The impact of sports on mental health

How wearable technology can help us control and maintain our state of mind

W.J. van der Rest

Master's Thesis
Artificial Intelligence
University of Groningen

25 Oct 2022

supervisors:

Dr. M. K. van Vugt
Artificial Intelligence - Univeristy of Groningen

Dr. Ir. G.J.W. van Dijk
Artificial Intelligence - Phaitality

university of groningen

wefitter

Phaitality

Abstract

How ideal would it be if a device small and non-intrusive enough would be able to help you live a healthier life? Over the last couple of years wearable technology has risen in popularity and accuracy. These 'wearables' promise to be useful for recording physical activity data such as daily activity, workouts, sleep and heart rate information. Earlier studies have shown that medical grade devices could be used to indicate the correlation between physical and mental health. This study aims to see if this correlation is also visible when using popular consumer wearables and if data from these wearables could be used to make a predictive model about the current mental state of an individual.
An experiment was conducted in which participants were asked to fill in a short mood assessment twice a day which scored the mental well-being of the participants based on positive affective state, rumination, resilience, cognitive complaints and depression. The answers to these questionnaires were used to capture mental health information, while wearables were used to capture the physical health information.
Linear mixed effects models showed that mental well-being is positively affected by more daily steps, a higher workout intensity and frequency and lower mean stress scores. The sleep information did not show to have a significant effect on mental well-being.
To see if this correlation could also be implemented in a predictive model, multiple models were trained using different classification methods. These models showed to be able to reach a maximum predictive accuracy of 75% when only predicting either a positive or a negative class of mental well-being. This accuracy could only be reached when some prior mental health information about depression and anxiety scores was added to the models.
This study partially confirms that wearables can be used to show the correlation between aspects of mental health and several physical health variables. Using the data and data types available in this study, the predictive model showed to be too inaccurate for reliable usage.

Contents

List of Figures

List of Tables

Introduction 1

How ideal would it be if a device small and non-intrusive enough would be able to help you live a healthier life? This could be helpful in, for example, the situation where someone has a dangerously infrequent heart rate. This device could be able to indicate that you should seek medical help. When someone has a panic attack the device could instruct someone with, for example, breathing exercises. Even when someone is feeling down/depressed, how great would it be if the device could help them by suggesting to take a walk.

Such a device would be put in the domain of wearable technology, a device that can be worn close to or on the surface of your skin. The interest in wearable technology has risen over the last couple of years [1][2] and so has the accuracy of recording physical information using this technology [3]. These 'wearables' promise to be useful in making ones life easier by being able to detect, analyze and provide information regarding signals from the body of the individual wearing that wearable. At the time of writing this paper, a wearable is generally capable of detecting and analyzing information like amount of steps taken, current heart rate, workout information, and sleep information. Some wearables are in addition also capable of processing heart rate variability (HRV) and sleep phase information. Most devices are accompanied by an app which is not only intended to provide extra information or to provide a more user friendly way of interpreting the wearables its data, but is also intended to be able to ask for extra information such as weight, length and age. it is possible to make an estimation on how physically healthy an individual is by combining all this data. If the person is not taking enough steps every day, the wearable will advise the user to take a walk, and if the user is not sleeping enough, the wearable will advise that the user should go to bed earlier.

Even though this sounds very promising, it mainly focuses on the physical aspect of health while the mental aspect is widely ignored within the wearable domain. This is unfortunate since mental health has been seen to have an effect on physical health and vice versa [4][5][6][7][8].

Poor mental health is often associated with mental disorders such as anxiety and depression. These mental health disorders are shown to be accompanied with sudden and uncontrollable mood changes [9][10][11]. If your mood is constantly negative, this might be an indication for depression [12], while a frequently changing mood might indicate a bipolar disorder [13][14].

Research has shown that a correlation can be found between mood and different types of physical information such as sleep data, activity data and heart rate information [15][16][17][18][19]. A positive mood is associated with enough and a regulated sleep cycle, enough physical activity and normal heart rate information without any signs of stress or other irregularities (per subject more detail later). Although mood alone cannot be used to diagnose a mental health disorder, it can be used as a motivational tool to go and find help from an expert.

