, accuracy, and recall, while the specificity and precision worsened. The reportable rate decreased to 0.73. The machine learning model predicting sleep disturbances achieved improvements in F1-score, precision, and recall after removing equivocal zones, although the accuracy and specificity decreased. The reportable rate remained at 1. For both outcomes, the AUC remained unchanged before and after the removal of equivocal zones.

TABLE IX: Performance metrics comparison before and after the removal of equivocal zones for the best performing models

Model	F1-score	Accuracy	Specificity	Precision	Recall	AUC
Depression, original (K-nn)	0.364	0.646	0.623	0.235	0.8	0.81
Depression, after removal (K-nn)	0.471	0.739	0.729	0.333	0.8	0.81
Depression, original (RF)	0.267	0.861	0.957	0.4	0.2	0.72
Depression, after removal (RF)	0.462	0.879	0.906	0.375	0.6	0.72
Sleep disturbances, original	0.431	0.532	0.509	0.341	0.583	0.54
Sleep disturbances, after removal	0.466	0.506	0.418	0.347	0.708	0.54

AUC = Area Under the Curve, K-nn = K-nearest neighbors, RF = Random Forest

G. Sensitivity analysis

Results from the sensitivity cohort differed from the primary cohort. Table X presents the performance metrics of the best-performing models predicting depression and sleep disturbances in the sensitivity cohort, both before and after the removal of equivocal zones. The entire sensitivity analysis can be found in Appendix C.

Initially, the best-performing models predicting depression showed comparable performance across the primary and sensitivity cohorts, looking at the key metrics. After the removal of equivocal zones, the F1-score and recall were notably higher in the primary cohort compared to the sensitivity cohort (0.462/0.471-0.364 and 0.6/0.8-0.25, respectively). In contrast, the accuracy (0.879/0.739-0.887), specificity (0.906/0.729-0.981), and precision (0.375/0.333-0.667) were all higher in the sensitivity cohort after the removal of equivocal zones. When examining the best-performing models predicting sleep disturbances before the removal of equivocal zones, the sensitivity cohort outperformed the primary cohort across all key metrics except recall (e.g., F1-score (0.458-0.431), AUC (0.60-0.54), and accuracy (0.639-0.532)). After the removal of equivocal zones, the sensitivity cohort continued to show better performance across all key metrics except specificity (e.g., F1-score (0.581-0.466) and accuracy (0.536-0.506)). For both outcomes, the AUC remained the same before and after the removal of equivocal zones.

Given that the best model predicting depression showed no substantial differences in key metrics between the primary and sensitivity cohorts before the removal of equivocal zones, the models from the primary cohort demonstrated robustness against the inclusion of extremely inconsistent (IS) and fragmented (IV) values. Moreover, after the removal of equivocal zones, model performance within the primary cohort improved more than in the sensitivity cohort, especially when looking at the F1-score (0.462/0.471-0.364) and the recall (0.6/0.8-0.25). In contrast, the model predicting sleep disturbances performed better in the sensitivity cohort, both before and after the removal of equivocal zones, especially when looking at the F1-score (0.581-0.466), which indicates that the sensitivity cohort had added value for model performance on sleep disturbances.

TABLE X: Performance metrics comparison before and after the removal of equivocal zones for the best performing models (sensitivity cohort)

Model	F1-score	Accuracy	Specificity	Precision	Recall	AUC
Depression, original	0.267	0.847	0.952	0.4	0.2	0.74
Depression, after removal	0.364	0.887	0.981	0.667	0.25	0.74
Sleep disturbances, original	0.458	0.639	0.7	0.423	0.5	0.60
Sleep disturbances, after removal	0.581	0.536	0.333	0.429	0.9	0.60

AUC = Area Under the Curve

IV. Discussion

A. Summary of key findings

During this study, several machine learning models were investigated to explore patterns in accelerometer data (circadian rhythm features) and covariate data to predict sleep disturbances and depression. For depression, the covariate-including k-nearest neighbors model performed the best across most key performance metrics (e.g., F1-score = 0.364, AUC = 0.81). After removing equivocal zones, all key performance metrics improved, with the F1-score increasing from 0.364 to 0.471 and the accuracy from 0.646 to 0.739. When comparing models that included covariates to those that excluded them, it is clear that the covariate-inclusive models performed better. This pattern is also evident in the variable importance plot, where covariate variables are ranked highest and circadian features are among the lowest. Sensitivity analysis showed no notable effect on the depression predictions: the best-performing model achieved an F1-score of 0.267, which improved to 0.364 after removing equivocal zones, with an AUC of 0.74.

In the case of sleep disturbances, significant differences were observed in amplitude and IV between the sleep disturbances group and the healthy controls. The covariate-inclusive logistic regression performed the best, achieving the highest F1-score (0.431) and AUC (0.54). The F1-score (0.466), precision, and recall improved after removing equivocal zones, although this came at the expense of the other key performance metrics. Overall, the key performance metrics were similar across all models, regardless of whether covariates were included. However, the variable importance plot again showed that the covariates were the most important predictors. In this case, the sensitivity analysis proved useful. It initially improved nearly all key performance metrics except recall. After removing equivocal zones, all metrics showed larger improvements than those from the primary cohort, except for accuracy and specificity. Still, the sensitivity cohort outperformed the primary cohort on all key performance metrics except specificity.

Model calibration did not enhance the reliability of the predicted probabilities, as the calibration curves still showed substantial deviation from the diagonal line. Overall, accelerometer-derived circadian rhythm features alone offered limited predictive power for both depression and sleep disturbances. When including covariate features, however, model performance improved considerably, especially for predicting depression.

B. Interpretation and comparison with existing literature

One of the primary research questions addressed in this study was *"To what extent is there an association between circadian rhythm disruptions and the risk of depression?"*, which was hypothesised as a significant association. Although no clear association between patterns in accelerometer data (i.e., the circadian rhythm parameters) and the presence of depression was found (as all AUC values from the covariate-excluding models were below 0.50), several steps were taken to mitigate this limitation. These steps included the inclusion of covariates, the testing of multiple classification algorithms, the use of oversampling techniques and predictor interactions, and the use of a sensitivity analysis. Except for the sensitivity analysis, these measures improved model performance, indicating that combining circadian rhythm features with covariates, and having interacting terms between several circadian rhythm features (e.g., IS and IV) could still hold some predictive value. Furthermore, the F1-score of the covariate-exclusive random forest model (0.308) was higher than that of the covariate-inclusive random forest model (0.267), indicating that circadian rhythm features still hold meaningful predictive value. Including covariates in the models improved the AUC values, which also applied to the other key metrics, except for recall. The variable importance plot of the best-performing model also confirmed that the covariates improved predictions: all circadian rhythm predictors had relatively little influence on the prediction of depression, and the prediction relied almost entirely on covariate predictors. Moreover, no significant differences between the depression group and healthy control group were observed in terms of circadian rhythm parameters. In contrast, significant differences were found in education, employment, household size, alcohol consumption, and race. These findings contrast with previous studies. For example, a similar UK Biobank study found that a 20% reduction in RA was associated with an increased risk of depression and other mood disorders [3]. Another actigraphy-using research, based in the Netherlands, has shown that having depression leads to lower MESOR and lower amplitude, but that it does not have any effect on the acrophase [69]. While the present study's findings somewhat support those conclusions, the differences observed in this study were not statistically significant and had little impact on the machine learning models. Finally, a study by Wescott et al. has shown that there is no significant difference in IV, IS, and RA between participants with seasonal affective disorder or subsyndromal seasonal affective disorder and healthy controls [70]. Similarly, the present study

found no significant differences between the two groups across the same key performance metrics. Therefore, the findings of this study support the findings reported by Wescott et al.

The other primary research question explored in this study was *"To what extent are circadian rhythm disruptions directly associated with sleep disturbances?"*. Significant differences were found in amplitude and IV between participants with sleep disturbances and healthy controls. This significance also extends to the MESOR in the sensitivity cohort. For the covariates, significance was found in household size, race, and age. Key performance metrics of the covariate-exclusive models were comparable to those of the covariate-inclusive models. While the covariate-exclusive models showed higher accuracies, the covariate-inclusive models showed higher F1-scores and AUCs. However, the AUC was at or below 0.50 for all covariate-exclusive models, which indicates that those models performed worse than random chance and could not distinguish between participants with sleep disturbances and healthy controls [64]. Moreover, the covariate-inclusive models did not achieve an AUC higher than 0.54, indicating poor discriminatory ability. All models selected for the initial analysis included either no interactions or interactions between circadian rhythm features (e.g., MESOR and amplitude), demonstrating that the circadian rhythm features do somewhat contribute to the predictive value of the models, with the covariate-exclusive random forest model (which had interactions between the IS and IV, and MESOR and amplitude) even showing the highest accuracy (0.570), specificity (0.727), and true negatives (35-20). The significant differences in specific circadian rhythm parameters and the similar key performance metrics scores of models with and without covariates suggest that circadian rhythm disruptions influence sleep disturbances. However, the variable importance plot showed that all circadian features barely influenced the prediction. Key performance metrics barely improved after removing the equivocal zones, while the accuracy and specificity declined. Sensitivity analysis proved helpful, as the best model from the sensitivity analysis reached an AUC of 0.60, which, while considered "poor", indicates a small predictive value [64]. Moreover, the F1-score was above 0.5 (0.581), which, although suboptimal, still indicated a small predictive value of the model [71]. However, the variable importance plots from the sensitivity analysis revealed that the model's predictive power mainly came from covariate features and not from circadian rhythm features, apart from a slight influence from the acrophase. These results partially support findings from other studies, with a study from Fossion et al. finding significant differences in acrophase and no significant differences in MESOR, amplitude, and IV between participants with acute insomnia and the control group [72]. In contrast, a study by Zhao et al. found that the circadian amplitude is associated with sleep disturbances, which supports the findings in the present study, as there were significant differences in amplitude between the sleep disturbances group and the healthy controls [73].