Previous research has mostly been performed using medical grade devices, which are usually big bulky devices, so these cannot really be called wearables. There is one big problem with using such devices for concluding anything about an individuals mental/physical health over a period of time. The recordings that are made during this period of time are not reliable to be representative for the typical day of an average person. Either the recordings have to be made on site at a medical center once every couple of days/weeks, or the devices are intrusive enough that it could interfere with the outcome.
Another shortcoming of previous research in this field is that conclusions about mental health and physical health were solely made using the outcomes from self-reflection questionnaires [17][18]. This is problematic if participants have to fill in physical activity information such as how many activities they have performed and how intensive they rated these activities. This might simply be hard to correctly remember and it can be very confrontational. If someone has to fill in that they only made a couple of hundred of steps a day, it can in turn affect mood by being disappointed with themselves. It is also not easily possible to accurately measure the exact amount of steps taken, heart rate or stress values or the specific total duration of activities when solely using questionnaires.

1.1 The Importance of Mental Health

While disorders such as anxiety and depression are widely known to cause difficulties in life, the majority of people suffering from mental disorders will not seek professional help or are not even aware of the impact that a mental, behavioral, or emotional disorder can have on a person [20][21][22]. The World Health Organization (WHO) estimated that 1 in 5 people suffer from a mental disorder of which two-thirds is not effectively trying to fix it . The WHO says that this comes from stigma, discrimination and neglect. While seeking medical help should be seen as a healthy and smart idea, often barriers such as disgrace and embarrassment are still a problem [23].
Poor mental health can have a (silent) impact on multiple aspects of life. Next to having an impact on an individual its social life by self isolation and breaking social connections [24], it also decreases productivity and thus it can have a financial impact. In the worst case, poor mental health may result into suicide, which is one of the leading causes of death worldwide [25]. Every form of making the process easier of identifying and tackling mental disorders can have a positive impact on a persons life. Since previous research showed the correlation between mental and physical health, there is a real potential in wearables that could possibly be used to give instructions on how to become mentally healthier.

1.2 Mental Health in a Diary Study

This study uses symptoms from mental diseases such as depression and anxiety. These symptoms are used to find the correlation with data originating from wearables, and to predict the mental state of an individual. These mental diseases develop through the time span of months or years [26][27], so the scope of this study does not include to find the correlation

between mental diseases and wearable data. The mentioned symptoms are used to relate to the term 'mental state'. Mental state is used to explain the positive mental well-being of a person, how is their emotional, psychological and social well-being at a specific moment. This mental well-being score is represented as several different aspects; positive affective state (PAS), rumination (R), resilience (RES), cognitive/physical complaints (COG), and depression (D). The term positive affective state explains how much your positive emotions such as excitement, cheerfulness and energy weigh out against negative emotions such as sadness, fear or distress[28]. Together with cognitive/physical complaints, and the depression aspect, these are the most obvious aspects when thinking about mental health. When a person suffers from cognitive or physical complaints this could include a lack of attention, difficulty to focus, and physical pains such as a headache or muscle strain. The possible causes of these problems are often associated to issues with sleep and overuse (or improper use of) muscles. Next to these causes related to physical health, the cause could be related to mental health in the form of depression [29][30].
While having a significantly lower prevalence than depression [31], anxiety still affects millions of people worldwide. Similar to depression, anxiety can present itself with cognitive and physical complaints, while the act of rumination and a poor resilience capability are likely to also be present [32][33]. Rumination is the act of having thoughts that influence your depressive mood by interfering with attention and concentration. By enhancing the recall of negative events, and by increasing the likelihood of using depressogenic explanations for negative life events [34], which can cause stress levels to rise. With resilience is meant how capable you are at coping with negative life events and stress. Higher scores of depression and anxiety usually correlate with lower resilience scores [33].
Another important aspect about mental health is the company that you are with or the location where you are at. For example, the environment at work can be quite stressful and purely professional, while being with friends might make you feel more relaxed and more open about your feelings.
In contrast to the earlier mentioned use of self-reflection questionnaires, there are some reliable questionnaires that can make a conclusion about mental health. In the mentioned studies [17][18] implementations (or variations) of the Patient Health Questionnaire (PHQ-9) and the Generalized Anxiety Disorder (GAD-7) scales were used. The first is a depression module based on the PRIME-MD diagnostic instrument for common mental disorders. In comparison to the PRIME-MD instrument, the Patient Health Questionnaire (PHQ) can be self-administrated while also resulting into a reliable and valid measure of depression severity [35][36]. This questionnaire asks to grade how frequent some questions were applicable to you in the last two weeks, with questions like; "How often were you bothered to have little to no interest or pleasure in doing things?". This question has to be answered on a 4 point scale about frequency. Eventually the outcome of the questionnaire will give a suggestion on where you are on a scale of depression.
The Generalized Anxiety Disorder (GAD) questionnaire works in the same way and was proven to be a reliable method of diagnosing anxiety [37][38]. Since the mentioned mental well-being variable might be heavily influenced by the depression or anxiety scores, it's important

to have some prior information about mental health before looking for correlations with mental state.