The secondary question explored in this study was *"Which machine learning model can best uncover patterns in the data associated with depression, and how do covariates affect these patterns?*". After exploring three types of machine learning algorithms (random forest, logistic regression, and k-nearest neighbors), the random forest models consistently showed the highest accuracy, specificity, and precision. In contrast, the logistic regression and k-nearest neighbors models had the highest F1-scores and recall. The random forest and k-nearest neighbors models had similar AUC values, with those of the logistic regression models being lower. For classifying depression, it was essential to look at the F1-score since this study dealt with imbalanced datasets in which depression was the minority class. The F1-score gives a balanced performance of the model and takes false negatives into account [71]. Among the covariate-exclusive models, the random forest model achieved the best F1-score and AUC, while for the covariate-inclusive models, the k-nearest neighbors model performed the best. Covariates generally improved key performance metrics across models and were the most important predictors in the variable importance plots.

C. Strengths and limitations of this study

Although no formal procedures were applied to assess bias, potential limitations such as selection/information bias or misclassification should be considered when interpreting the results. For example, this study only included participants between 20 and 45 years old, a relatively healthy and active age group. This might have led to an underestimation of circadian disruptions. Additionally, participants with one or more data quality flags or fewer than four days with more than 16 hours of actigraphy data were excluded, resulting in the removal of around 4.000 additional participants. This could have introduced selection bias and reduced the generalisability of the findings. Furthermore, the data available per participant (4-7 days) may have been insufficient to determine circadian rhythm features reliably. The limitation becomes even more relevant since the first and last days of data were excluded from the circadian rhythm parameters extraction.

There is also a high risk of information bias, since all questionnaires are self-reported, which raises concerns about the reliability of the covariate predictors and outcomes used. Covariates, depressive symptoms, and sleep disorders were all self-reported at a single point in time rather than being a clinical diagnosis or a follow-up assessment, making them susceptible to momentary emotional states. This could have introduced misclassification in the study, as having sleep disturbances and/or heightened depressive symptoms was dependent on how participants felt on the day of taking the questionnaires. Moreover, in this study, depression was seen as a simple yes/no classification, whereas in reality, depression is a spectrum of severity (mild, moderate, severe), and there are multiple depressive subgroupings, such as atypical depression or seasonal affective disorder [74].

In total, 392 participants were included in the final analysis. In retrospect, this sample size was insufficient for the machine learning models to reliably detect patterns in the data. This limitation became even more apparent due to the class imbalance in the dataset for both outcomes: 9.4% of the primary cohort had heightened depressive symptoms, and 28.2% of the primary cohort had sleep disturbances. This resulted in a class imbalance of 1:9.6 and 1:1.9, respectively. Although the prevalence of depression in the dataset is comparable to that from the United States (8.3%), the absolute number of participants with heightened depressive symptoms and/or sleep disturbances in the dataset is still minimal, respectively 50 and 120 [75]. Although the machine learning models predicting depression had higher accuracies than those predicting sleep disturbances, the F1-score was higher on average in the models predicting the latter. This suggests that relatively few participants were correctly classified as having heightened depressive symptoms. The depression test set of the primary cohort only had 10 participants with depression and 69 healthy controls, a ratio of 1:6.9. This imbalance could have caused unreliable classification results and not meaningful key performance metrics. Notably, the accuracy and AUC tend to overestimate performance on imbalanced datasets [76]. Although Synthetic Minority Oversampling Technique (SMOTE) was applied to address class imbalance by oversampling the minority class, prior research has shown that it can lead to poor model calibration and may degrade model performance [77].

This study utilized a cross-sectional dataset, making it impossible to look at causality. It is not possible to determine whether disruptions in the circadian rhythm ultimately lead to depression and sleep disturbances. In cross-sectional research, other measures may have influenced the association between circadian rhythm disruptions and depression or sleep disturbances during data collection. Therefore, it cannot be assumed that only circadian rhythm disruptions cause the two outcomes. Given the chosen age range of the study population (20-45 years old), the chance of other mental disorders playing a role is high, as the average age of onset for conditions such as PTSD, mood disorders, anxiety disorders, bipolar disorders, and panic disorders also falls within this range [22]. Furthermore, the prevalence of mental health disorders is highest among people aged 18-25 (36.2%), followed by those aged 26-49 (29.4%), and lowest among older adults (13.9%) [78]. This means that participants might have circadian rhythm disruptions, sleep disturbances, or depression due to other mental health disorders, especially since depression is comorbid with other mental health disorders such as anxiety and panic disorders [75]. These other mental health disorders were not included in this study as covariates, since mental health disorder diagnoses were not available in the 2011-2012 NHANES cycle.

Despite its limitations, this study also has several strengths. Actigraphy data were objectively collected to determine circadian rhythm disruptions. As this method was the same for every participant, there was no risk of information bias here. Furthermore, by excluding participants with missing covariate data, no imputation was necessary, and there was no risk of issues with missing data. A sensitivity analysis was conducted to assess the robustness of the machine learning models when also including participants with extremely fragmented (IV $>$ 95th percentile) or inconsistent circadian rhythms (IS $<$ 5th percentile). The results showed that, in particular, the machine learning models predicting depression were robust to including those participants.

Another strength of this study is the use of multiple machine learning algorithms, which were compared to each other before selecting the best one for further analysis. This approach ensured a transparent decision regarding the choice of models. In addition, using multiple oversampling ratios and including variable interactions for each algorithm contributed to a robust model selection process. Both covariate-inclusive and covariate-exclusive models were used, which provided valuable insights into the utility of the different models. Variable importance plots further highlighted the most important predictors, showing which predictors the particular model relied on the most. Moreover, the use of cross-validation prevented overfitting, and the use of both a train-test split and the removal of equivocal zones both improved key performance metrics. Finally, calibrating the best-performing models by applying logistic

calibration ensured that the predicted probabilities better reflect the actual event rates, although the calibration results were not ideal.

D. Future research directions

To better understand the causal relationship between circadian rhythm disruptions and depression/sleep disturbances, future research should implement a longitudinal (i.e., follow-up) study design. With such a study design, it could be researched whether circadian disruptions precede the onset of depression and sleep disturbances. As follow-up data is unavailable with data from the NHANES, other datasets such as those from the UK Biobank should be looked at instead.

Additionally, future research should include a larger sample size in the analysis to improve model robustness and generalisability. One approach could be to increase the age range to (at least) 65, better representing the American population. Another approach is not to exclude participants with missing covariate data, and instead use imputation techniques on those participants. Finally, using multiple NHANES cycles would also increase the sample size.

As previously mentioned, participants were classified as having heightened depressive symptoms based on a single, self-reported PHQ-9 questionnaire. Although this questionnaire is a reliable and valid measure, it is not the same as a formal, clinical diagnosis [79]. Using clinical diagnoses rather than a single, self-reported questionnaire could improve the validity of the results. Future research should use a later NHANES cycle (for example, the 2013-2014 cycle), since mental health disorders were included. Besides resulting in a more reliable classification of depression, this dataset could also be used to include other mental health disorders, like anxiety and panic disorders, as covariates or as different possible outcomes. Moreover, for heightened depressive symptoms, gradations should be taken into account while classifying participants, instead of only using a "yes/no" binary classification. This may result in more visible differences between participants with no depression and participants on the depression spectrum. However, this approach is only viable if the number of participants is increased as well, since the class imbalance (between healthy controls and participants with any depression severity) would otherwise become even more pronounced.

Future research should include supplementary sleep diaries or questionnaires to validate the actigraphy data, which could provide a clearer picture of whether participants had experienced disrupted circadian rhythms on a day-to-day basis. As the NHANES does not provide such data, future research should utilize a different national survey or design a new study entirely. Additionally, future studies could incorporate salivary melatonin measurements alongside actigraphy data to validate circadian rhythms, as melatonin measurements are the gold standard for assessing circadian rhythms, and the use of saliva offers a less invasive alternative compared to the use of blood plasma [14, 16].

External validation through an external dataset should be implemented in future research. Validating the machine learning models on a different cohort would provide insights into their generalisability and the class imbalance in this study. Finally, future work should explore more machine learning algorithms, more oversampling ratios of the minority classes, and more interactions between the predictors. A different number of predictors could also be looked at instead of just the two options used in this study.

E. Conclusion

Using multiple machine learning models, preprocessing steps, and various circadian rhythm features, this study investigated whether circadian rhythm disruptions were associated with depression and/or sleep disturbances. However, no clear association was found between circadian rhythm disruptions and the risk of depression during this study. Similarly, the results also showed no strong association between circadian rhythm disruptions and sleep disturbances, apart from significant differences in amplitude and IV, as well as MESOR in the sensitivity cohort. Among the covariate-exclusive machine learning models, the random forest model showed the highest F1-score and AUC, while the k-nearest neighbors model showed the highest performance across key performance metrics when covariates were included. These two models were best at uncovering patterns in the data that are associated with depression. Key performance metrics generally improved with the inclusion of covariates and the exclusion of equivocal predictions. However, it is difficult to draw any binding conclusions from this research due to the small number of participants included in the analysis and the strong class imbalance concerning depression.

References

[1] Goodwin R. D. et al. “Trends in U.S. Depression Prevalence From 2015 to 2020: The Widening Treatment Gap”. In: *Am J Prev Med* 63.5 (2022), pp. 726–733. ISSN: 0749-3797. DOI: 10.1016/j.amepre.2022.05.014.

[2] Centers for Disease Control and Prevention. *Sleep Difficulties in Adults: United States, 2020*. URL: https://www.cdc.gov/nchs/products/databriefs/db436.htm. (accessed: 18.06.2025).

[3] Laura M Lyall et al. “Association of disrupted circadian rhythmicity with mood disorders, subjective well-being, and cognitive function: a cross-sectional study of 91-105 participants from the UK Biobank”. In: *The Lancet Psychiatry* 5.6 (2018), pp. 507–514. ISSN: 2215-0366. DOI: https://doi.org/10.1016/S2215-0366(18)30139-1. URL: https://www.sciencedirect.com/science/article/pii/S2215036618301391.

[4] Epstein L. *Why your sleep and wake cycles affect your mood*. URL: https://www.health.harvard.edu/blog/why-your-sleep-and-wake-cycles-affect-your-mood-2020051319792. (accessed: 18.06.2025).

[5] Cleveland Clinic. *Circadian Rhythm*. URL: https://my.clevelandclinic.org/health/articles/circadian-rhythm. (accessed: 24.04.2025).

[6] Yanyan Xu et al. “Blunted Rest–Activity Circadian Rhythm Is Associated With Increased Rate of Biological Aging: An Analysis of NHANES 2011–2014”. In: *The Journals of Gerontology: Series A* 78.3 (Sept. 2022), pp. 407–413. ISSN: 1758-535X. DOI: 10.1093/gerona/glac199. eprint: https://academic.oup.com/biomedgerontology/article-pdf/78/3/407/49378785/glac199.pdf. URL: https://doi.org/10.1093/gerona/glac199.

[7] Jinjoo Shim, Elgar Fleisch, and Filipe Barata. “Circadian rhythm analysis using wearable-based accelerometry as a digital biomarker of aging and healthspan”. In: *npj Digital Medicine* 7.1 (2024), p. 146. DOI: 10.1038/s41746-024-01111-x.

[8] A. R. Neves et al. “Circadian rhythm and disease: Relationship, new insights, and future perspectives”. In: *J Cell Physiol* 237.8 (2022), pp. 3239–3256. DOI: 10.1002/jcp.30815.

[9] Taylor Stowe and Colleen Mcclung. “How Does Chronobiology Contribute to the Development of Diseases in Later Life”. In: *Clinical Interventions in Aging* Volume 18 (Apr. 2023), pp. 655–666. DOI: 10.2147/CIA.S380436.

[10] YongMin Cho et al. “Effects of artificial light at night on human health: A literature review of observational and experimental studies applied to exposure assessment”. In: *Chronobiology International* 32.9 (2015). PMID: 26375320, pp. 1294–1310. DOI: 10.3109/07420528.2015.1073158. eprint: https://doi.org/10.3109/07420528.2015.1073158. URL: ttps://doi.org/10.3109/07420528.2015.1073158.

[11] Chi Nguyen et al. “In vivo molecular chronotyping, circadian misalignment, and high rates of depression in young adults”. In: *Journal of Affective Disorders* 250 (2019), pp. 425–431. ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2019.03.050. URL: https://www.sciencedirect.com/science/article/pii/S0165032718320299.

[12] Yundan Liao et al. “Associations between rest–activity/light-exposure rhythm characteristics and depression in United States adults: A population-based study”. In: *Journal of Affective Disorders* 369 (2025), pp. 1004–1012. ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2024.10.073. URL: https://www.sciencedirect.com/science/article/pii/S0165032724017658.

[13] Joanne S. Carpenter et al. “Actigraphy-derived circadian rhythms, sleep-wake patterns, and physical activity across clinical stages and pathophysiological subgroups in young people presenting for mental health care”. In: *Journal of Psychiatric Research* (2025). ISSN: 0022-3956. DOI: https://doi.org/10.1016/j.jpsychires.2025.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0022395625001530.

[14] Kathryn J. Reid. “Assessment of Circadian Rhythms”. In: *Neurologic Clinics* 37.3 (2019). Circadian Rhythm Disorders, pp. 505–526. ISSN: 0733-8619. DOI: https://doi.org/10.1016/j.ncl.2019.05.001. URL: https://www.sciencedirect.com/science/article/pii/S0733861919300404.

[15] H. Hymczak et al. “Core Temperature Measurement-Principles of Correct Measurement, Problems, and Complications”. In: *Int J Environ Res Public Health* 18.20 (2021). ISSN: 1660-4601. DOI: 10.3390/ijerph182010606.

[16] Antonio Almendros-Ruiz et al. “Melatonin Secretion and Impacts of Training and Match Schedules on Sleep Quality, Recovery, and Circadian Rhythms in Young Professional Football Players”. In: *Biomolecules* 15.5 (2025). ISSN: 2218-273X. DOI: 10.3390/biom15050700. URL: https://www.mdpi.com/2218-273X/15/5/700.

[17] Md Mobashir Hasan Shandhi, Will Ke Wang, and Jessilyn Dunn. “Taking the time for our bodies: How wearables can be used to assess circadian physiology”. In: *Cell Reports Methods* 1.4 (2021). ISSN: 2667-2375. DOI: 10.1016/j.crmeth.2021.100067. URL: https://doi.org/10.1016/j.crmeth.2021.100067.

[18] Tomasz Cudejko, Kate Button, and Mohammad Al-Amri. “Validity and reliability of accelerations and orientations measured using wearable sensors during functional activities”. In: *Scientific Reports* 12.1 (2022), p. 14619. ISSN: 2045-2322. DOI: 10.1038/s41598-022-18845-x. URL: https://doi.org/10.1038/s41598-022-18845-x.

[19] Jonathan A Mitchell et al. “Variation in actigraphy-estimated rest-activity patterns by demographic factors”. In: *Chronobiology International* 34.8 (2017). PMID: 28650674, pp. 1042–1056. DOI: 10.1080/07420528.2017.1337032. eprint: https://doi.org/10.1080/07420528.2017.1337032. URL: https://doi.org/10.1080/07420528.2017.1337032.

[20] Margaret M. Doyle et al. “Enhancing cosinor analysis of circadian phase markers using the gamma distribution”. In: *Sleep Medicine* 92 (2022), pp. 1–3. ISSN: 1389-9457. DOI: https://doi.org/10.1016/j.sleep.2022.01.015. URL: https://www.sciencedirect.com/science/article/pii/S1389945722000181.

[21] Dongju Lim et al. “Accurately predicting mood episodes in mood disorder patients using wearable sleep and circadian rhythm features”. In: *npj Digital Medicine* 7.1 (2024), p. 324. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01333-z. URL: https://doi.org/10.1038/s41746-024-01333-z.

[22] Marco Solmi et al. “Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies”. In: *Molecular Psychiatry* 27.1 (2022), pp. 281–295. ISSN: 1476-5578. DOI: 10.1038/s41380-021-01161-7. URL: https://doi.org/10.1038/s41380-021-01161-7.

[23] Jiahui Yin et al. “Nonlinear relationship between sleep midpoint and depression symptoms: a cross-sectional study of US adults”. In: *BMC Psychiatry* 23.1 (2023), p. 671. ISSN: 1471-244X. DOI: 10.1186/s12888-023-05130-y.

[24] S. Cuschieri. “The STROBE guidelines”. In: *Saudi J Anaesth* 13 (2019), S31–S34. ISSN: 1658-354X. DOI: 10.4103/sja.SJA_543_18.

[25] National Center for Health Statistics. *About NHANES*. URL: https://www.cdc.gov/nchs/nhanes/about/index.html. (accessed: 24.04.2025).

[26] Centers for Disease Control and Prevention. *About CDC*. URL: https://www.cdc.gov/about/cdc/index.html. (accessed: 29.04.2025).

[27] National Center for Health Statistics. *Who Participates In NHANES*. URL: https://www.cdc.gov/nchs/nhanes/about/who-participates.html. (accessed: 29.04.2025).

[28] National Center for Health Statistics. *Ethics Review Board Approval*. URL: https://www.cdc.gov/nchs/nhanes/about/erb.html. (accessed: 29.04.2025).

[29] National Center for Health Statistics. *2003-2004 Data Documentation, Codebook, and Frequencies: Physical Activity Monitor*. URL: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2003/DataFiles/PAXRAW_C.htm. (accessed: 20.05.2025).

[30] A. M. Hashmi et al. “Insomnia during pregnancy: Diagnosis and Rational Interventions”. In: *Pak J Med Sci* 32.4 (2016), pp. 1030–7. ISSN: 1682-024X. DOI: 10.12669/pjms.324.10421.

[31] K. Bakrania et al. “Intensity Thresholds on Raw Acceleration Data: Euclidean Norm Minus One (ENMO) and Mean Amplitude Deviation (MAD) Approaches”. In: *PLoS One* 11.10 (2016), e0164045. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0164045.

[32] John Muschelli. *Summarizing Actigraphy Data*. URL: https://johnmuschelli.com/SummarizedActigraphy/articles/Summarizing_Actigraphy_Data.html. (accessed: 09.05.2025).

[33] Bruno S.B. Gonçalves et al. “Nonparametric methods in actigraphy: An update”. In: *Sleep Science* 7.3 (2014), pp. 158–164. ISSN: 1984-0063. DOI: https://doi.org/10.1016/j.slsci.2014.09.013.

[34] Christine Blume, Nayantara Santhi, and Manuel Schabus. “‘nparACT’ package for R: A free software tool for the non-parametric analysis of actigraphy data”. In: *MethodsX* 3 (2016), pp. 430–435. ISSN: 2215-0161. DOI: https://doi.org/10.1016/j.mex.2016.05.006. URL: https://www.sciencedirect.com/science/article/pii/S221501611630022X.

[35] Germaine Cornelissen. “Cosinor-based rhythmometry”. In: *MethodsX* 11 (2014), p. 16. ISSN: 1742-4682. DOI: 10.1186/1742-4682-11-16. URL: https://doi.org/10.1186/1742-4682-11-16.

[36] Chih-Liang Wang, Cheng-Xue Li, and Sheng-Fu Liang. “The lifestyle of new middle-aged and older adults in Taiwan described by wearable device: age and gender differences”. In: *European Journal of Ageing* 21 (Sept. 2024). DOI: 10.1007/s10433-024-00824-y.

[37] Government of British Columbia. *Patient Health Questionnaire (PHQ-9)*. URL: https://www2.gov.bc.ca/assets/gov/health/practitioner-pro/bc-guidelines/depression_patient_health_questionnaire.pdf. (accessed: 20.05.2025).

[38] Linda Theron et al. “Factors that affect the resilience of young adults to depression: a systematic review”. In: *The Lancet Psychiatry* 12.5 (2025), pp. 377–383. ISSN: 2215-0366. DOI: 10.1016/S2215-0366(25)00044-6. URL: https://doi.org/10.1016/S2215-0366(25)00044-6.

[39] Valérie Cochen et al. “Sleep disorders and their impacts on healthy, dependent, and frail older adults”. In: *The Journal of nutrition, health and aging* 13.4 (2009), pp. 322–329. ISSN: 1279-7707. DOI: https://doi.org/10.1007/s12603-009-0030-0. URL: https://www.sciencedirect.com/science/article/pii/S1279770723022510.