1.3 Physical Aspects Correlated to Mental Health

The physical aspects correlated to mental health that could be recorded using wearables can be classified into three subjects; sleep data, activity data, and heart rate information. The difference between the three is that the first two are behavioral data types, while the last is a body function. This difference and the correlations of these aspects to mental health will be discussed in the following sections. Next to these aspects there are other factors that have an influence on mental well-being. These factors can be both genetical and environmental [39][40]. One of these environmental factors can be the weather for example [41] or the amount of access to sunlight over the day [42][43]. Only physical health parameters originating from consumer wearables will be used in this study.

The Relationship Between Sleep and Mental Health

As mentioned, previous research showed a correlation between mental health and sleep. A properly regulated and long enough sleep schedule showed to have a positive effect on mental health [18][44][45][46]. Mental disorders such as anxiety and depression are shown to be intertwined with insomnia over time. Anxiety can cause higher levels of rumination, which can increase stress, which can in turn decrease the quality of sleep [34]. While anxiety causes your thoughts to race, thus making it harder to fall asleep, depression can lead to oversleeping and irregularity in sleep rhythm [47][48][49].
Several aspects of sleep have shown to affect mental health. Correlations were found between mental health and sleep continuity, sleep depth, and excessive daytime sleepiness [44][50][51][18][45]. Sleep continuity is defined as a longer sleep duration combined with a reduced number of awakenings.
The sleep depth refers to the different sleep stages someone can be in during a sleep session. There are 5 different stages; wake, NREM1, NREM2, NREM3 and REM. Sequentially these stages show a slower heart beat, a lower breathing frequency and the body begins to relax more. The NREM3 (deep sleep) stage houses several health-promoting tasks such as tissue repair and immune system strengthening [52][53][54]. The REM stage shows a more variable heart rate, more variable breathing frequency and the brain activity increases (dreams occur during this stage). An increase in percentage of time spent in NREM3 and REM phase and a decrease of percentage of time spent in wake, NREM1 or NREM2 has shown to have a positive effect on mental health [51][55].
Next to the effects of 'normal' sleep sessions on mood, excessive daytime sleepiness was shown to correlate with a depressive disorder [56] and thus a poor mental health.
In addition to these symptoms that can be recorded on a daily basis, mental health showed to be correlated with sleep patterns [57]. An

infrequent sleep pattern, for both sleep onset and sleep duration is correlated with poor mental health.

Working Out to Improve Mental Health

Previous research showed a correlation between mental health and physical activity[58][59][60]. People who suffer from mental health diseases such as anxiety and depression generally tend to be less physically active [59], while frequent exercises and strength training has shown to reduce symptoms of these mental health diseases [61][62][63][64][65]. A study by Martinsen et al. (1985) [66] showed that individuals who suffer from depression can use aerobic exercises as an antidepressant.
In studies that used a combination of self-reflection questionnaires and medical grade devices, higher workout intensity and total workout duration showed to have a positive impact on mental health [67][68][69][70]. The workout frequency showed to be beneficial for mental health and cognition, where the optimal value is between 3 and 5 workouts a week [71][72].
The time since the last workout also showed to have a negative relation with mental health [73][74][75]. During workouts the production of several hormones such as dopamine is increased. In the period after the workouts this increased productivity slowly decreases.
The amount of steps taken showed to have a positive impact on mental health [76][77][78]. It was shown that people who reported to take longer walks, more frequently and at a more intensive level also reported to have an improved mental health. The level of intensity is based on the speed and distance of the walk and is reflected in the amount of calories burned.