[40] J. F. Duffy, K. M. Zitting, and E. D. Chinoy. “Aging and Circadian Rhythms”. In: *Sleep Med Clin* 10.4 (2015), pp. 423–434. ISSN: 1556-407X. DOI: 10.1016/j.jsmc.2015.08.002.

[41] University of Southampton. *Research uncovers differences between men and women in sleep, circadian rhythms and metabolism.* 2024. URL: www.sciencedaily.com/releases/2024/04/240410112643.html. (accessed: 08.05.2025).

[42] European Institute for Gender Equality. *Gender differences in mental disorders begin early in life*. 2021. URL: https://eige.europa.eu/publications-resources/toolkits-guides/gender-equality-index-2021-report/gender-differences-mental-disorders-begin-early-life?language_content_entity=en. (accessed: 08.05.2025).

[43] K. J. Egan et al. “The role of race and ethnicity in sleep, circadian rhythms and cardiovascular health”. In: *Sleep Med Rev* 33 (2017), pp. 70–78. ISSN: 1087-0792. DOI: 10.1016/j.smrv.2016.05.004.

[44] R. K. Bailey, J. Mokonogho, and A. Kumar. “Racial and ethnic differences in depression: current perspectives”. In: *Neuropsychiatr Dis Treat* 15 (2019), pp. 603–609. ISSN: 1176-6328. DOI: 10.2147/ndt.S128584.

[45] Baojing Li et al. “Educational level and the risk of mental disorders, substance use disorders and self-harm in different age-groups: A cohort study covering 1,6 million subjects in the Stockholm region”. In: *International Journal of Methods in Psychiatric Research* 32.4 (2023), e1964. DOI: https://doi.org/10.1002/mpr.1964. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mpr.1964. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mpr.1964.

[46] James Pagel and Carol Kwiatkowski. “Sleep Complaints Affecting School Performance at Different Educational Levels”. In: *Frontiers in Neurology* 1 (Nov. 2010), p. 125. DOI: 10.3389/fneur.2010.00125.

[47] Michael A. Grandner. “Sleep, Health, and Society”. In: *Sleep Medicine Clinics* 12.1 (2017), pp. 1–22. ISSN: 1556-407X. DOI: 10.1016/j.jsmc.2016.10.012. URL: https://doi.org/10.1016/j.jsmc.2016.10.012.

[48] Marta Wilk et al. “Associations between household overcrowding and adult mental illness in an ethnically diverse urban population: cross-sectional study using linked primary care and housing records.” In: *International Journal of Population Data Science* 9.5 (2024). DOI: 10.23889/ijpds.v9i5.2583. URL: https://ijpds.org/article/view/2583.

[49] S. Numaguchi et al. “Passive cigarette smoking changes the circadian rhythm of clock genes in rat intervertebral discs”. In: *J Orthop Res* 34.1 (2016), pp. 39–47. ISSN: 0736-0266. DOI: 10.1002/jor.22941.

[50] Haoxiong Sun and Sijia Li. “Exploring the relationship between smoking and poor sleep quality: a cross-sectional study using NHANES”. In: *Frontiers in Psychiatry* 15 (2024). ISSN: 1664-0640. DOI: 10.3389/fpsyt.2024.1407741. URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2024.1407741.

[51] A. M. Rosenwasser. “Alcohol, antidepressants, and circadian rhythms. Human and animal models”. In: *Alcohol Res Health* 25.2 (2001), pp. 126–135. ISSN: 1535-7414. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6707123/.

[52] K. M. Keyes et al. “Alcohol consumption predicts incidence of depressive episodes across 10 years among older adults in 19 countries”. In: *Int Rev Neurobiol* 148 (2019), pp. 1–38. ISSN: 0074-7742. DOI: 10.1016/bs.irn.2019.09.001.

[53] S. He, B. P. Hasler, and S. Chakravorty. “Alcohol and sleep-related problems”. In: *Curr Opin Psychol* 30 (2019), pp. 117–122. ISSN: 2352-250x. DOI: 10.1016/j.copsyc.2019.03.007.

[54] National Institute on Alcohol Abuse and Alcoholism. *Alcohol’s Effects on Health*. URL: https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-drinking-patterns. (accessed: 12.05.2025).

[55] Sohrab Amiri. “Unemployment associated with major depression disorder and depressive symptoms: a systematic review and meta-analysis”. In: *International Journal of Occupational Safety and Ergonomics* 28.4 (2022). PMID: 34259616, pp. 2080–2092. DOI: 10.1080/10803548.2021.1954793. eprint: https://doi.org/10.1080/10803548.2021.1954793. URL: https://doi.org/10.1080/10803548.2021.1954793.

[56] M. Maeda et al. “Association between unemployment and insomnia-related symptoms based on the Comprehensive Survey of Living Conditions: a large cross-sectional Japanese population survey”. In: *Ind Health* 57.6 (2019), pp. 701–710. ISSN: 0019-8366. DOI: 10.2486/indhealth.2018-0031.

[57] David G. Blanchflower and Alex Bryson. “Unemployment and sleep: evidence from the United States and Europe”. In: *Economics Human Biology* 43 (2021), p. 101042. ISSN: 1570-677X. DOI: https://doi.org/10.1016/j.ehb.2021.101042. URL: https://www.sciencedirect.com/science/article/pii/S1570677X21000666.

[58] Xueting Guan. “Circadian Rhythm and Obesity”. In: *Highlights in Science, Engineering and Technology* 66 (Sept. 2023), pp. 74–83. DOI: 10.54097/hset.v66i.11624.

[59] N. Badillo et al. “Correlation Between Body Mass Index and Depression/Depression-Like Symptoms Among Different Genders and Races”. In: *Cureus* 14.2 (2022), e21841. ISSN: 2168-8184. DOI: 10.7759/cureus.21841.

[60] S. Amiri. “Body mass index and sleep disturbances: a systematic review and meta-analysis”. In: *Postep Psychiatr Neurol* 32.2 (2023), pp. 96–109. ISSN: 1230-2813. DOI: 10.5114/ppn.2023.129067.

[61] Cleveland Clinic. *Body Mass Index (BMI)*. URL: https://my.clevelandclinic.org/health/articles/9464-body-mass-index-bmi. (accessed: 09.05.2025).

[62] themis. *Apply SMOTE Algorithm*. URL: https://themis.tidymodels.org/reference/step_smote.html. (accessed: 18.06.2025).

[63] O. Rainio, J. Teuho, and R. Klén. “Evaluation metrics and statistical tests for machine learning”. In: *Sci Rep* 14.1 (2024), p. 6086. ISSN: 2045-2322. DOI: 10.1038/s41598-024-56706-x.

[64] K. Çorbacıoğlu Ş and G. Aksel. “Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value”. In: *Turk J Emerg Med* 23.4 (2023), pp. 195–198. ISSN: 2452-2473. DOI: 10.4103/tjem.tjem_182_23.

[65] Abzu. *An introduction to calibration (part I): Understanding the basics*. URL: https://www.abzu.ai/data-science/calibration-introduction-part-1/. (accessed: 02.06.2025).

[66] Abzu. *An introduction to calibration (part II): Platt scaling, isotonic regression, and beta calibration.* URL: https://www.abzu.ai/data-science/calibration-introduction-part-2/. (accessed: 02.06.2025).

[67] D. Vaughan. *Equivocal zones*. URL: https://probably.tidymodels.org/articles/equivocal-zones.html. (accessed: 18.06.2025).

[68] Sanchinsoni. *K Nearest Neighbours — Introduction to Machine Learning Algorithms*. URL: https://medium.com/@sachinsoni600517/k-nearest-neighbours-introduction-to-machine-learning-algorithms-9dbc9d9fb3b2. (accessed: 18.06.2025).

[69] Olga Minaeva et al. “Level and timing of physical activity during normal daily life in depressed and non-depressed individuals”. In: *Translational Psychiatry* 10.1 (2020), p. 259. ISSN: 2158-3188. DOI: 10.1038/s41398-020-00952-w.

[70] Delainey L. Wescott et al. “Sleep and circadian rhythm profiles in seasonal depression”. In: *Journal of Psychiatric Research* 156 (2022), pp. 114–121. ISSN: 0022-3956. DOI: https://doi.org/10.1016/j.jpsychires.2022.10.019. URL: https://www.sciencedirect.com/science/article/pii/S0022395622005520.

[71] Futurense. *F1 Score in Machine Learning: All You Need To Know in 2025*. URL: https://www.futurense.com/uni-blog/f1-score-machine-learning. (accessed: 03.06.2025).

[72] R. Fossion et al. “Multiscale adaptive analysis of circadian rhythms and intradaily variability: Application to actigraphy time series in acute insomnia subjects”. In: *PLoS One* 12.7 (2017), e0181762. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0181762.

[73] Xuan Zhao et al. “Circadian Amplitude Regulation via FBXW7-Targeted REV-ERBa Degradation”. In: *Cell* 165.7 (2016), pp. 1644–1657. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.05.012.

[74] National Collaborating Centre for Mental Health (UK). *Depression in Adults with a Chronic Physical Health Problem: Treatment and Management*. Vol. 91. NICE Clinical Guidelines. Appendix 12: The classification of depression and depression rating scales/questionnaires. Leicester (UK): British Psychological Society, 2010. URL: https://www.ncbi.nlm.nih.gov/books/NBK82926/.

[75] NIH. *Major Depression*. URL: https://www.nimh.nih.gov/health/statistics/major-depression. (accessed: 04.06.2025).

[76] F. Movahedi, R. Padman, and J. F. Antaki. “Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores”. In: *J Thorac Cardiovasc Surg* 165.4 (2023), pp. 1433–1442. ISSN: 0022-5223. DOI: 10.1016/j.jtcvs.2021.07.041.

[77] R. van den Goorbergh et al. “The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression”. In: *J Am Med Inform Assoc* 29.9 (2022), pp. 1525–1534. ISSN: 1067-5027. DOI: 10.1093/jamia/ocac093.

[78] NIH. *Mental Illness*. URL: https://www.nimh.nih.gov/health/statistics/mental-illness. (accessed: 03.06.2025).

[79] K. Kroenke, R. L. Spitzer, and J. B. Williams. “The PHQ-9: validity of a brief depression severity measure”. In: *J Gen Intern Med* 16.9 (2001), pp. 606–613. ISSN: 0884-8734. DOI: 10.1046/j.1525-1497.2001.016009606.x.

APPENDIX A
AI STATEMENT

During the preparation of this work, the author used ChatGPT, DeepSeek, and Grammarly Pro as supportive tools. ChatGPT and DeepSeek were used to debug code and explore new code approaches when the author faced difficulties, while Grammarly Pro was used to check and refine spelling, grammar, and phrasing after writing the text. After using these tools/services, the author thoroughly reviewed and edited the content as needed, taking full responsibility for the final outcome.

APPENDIX B
CODE AVAILABILITY

The code is available on the following website: https://github.com/nikkiov/Bachelor-Thesis

APPENDIX C
FULL SENSITIVITY ANALYSIS

Note: The structure of this appendix (sensitivity analysis) mirrors that of the results section (primary analysis).

As shown in Table XI, the circadian rhythm parameters differ (although not significantly) between participants in the sensitivity cohort who met the criteria for depression and healthy controls. The p value is also shown, determined using Welch's Two Sample t-test. Significance was set at $p < 0.05$. In addition, Cohen's d is also shown in the table. Depression resulted in a lower average MESOR, amplitude, IS, and RA, a later average acrophase, and a higher average IV. However, looking at the p value, these differences were not significant. Cohen's d also reflected negligible differences across all parameters. A similar pattern is also evident in the density plots shown in Fig. 20. There were also apparent differences in covariates between participants who met the depression criteria and healthy controls, as shown in Table XII, the bar plots in Fig. 21, and the density plot in Fig. 22. As highlighted in these figures, the incidence of depression was relatively the highest among older female participants of black descent who were underweight, low educated, unemployed, had larger households, smoked and were heavy drinkers. The table shows that the differences in education, employment, household size, alcohol consumption, and race were significant.

TABLE XI: Circadian rhythm parameters of participants with depression and healthy controls, sensitivity cohort

Variable	**Depression**		**Healthy controls**		p	**Cohen's** d
	Mean	SD	Mean	SD		
MESOR (g)	0.0295	0.0104	0.0305	0.00927	0.517	0.11
Acrophase (clock hour)	15:14:26	03:21:13	15:12:09	02:13:23	0.928	-0.02
Amplitude (g)	0.0181	0.00810	0.0187	0.00752	0.642	0.08
IS	0.411	0.133	0.424	0.136	0.535	0.10
IV	0.860	0.249	0.850	0.252	0.805	-0.04
RA	0.627	0.226	0.661	0.212	0.327	0.16

Statistical significance (*) = $p < 0.05$, according to Welch's Two Sample t-test.
SD = Standard Deviation, MESOR = Midline Estimating Statistic Of Rhythm,
IV = Intradaily Variability, IS = Interdaily Stability, RA = Relative Amplitude.

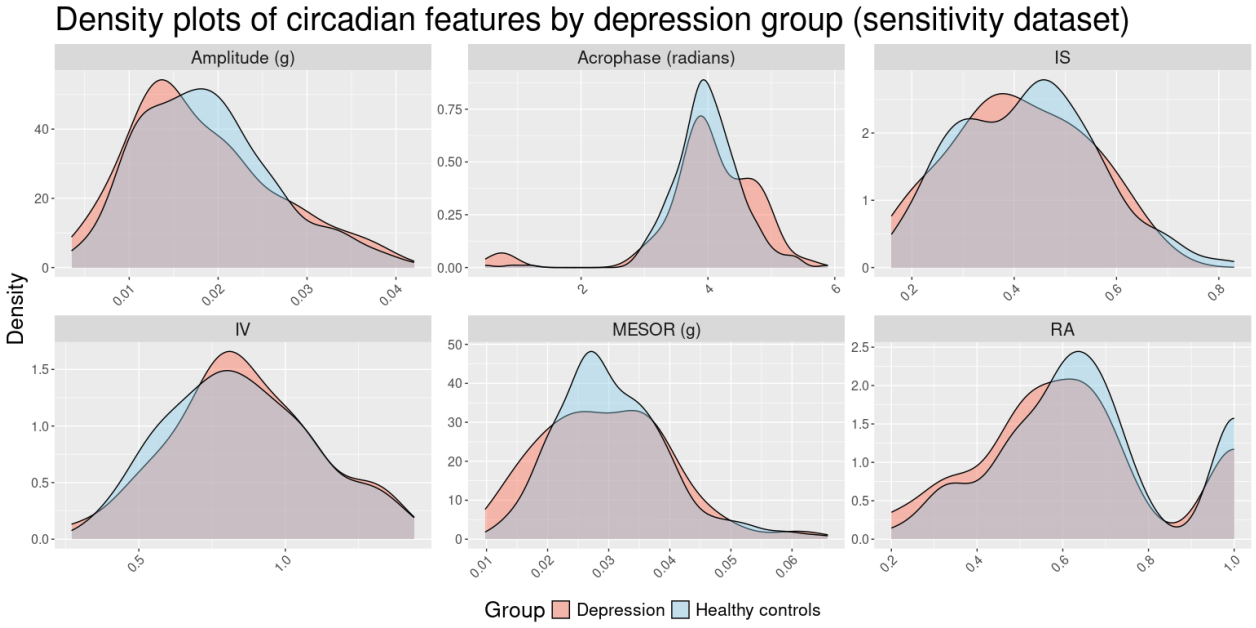


Fig. 20: *Density plots of the circadian features by depression group in the sensitivity cohort. Variables include the amplitude, the acrophase (in radians), the IS (Interdaily Stability), the IV (Intradaily Variability), the MESOR (Midline Estimated Statistic Of Rhythm), and the RA (Relative Amplitude). The depression group is shown in red, and the healthy controls are shown in blue. The y-axis scale is not fixed.*

TABLE XII: Significance test of covariate parameters between participants with depression and healthy controls, sensitivity cohort

Variable	Test	Test statistic	p
BMI	Chi-squared	0.0105	0.28
Education	Chi-squared	11.08	0.0008 (*)
Employment	Chi-squared	15.08	0.0001 (*)
Sex	Chi-squared	0.032	0.86
Smoking behaviour	Chi-squared	0.47	0.49
Household size	Chi-squared	8.14	0.02 (*)
Alcohol consumption	Chi-squared	19.16	0.00006 (*)
Race	Chi-squared	11.93	0.04 (*)
Age	t-test	-0.08	0.94

Statistical significance (*) = $p < 0.05$.

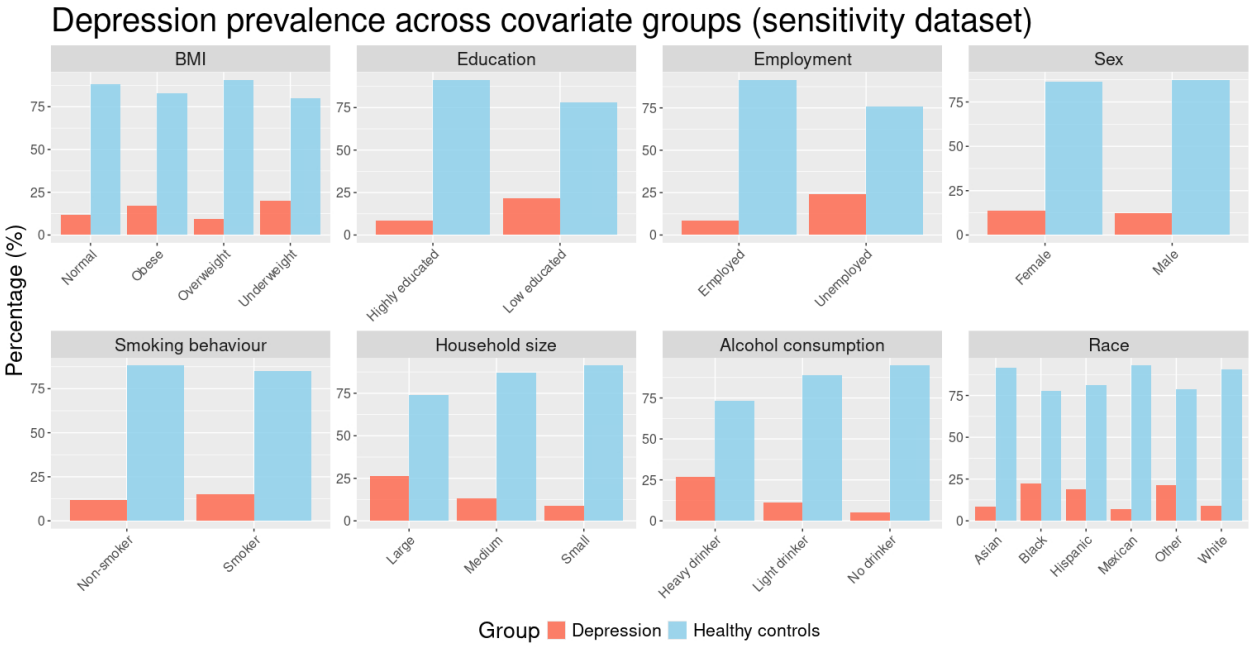


Fig. 21: *Bar plots of covariates by depression group in the sensitivity cohort. The depression group is shown in red, and the healthy controls are shown in blue. The y-axis scale is not fixed and shows the percentage of participants within each covariate group with and without depression.*