Heart Rate Variability and Mental Health

Where you are able to choose to start a workout or try to go to bed, most people do not have a lot of control over their heart. In some ways it is convenient that your heart is automatically regulated, for example you do not have to actively think about letting your heart make a beat frequently enough. In other ways missing some control might be problematic for people. Next to the physical importance of your heart by keeping all organs running, the heart plays an essential role for your mental health. Especially heart rate variability showed to play a major role in indexing your stress levels and your resilience to stress [79][80][81]. Heart rate variability (HRV) is the variability of the time between consecutive heart beats. If there is a bigger difference between these times (so HRV is higher) stress is naturally lower and vice versa [82].
When stress is at a higher level it in turn can cause an imbalance in your quality of cognition, decision making and mood [83][84]. High stress can therefore cause signs of depression, anxiety and other mental illnesses. Multiple studies compared individuals who suffer from these mental health disorders to mentally healthy individuals [85][86]. These studies found a significant effect between higher stress levels (so lower HRV scores) and mental health disorders [87][88][89].

1.4 Why Wearables Could Be Useful

As mentioned, the popularity, acceptance and accuracy of wearables has risen over the years [1][2][3][15]. Compared to wearables, questionnaires about physical health information are less reliable due to the inability to capture real-time data. The correlations found between mental health and sleep, workout or heart rate information were indicated to exist through the use of self-reflection questionnaires and medical grade devices [15][16][17][18][19]. What makes the use of this method an even bigger problem is that mental disorders are shown to cause more unreliability in self-reflection questionnaires [90]. The main reason for this is that people with mental disorders see themselves differently than how others see them, Anthony et. al. (2017) [90] described this with the term "lack of insight". Since wearables make it possible for information such as sleep data, workout data and heart rate information to be collected in a non-intrusive way, these problematic experimental design flaws can be resolved, while also improving accuracy (in comparison to only self-reflection questionnaires).
The use of wearables in mental health related studies has been explored before [91]. Smith et. al. (2020) [92] used data originating from wearables for stress management intervention and Coutts et. al. (2020) [93] showed that a prediction could be made about mental health diseases such as depression and anxiety using HRV recordings from wearables. These studies used wearables that are only focused on tracking specific measurements, like HRV or even EEG recordings [94]. These are not popular consumer wearables so currently these prediction models are not directly applicable to the masses. Next to that, these earlier studies focused on mental health diseases, such as anxiety and depression. Because these diseases do not arise in a time span of 2 weeks, these studies had a time span of multiple weeks or months.
Most research that made use of wearables mostly used these wearables as a measurement for validation (compared to medical grade devices) or as a means to set physical health goals and track progress for these goals. Liau et. al. (2018) [95] used Fitbit wearables to see the effects of setting physical health goals on mental health and the Apple Watch was used to monitor mental health related symptoms, such as resilience [96], energy expenditure [97][98] and sleep monitoring [99]. The biggest difference to these earlier studies is that this study is not looking to validate the accuracy of wearables against medical grade devices or to see if monitoring certain health related symptoms are possible. This study is looking into whether popular consumer wearables can be used to find the correlation/make a prediction about mental state.

1.5 Research Outline

This study aims to use popular consumer wearables from brands such as Apple, Fitbit, Samsung and Garmin to handle the physical aspect of health, while using frequent 'positive mental well-being' assessments to capture the mental aspect of health. Earlier research has shown that there is a correlation between physical health and mental health [15][16][17][18][19] and that the prediction of mental health diseases based on wearable data is promising [93]. In contrast to this earlier research this study will use

popular consumer wearables and will try to find the correlation and make a prediction on the current mental state of an individual.
With this combination of mental health data and physical health data the aim was to answer two research questions. The first, *Can we use physical health data originating from consumer wearable technology to note a correlation between mental and physical health?* Since information such as sleep data, workout data and heart rate information was shown to impact mental health using questionnaires or medical grade devices, it would be interesting to see if the usage of consumer wearables could introduce a method which is more reliable (instead of self-reflection) and less intrusive (instead of medical grade devices).
A side note to this question is to see how much each wearable or wearable brand differs compared to each other, since every brand can use a different heart rate sensor or step count sensor. Once wearables have shown to be a reliable source of measuring this correlation we want to be able to use this information to create a 'personal health plan' in a predictive way: *How can the data from these wearables be used to create a model which is capable of predicting the mental state of an individual according to physical activity information?* If the data coming from the wearables can not only be used to keep track of physical health but correlation wise also 'conclude' something about mental state, wearables might become something more than just fun physical health trackers or digital fashion accessories.
The hypothesis states that a better/more positive mental state correlates with more activity, more sleep and lower stress values. More specifically for activity the amount of steps, the amount of workouts and the intensity of these workouts are expected to have a positive relation with mental well-being. For sleep the total sleep duration and the percentage of time spent in the NREM3 and REM sleep phases is expected to have a positive relation with mental well-being. More sleep sessions during the day are expected to correlate with a poorer/more negative mental well-being. For the heart rate information it is expected that a higher percentage of time spent with a higher stress level has a negative relation with an individual its mental well-being.
Since the frequency of the sleep and activity information showed to be correlated with mental health, we hypothesize that a more infrequent sleep pattern and more infrequent walking/workout pattern has a negative effect on mental well-being.