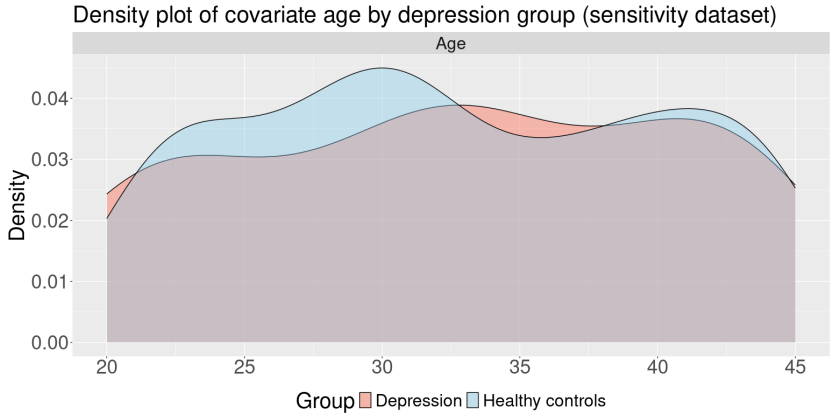


Fig. 22: *Density plot of the covariate age by depression group in the sensitivity cohort. The depression group is shown in red, and the healthy controls are shown in blue.*

Table XIII shows the circadian rhythm parameters of participants in the sensitivity cohort with sleep disturbances and healthy controls. Sleep disturbances resulted in a lower MESOR, amplitude, IS, and RA, an earlier acrophase, and a higher IV. The MESOR, amplitude, and the IV were significantly different, considering the p value. Cohen's d reflected a small effect size for the MESOR, amplitude, and the IV. Negligible effect sizes were found for the other parameters. Fig. 23 visualizes this difference, with the amplitude and the MESOR being lower in the sleep disturbances group than in the healthy controls. At the same time, the IV was higher in the sleep disturbances group. As shown in Table XIV, the bar plots in Fig. 24, and the density plot in Fig. 25, the distribution of the covariates varied considerably between participants with and without sleep disturbances. The incidence of sleep disturbances was relatively highest among older female participants of other racial backgrounds who were obese, highly educated, unemployed, had smoked, lived in smaller households, and drank no alcoholic beverages. The differences in household size, race, and age were significant between the groups.

TABLE XIII: Circadian rhythm parameters of participants with sleep disturbance and healthy controls, sensitivity cohort

Variable	**Sleep disturbances**		**Healthy controls**		p	**Cohen's** d
	Mean	SD	Mean	SD		
MESOR (g)	0.0288	0.00857	0.0311	0.00970	0.030 (*)	0.24
Acrophase (clock hour)	15:02:59	01:42:40	15:14:26	02:38:49	0.419	-0.02
Amplitude (g)	0.0170	0.00661	0.0194	0.00788	0.003 (*)	0.32
IS	0.414	0.131	0.427	0.137	0.405	0.09
IV	0.908	0.241	0.826	0.252	0.004 (*)	-0.33
RA	0.653	0.225	0.658	0.209	0.827	0.03

Statistical significance (*) = $p < 0.05$, according to Welch's Two Sample t-test.
SD = Standard Deviation, MESOR = Midline Estimating Statistic Of Rhythm,
IV = Intradaily Variability, IS = Interdaily Stability, RA = Relative Amplitude.

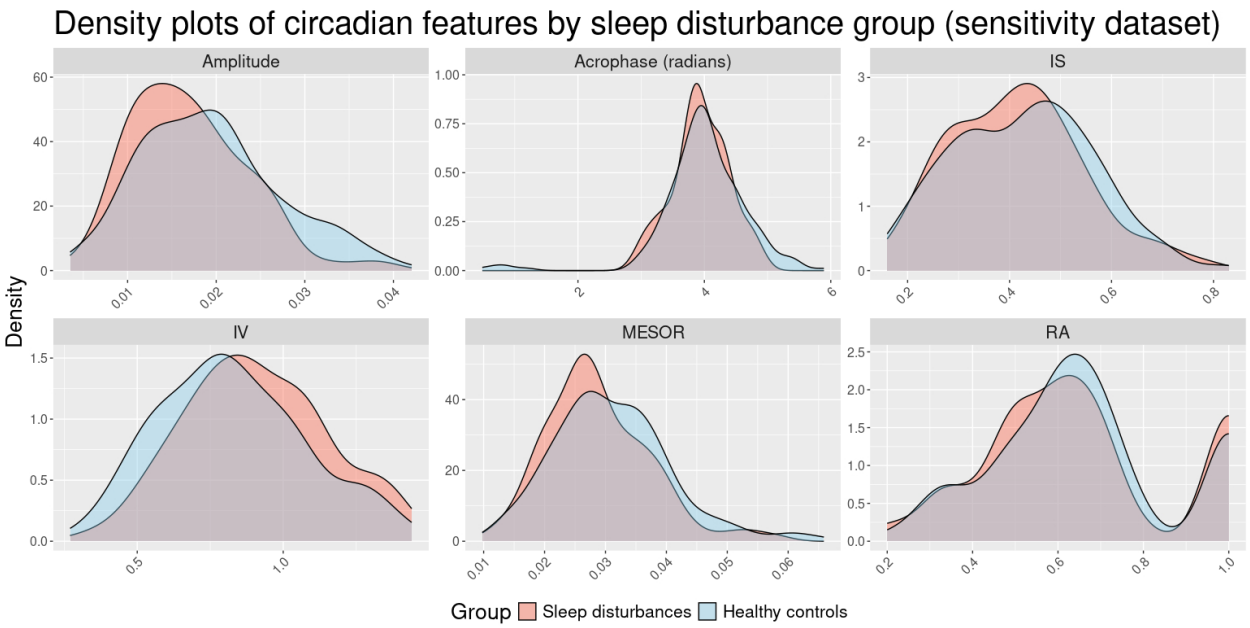


Fig. 23: *Density plots of the circadian features, by sleep disturbance group in the sensitivity cohort. Variables include the amplitude, the acrophase (in radians), the IS (Interdaily Stability), the IV (Intradaily Variability), the MESOR (Midline Estimated Statistic Of Rhythm), and the RA (Relative Amplitude). The sleep disturbances group is shown in red and the healthy controls are shown in blue. The y-axis scale is not fixed.*

TABLE XIV: Significance test of covariate parameters between participants with sleep disturbances and healthy controls, sensitivity cohort

Variable	**Test**	**Test statistic**	p
BMI	Chi-squared	4.42	0.22
Education	Chi-squared	0.13	0.72
Employment	Chi-squared	3.77	0.05
Sex	Chi-squared	0.88	0.35
Smoking behaviour	Chi-squared	2.17	0.14
Household size	Chi-squared	7.63	0.02 (*)
Alcohol consumption	Chi-squared	1.04	0.60
Race	Chi-squared	22.50	0.0004 (*)
Age	t-test	-2.71	0.007 (*)

Statistical significance (*) = $p < 0.05$.

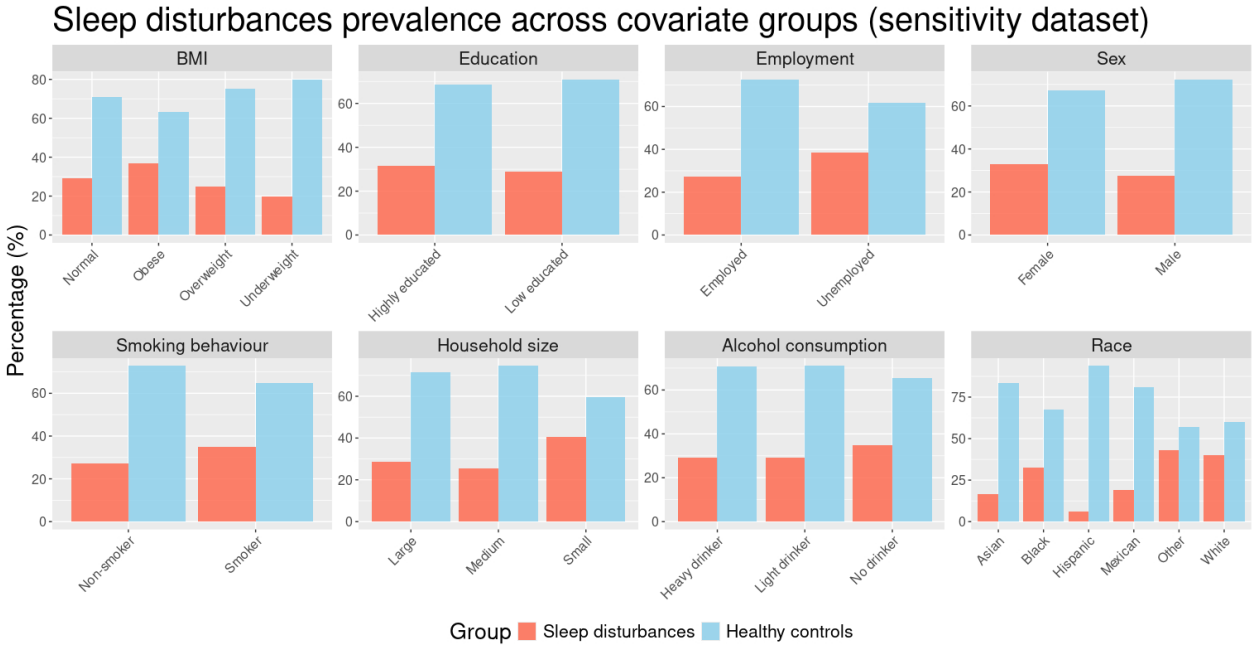


Fig. 24: *Bar plots of covariates by sleep disturbance group in the sensitivity cohort. The sleep disturbance group is shown in red, and the healthy controls are shown in blue. The y-axis scale is not fixed and shows the percentage of participants within each covariate group with and without sleep disturbances.*

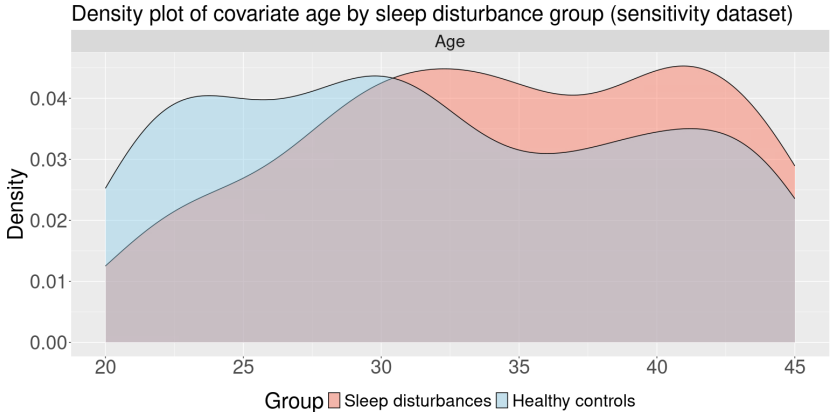


Fig. 25: *Density plot of the covariate age by sleep disturbance group in the sensitivity cohort. The sleep disturbance group is shown in red, and the healthy controls are shown in blue.*

A. Performance of predictive machine learning models

Fig. 26 displays the cross-validated performance of trained machine learning models predicting depression in the sensitivity cohort, based on the metrics accuracy, F1-score, precision, recall, and specificity. This figure shows that the k-nearest neighbors models had the best recall and F1-score, while the random forest models had the best accuracy, precision, and specificity. The cross-validated performance of trained machine learning models predicting sleep disturbances in the sensitivity cohort is shown in Fig. 27, based on the same metrics. As shown in the figure, the k-nearest neighbors models had the best F1-score and recall, the random forest models had the best accuracy and specificity, and the precision was similar across the models.