Methods 2

2.1 Participants

In total 32 participants participated in the experiment (13 woman, 19 men, aged between 18 and 62 with an average age of 28.9 and a standard deviation of 10.46). In total 43 individuals registered to participate in the experiment but not all completed the initial setup or participated after the initial setup (18 woman, 25 men, aged between 18 and 62 with an average age of 29 and a standard deviation of 10.38). These participants were mostly gathered through family, friends, and colleagues. Furthermore, advertisements were placed at public properties of the University of Groningen, as well as online social media platforms such as LinkedIn and Instagram.
The participants who completed the experiment received a mental and physical health summary after the experiment in which their physical activity and their answers to the mental well-being assessments are presented and evaluated (based on aspects in other studies such as average sleep duration, average amount of steps or average stress levels). To participate with the study participants had to fill in a consent form where they were instructed about the voluntary participation with the experiment, the data that would be collected and how this data would be used. In this consent form the participant agreed with wearing the wearable as much as possible during the day as well as while sleeping. In addition, the participants were informed about the duration of the experimental period, the anonymity of the gathered data and they were provided with contact details for possible questions. The participants were informed about how data would be collected and that they were free to leave the experiment without any negative consequences.
Derived from the PHQ-9 and GAD-7 questionnaires the participants had to fill in while setting up their account, 55% of the participants had a PHQ-9 score higher than 5, which already shows some mild signs of depression (see Table 2.1 for all PHQ-9 scores), and 36% of the participants had a GAD-7 score higher than 4, which shows some mild signs of anxiety (see Table 2.2 for all GAD-7 scores). The difference between genders showed to be neglectable.
It was decided to allow as much possible different wearable brands in the study since this introduces a bigger audience of possible participants. The division of wearables used in the study is visible in Figure 2.1. The biggest part of the participants used a Garmin wearable, which is capable of recording stress values.

Table 2.1: Summary of the PHQ-9 scores of the participants

Score	#	
0-4	15	Minimal signs of depression
5-9	14	Mild signs of depression
10-14	4	Moderate signs of depression
15-19	0	Moderate to severe signs of depression
≥20	0	Severe signs of depression

Table 2.2: Summary of the GAD-7 scores of the participants

Score	#	
0-4	21	No signs of an anxiety disorder
5-9	5	Mild signs of an anxiety disorder
10-14	4	Moderate signs of an anxiety disorder
≥15	3	Severe signs of an anxiety disorder

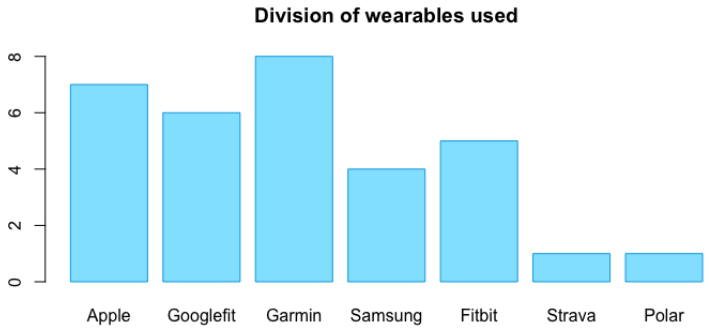


Figure 2.1: Division of wearable brands used by the participants of the study.
Garmin: 24%
Apple: 21%
Fitbit: 18%
Google Fit: 18%
Samsung: 12%
Strava: 3%
Polar: 3%

2.2 Materials

Several aspects were required to be able to conduct this experiment. We needed a method of receiving and storing the participant its physical activity information and we needed a secure method to fill in, store and receive the mental health information. Since this study is conducted solely online, we also needed to offer a safe way to participate online. To do this the platform Research-Health was created.