For each outcome and model type, the two best-performing models were chosen for further analysis. One of the best-performing models included the covariates in the preprocessing steps, and the other only included the circadian rhythm features. The best-performing models were chosen based on the F1-score and secondarily on the AUC value. The F1-score, accuracy, specificity, precision, recall, true positives, true negatives, false positives, false negatives, and AUC values of the test datasets for each of the best-performing models are shown in Table XV.

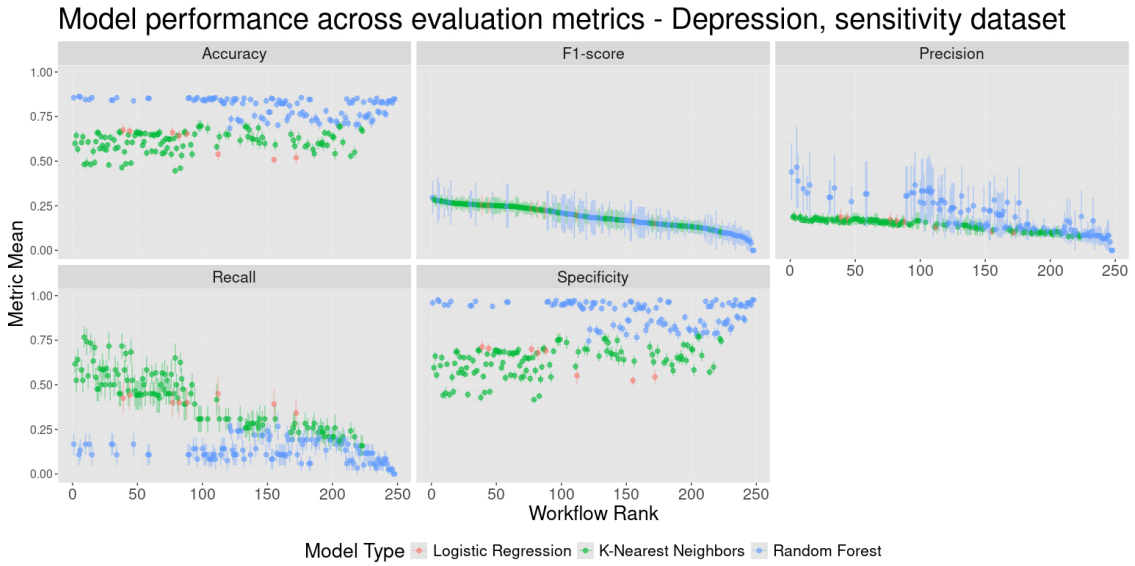


Fig. 26: *Cross-validated performance of the trained machine learning models predicting depression in the sensitivity cohort based on five evaluation metrics: accuracy, F1-score, precision, recall, and specificity. Shown in blue are the random forest models, in green the k-nearest neighbors models, and in red the logistic regression models. The x-axis shows the rank of the model, based on the F1-score. The y-axis shows the metric score, which ranges from 0.00 to 1.00.*

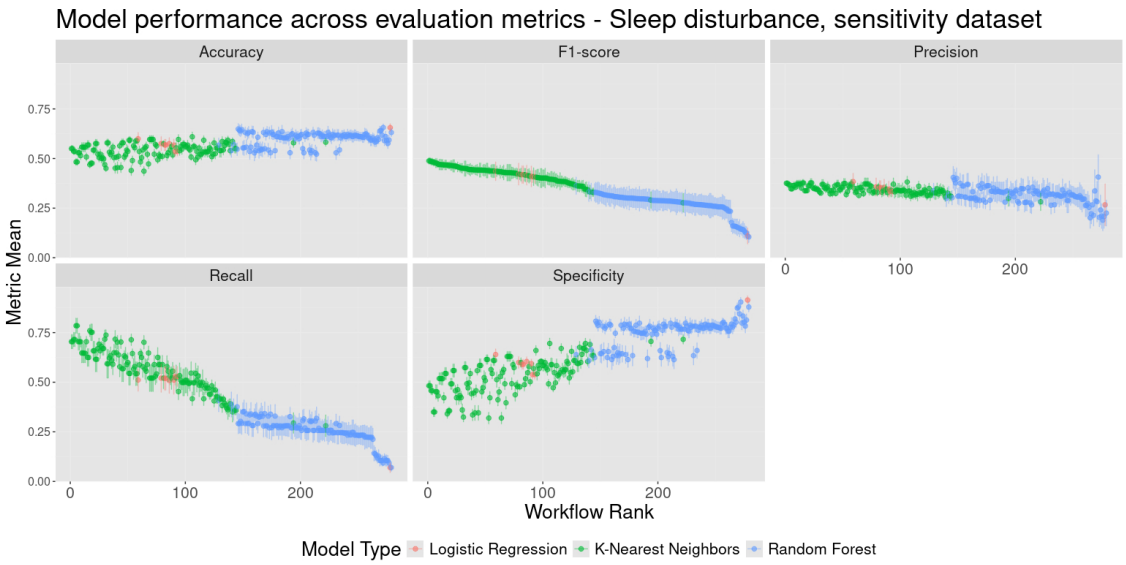


Fig. 27: *Cross-validated performance of the trained machine learning models predicting sleep disturbances in the sensitivity cohort based on five evaluation metrics: accuracy, F1-score, precision, recall, and specificity. Shown in blue are the random forest models, in green the k-nearest neighbors models, and in red the logistic regression models. The x-axis shows the rank of the model, based on the F1-score. The y-axis shows the metric score, which normally ranges from 0.00 to 1.00.*

TABLE XV: Performance of the best-performing models on the test datasets for depression and sleep disturbances outcomes, based on key evaluation metrics, per model type and the inclusion/exclusion of covariates

Key performance metrics	**Random forest**		**K-nearest neighbors**		**Logistic regression**	
	Included	Excluded	Included	Excluded	Included	Excluded
Depression						
F1-score	0.267	0.111	0.343	0.111	**0.452**	0.174
Accuracy	**0.847**	0.778	0.681	0.778	0.764	0.472
Specificity	**0.952**	0.887	0.694	0.887	0.774	0.484
Precision	**0.4**	0.125	0.24	0.125	0.333	0.111
Recall	0.2	0.1	0.6	0.1	**0.7**	0.4
True positives	2	1	6	1	**7**	4
True negatives	**59**	55	43	55	48	30
False positives	3	7	19	7	14	32
False negatives	8	9	4	9	3	6
AUC	0.74	0.48	0.71	0.45	**0.78**	0.37
Sleep disturbances						
F1-score	0.276	0.419	**0.515**	0.459	0.458	0.408
Accuracy	**0.708**	0.653	0.556	0.542	0.639	0.597
Specificity	**0.94**	0.76	0.46	0.5	0.7	0.66
Precision	**0.571**	0.429	0.386	0.359	0.423	0.370
Recall	0.182	0.409	**0.773**	0.636	0.5	0.455
True positives	4	9	**17**	14	11	10
True negatives	**47**	38	23	25	35	33
False positives	3	12	27	25	15	17
False negatives	18	13	5	8	11	12
AUC	**0.61**	0.54	0.60	0.51	0.60	0.57

AUC = Area Under the Curve

Values in **bold** represent the best performance for each metric across models.

1) Depression

All analysed machine learning models that predicted depression used a preprocessing step that created example samples of depression on a 1:1 ratio. The following models had additional preprocessing steps:

- **Random forest, covariates:** Interaction between IS and IV.
- **Random forest, no covariates:** Interaction between MESOR and amplitude.
- **Logistic regression, covariates:** Interaction between alcohol consumption and smoking.
- **K-nearest neighbors, no covariates:** Interaction between MESOR and amplitude.

Among the evaluated models, the random forest model that included covariates showed the highest accuracy (0.847), specificity (0.952), precision (0.4), and true negatives (59-3) at the expense of the recall (0.2) and true positives (2-8). On the other hand, the covariate-inclusive logistic regression model showed the highest F1-score (0.452), recall (0.7), true positives (7-3), and AUC (0.78). Since the covariate-inclusive logistic regression model had the highest F1-score and AUC, this model was seen as the best-performing model and included in the final analysis for predicting depression in the sensitivity cohort.

2) Sleep disturbances

The analysed machine learning models that predicted sleep disturbances had the following preprocessing steps:

- **Random forest, covariates:** Creating example samples of sleep disturbances on a 1:1 ratio & interaction between IS and IV.
- **Random forest, no covariates:** Creating example samples of sleep disturbances on a 1:1 ratio & interaction between IS and IV & interaction between MESOR and amplitude.
- **Logistic regression, covariates:** Creating example samples of sleep disturbances on a 1:1 ratio & interaction between alcohol consumption and smoking.
- **Logistic regression, no covariates:** Creating example samples of sleep disturbances on a 1:1 ratio & interaction between IS and IV & interaction between MESOR and amplitude.
- **K-nearest neighbors, covariates:** Creating example samples of sleep disturbances on a 1:1 ratio & interaction between MESOR and ampltiude & interaction between alcohol consumption, smoking, and BMI.
- **K-nearest neighbors, no covariates:** Creating example samples of sleep disturbances on a 1:1 ratio & interaction between MESOR and amplitude.