Research-Health

Research-Health is a Laravel [100] based online application which functions as the technical backbone of this experiment[1] . This platform was used for online registration (see Figure 2.2), secure authentication and data storage, mental well-being assessments, connecting to wearables as well as an admin panel to conduct the experiment. Every hour the server was scheduled to possibly send out notifications either by mail or through Firebase [101] to get notifications in the created Android app. This app is a Kotlin [102] based Android application which main functionality is to let the user receive notifications when it has to fill in a mental well-being assessment. The app is available on the Google Play Store[2] and can also be used to register into the experiment and start participating[3] [4] [5] .

WeFitter API

The WeFitter API [103] makes it possible to connect to wide range of wearable brands, and to more than 250 different wearable devices. At the time of conducting the experiment, WeFitter made it possible to connect to FitBit, Strava, Garmin, Google Fit, Withings, Polar, and Oura without any necessary development per connection. WeFitter handles the authentication to connect to these wearable brands and pulls data from connected wearables every ±10 minutes (depending on the wearable brand). This data is processed and returned as structured data in the form of daily summaries, biometric data, workouts, sleep summaries and stress samples/summaries. At the time of conducting this experiment these stress samples/summaries are only available on Garmin devices. If participants are wearing multiple wearables and did also connect multiple wearables, the data will be aggregated.
By using the WeFitter platform we do not need to implement the authentication methods, data retrieval methods and secure data storage methods for all the different wearable brands. This safes a lot of development time, server costs and maintenance.
Since both Apple and Samsung wearable devices do not offer an API but only an SDK connection, the decision was made to not implement these two brands into the research-health app (due to verification-, costs- and time-constraints). The already existing WeFitter app could be used together with the Research-Health platform. This meant that some manual work had to be done per participant to connect both platforms, which required more guidance initialising the participation with the experiment. This method was only used for participants who could be available for direct communication. This difference in app usage is not visible in the collected data and does not affect the outcome of this study in any way.

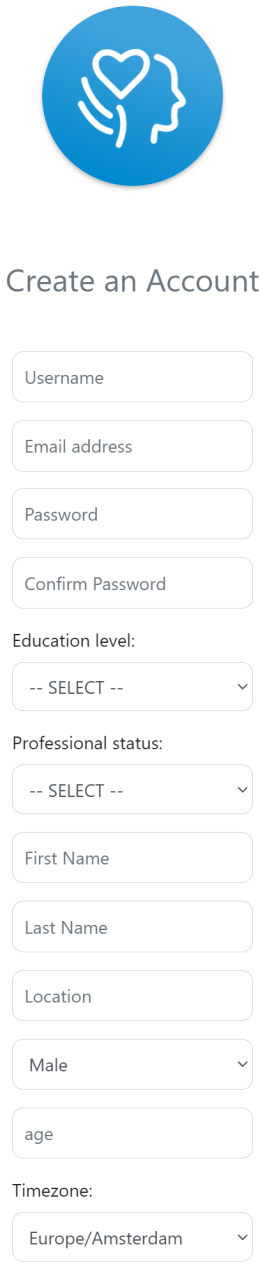


Figure 2.2: The registration screen of research-health.

1: Research-Health is available at https://www.research-health.com/

2: The Research-Health Android application is available at https://play.google.com/store/apps/details?id=com.researchhealth

3: In order to be transparent with the participants about the private data being shared by the app, a privacy statement is available on the website: https://research-health.com/privacy_policy_app

4: The website has its own privacy policy: https://research-health.com/privacy_policy

Physical Health Data

The health data from the wearables cover a wide variety of domains and different time intervals. To see if the data originating from wearables could be used to show a correlation between physical health and mental health, we will use the WeFitter models; daily summaries, workouts, heart rate summaries, sleep summaries and stress samples. All the different wearable brands have a different way of structuring their data. WeFitter restructures this uniformly without losing necessary data, so no conversion is required anymore on our side. The different models will be explained more in detail:

Daily Summaries

The daily summary objects from WeFitter offer a way to get a general sense of how active the participant was on a certain date. It included the amount of steps, calories and distance of a person a day, while some wearable devices include a heart rate summary with minimum, maximum and average heart rate values. The calories variable is an estimation based on the basal metabolic rate (BMR) and the amount of steps traveled and the heart rate during these steps. This daily summary

5: The website is also accompanied with a cookie policy to adhere to European guidelines: https://research-health.com/cookie_policy

can be used to cross validate earlier studies that showed a correlation between the amount of walks and mental health of a person [76][77].