Among the evaluated models, the random forest model that included covariates showed the highest accuracy (0.708), specificity (0.94), precision (0.571), true negatives (47-3), and AUC (0.61), at the expense of the recall (0.182) and true positives (4-18). On the other hand, the covariate-inclusive k-nearest neighbors model showed the highest F1-score (0.515), recall (0.773), and true positives (17-5). However, the logistic regression model that included covariates was chosen for the final analysis as this model had the second highest F1-score (0.458), accuracy (0.639), and AUC (0.60).

B. Variable importance

For both depression and sleep disturbances, it was determined which predictor(s) had the most influence on the final prediction of the outcome (i.e., depression or sleep disturbances) in the test sensitivity datasets. This was visualised by using variable importance plots. Only plots of the best performing machine learning models were visualised for both depression and sleep disturbances, as can be seen in Fig. 28:

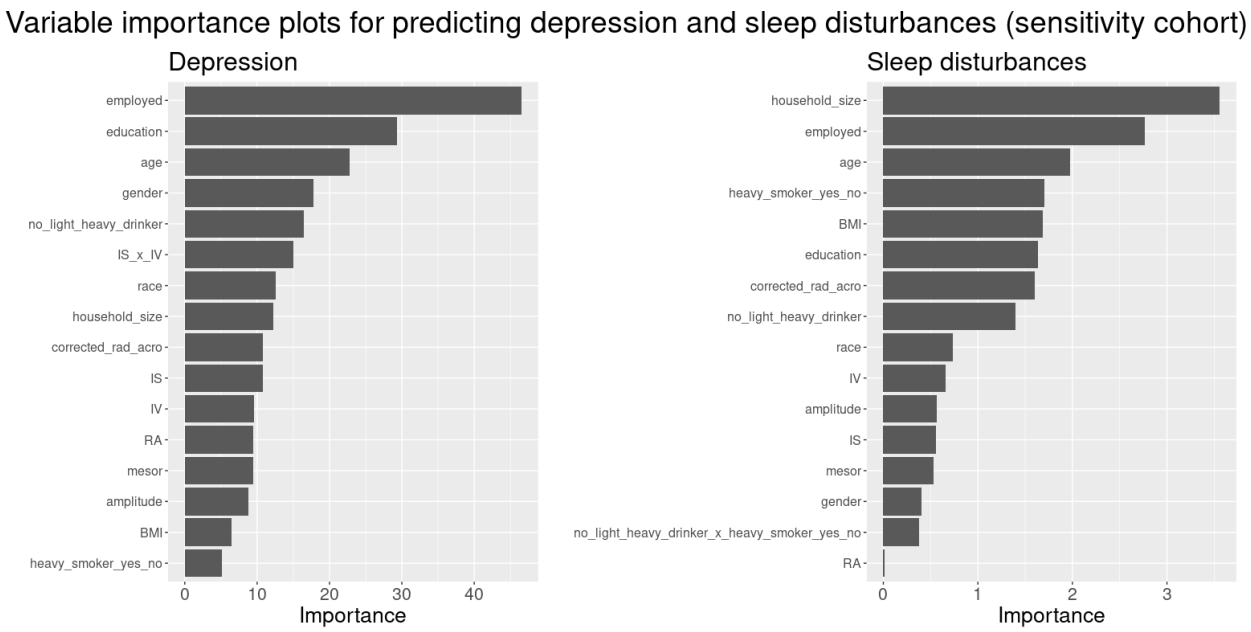


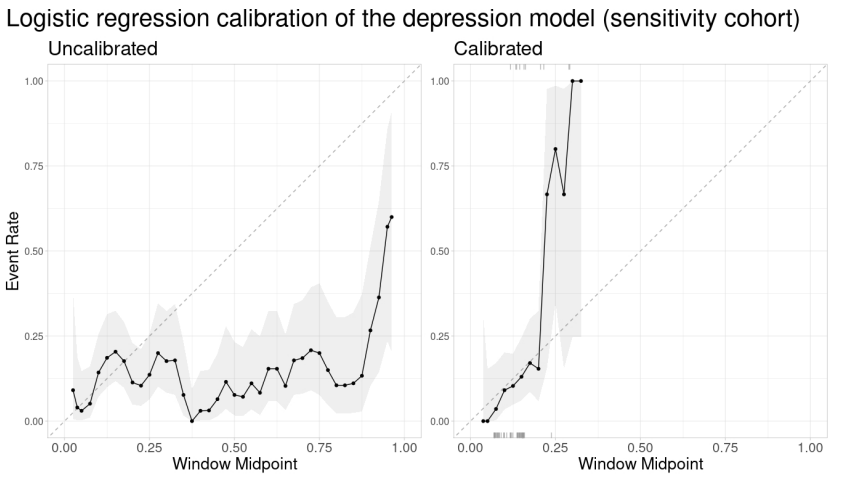
Fig. 28: *Variable importance plots of 2 different machine learning models predicting depression and sleep disturbances in the test sensitivity datasets. The figure on the left shows the variable importance plot of the machine learning model predicting depression. The model used is the covariate-inclusive logistic regression model. The figure on the right shows the variable importance plot of the machine learning model predicting sleep disturbances. The model used is the covariate-inclusive logistic regression model. The predictors (i.e., variables) are ranked from most to least important, with the most important variables at the top. IS = Interdaily Stability, IV = Intradaily Variability, RA, = Relative Amplitude, BMI = Body Mass Index, MESOR = Midline Estimated Statistic Of Rhythm.*

The variable importance plot of the covariate-inclusive logistic regression model predicting depression shows that the prediction was mainly based on the covariate predictors, in particular, employment and education. The most important circadian predictors were the interaction term between IS and IV. However, this interaction term does not have more influence than the IS and IV individually combined, so there is no significant interaction effect beyond their sum. From the predictors, smoking, BMI, and the amplitude had the least influence.

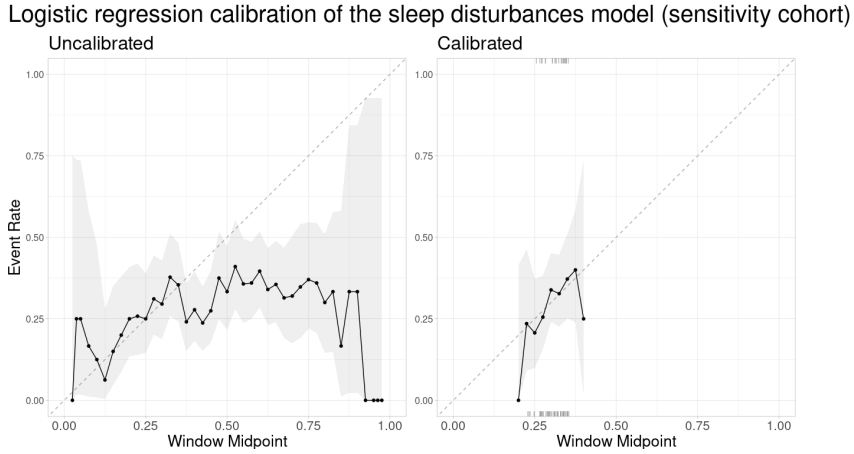
The variable importance plot of the covariate-inclusive logistic forest model predicting sleep disturbance reveals that the final prediction was mainly based on household size, employment, and age. From the circadian predictors, the acrophase had the most influence, followed by the IV. The RA had almost no impact on the final prediction of sleep disturbances in the sensitivity cohort, followed by the interaction term between alcohol consumption and smoking.

C. Calibration of the best performing models

Fig. 29a compares the uncalibrated and calibrated plots of the covariate-including logistic regression model predicting depression. The figure on the left (uncalibrated) shows that the probabilities were not showing good accuracy. The model over-predicted probabilities as the event rate fell below the diagonal line. The calibrated plot on the right visualises an event rate following the diagonal line until a window midpoint of around 0.25, after which the model started under-predicting the probabilities. In addition, this figure indicates that the model's predicted probability was never above 0.35. Fig. 29b compares the uncalibrated and calibrated plots of the covariate-including logistic regression model predicting sleep disturbances. The figure on the left reveals that the event rate fell below the diagonal line after a window midpoint of 0.35, which indicates that the model was over-predicting probabilities beyond this point. The calibrated figure on the right corrected the over-prediction. However, the calibrated model never had an event rate below 0.20 or above 0.45. These suggest that logistic regression calibration was insufficient in improving the best-performing models. The performance metrics of the uncalibrated and calibrated models are the same as the ones shown in Table XV.



(a)



(b)

Fig. 29: *Uncalibrated and calibrated plots of the best performing models (sensitivity cohort). The figures on the left show the uncalibrated plots, and those on the right show the logistic regression-calibrated plots. The x-axis shows the window midpoint (model's predicted probability), and the y-axis shows the event rate (fraction of positive cases).* ***(a)*** *Depression.* ***(b)*** *Sleep disturbances.*

D. Improvement of best performing models by removing equivocal values

Removing equivocal values improved the performance of the best-performing models predicting depression or sleep disturbances, as seen in Table XVI. For the machine learning model predicting depression, a threshold of 0.45 and a buffer of 0.1 improved all key metrics. The reportable rate decreased to 0.86. The machine learning model predicting sleep disturbances achieved the best performance metrics at a threshold of 0.40 and a buffer of 0.15. This improved F1-score, precision, and recall, while the accuracy and specificity declined. The reportable rate decreased to 0.39. The AUC remained the same for both outcomes before and after removing equivocal zones.

TABLE XVI: Performance metrics comparison before and after the removal of equivocal zones for the best performing models (sensitivity cohort)

Model	F1-score	Accuracy	Specificity	Precision	Recall	AUC
Depression, original	0.267	0.847	0.952	0.4	0.2	0.74
Depression, after removal	0.364	0.887	0.981	0.667	0.25	0.74
Sleep disturbances, original	0.458	0.639	0.7	0.423	0.5	0.60
Sleep disturbances, after removal	0.581	0.536	0.333	0.429	0.9	0.60

AUC = Area Under the Curve