Workouts

The workouts objects from WeFitter are manually or automatically tracked exercises such as cycling or taking a longer walk. A workout has a timestamp, duration, type (like running or cycling), tracking method (automatic or manual), distance, steps, calories, and a possible heart rate summary (min. max. avg.). A wearable automatically starts tracking a workout by looking at multiple aspects, such as heart rate, amount of steps, and GPS location. This means that most often these automatically tracked exercises are aerobic exercises.

Heart Rate Data

Value	Class
0-25	Rest
26-50	Low
51-75	Medium
76-100	High

Figure 2.3: The classification of stress scores according to Garmin.

At the time of conducting the experiment, the only wearable brand that uses heart rate variability to estimate a stress value is Garmin. Garmin is dividing this stress score into 4 separate stress levels classified as rest (0-25), low stress (26-50), moderate stress (51, 75), and high stress (76-100). This data is recorded and collected every 3 minutes (if possible). This stress score is determined by looking at the heart rate variability and the domain of these stress scores are dependant on the wearer of the wearable. If the person wears the wearable often, the stress scores are adapted to their personal condition, if not the stress scores are based on global averages. All participants with a Garmin connection already owned the watch for more than a month, which is enough to personalize the stress scores. This means that these scores can be used as an indication of stress levels over time.

Sleep Summaries

The sleep summary objects are recorded if the participant is wearing their wearable while they are sleeping. This object contains a timestamp, duration, duration awake (NREM1), duration in light sleep (NREM2), duration in deep sleep (NREM3), duration in rem sleep (REM), and total time in sleep. The difference in sleep stages is determined using the accelerometer and the heart rate of the person wearing the wearable. The accuracy of sleep tracking in the wearables differs per brand. Although reliable to summarize a night of sleep, the actual accuracy of the measurements compared to (current) medical grade devices is between 70% and 80% [104][105][106]. This lower accuracy is not expected to be problematic since the percentages of time spent in sleep stages will be used, this does not have to be accurate on a minute per minute basis.

Positive Mental Well-being Assessment

To get a good assessment about mental health looking at the 5 different aspects, positive affective state, rumination, resilience, cognitive complaints and depression, several items from multiple well-validated questionnaires were combined. First of all a part of the Positive and

Negative Affect Schedule (PANAS) was used, which involves repeated and frequent sampling of the current mood of the participant based on certain words. More specifically the PANAS-X scale was used since this makes it possible to measure at different time intervals [107][108]. In comparison to the PHQ-9 and GAD-7 questionnaires, these mental health assessments were asked twice a day and specifically ask for the current state of the participant instead of the past two weeks.
To test for rumination and resilience we respectively used parts of the Ruminative Response Scale (which also tests for depression) [109][110] and the Resilience Scale [111]. To test for cognitive complaints, statements about concentration and memory are used since these strongly correlate with symptoms of mental disorders [30]. Another symptom of mental illness can be physical complaints [29], so a question about how physically uncomfortable someone is was also added.
The statements had to be scored on a 5 point Likert scale where 5 means strongly agree and 0 means strongly disagree. The statement about the location can be answered by several options, see table 2.3. This statement was added to get a better idea about the participant its environment and state while filling in the questionnaire. The order of these statements were randomized when presented to the participant, so that the participants were less prone to recognize a pattern and fill in the same scores.

	question	**subject**
1	At this moment I feel Energetic	PAS
2	At this moment I feel Happy	PAS
3	At this moment I feel Satisfied	PAS
4	At this moment I am focused on my feelings	R
5	At this moment I am worried about problems	R
6	At this moment I feel like I would be able to effectively cope with a negative life event	RES
7	At this moment I feel forgetful	COG
8	At this moment I feel inattentive	COG
9	At this moment I experience physical discomfort	COG
10	At this moment I feel depressed	D
11	At this moment I feel hopeful about my future	D
12	At his moment I feel stressed / I am too busy	R
13	In the last 2 hours I spent most of my time: - At home - At School / Work - With Family / Friends - On the move - In nature - At the gym - On vacation - Somewhere else	Location

Table 2.3: Statements which are shown in the positive mental well-being assessments.
Statements 1 - 12 are scored on a 5 point Likert scale.
PAS = positive affective state.
R = rumination.
RES = resilience.
COG = cognitive complaints.
D = depression.

2.3 Procedure

A participant can create an account to participate online[6] , where it had to fill in some personal details (see Figure 2.2). On this screen the participant had to consent to participate, as well as agree to the terms and conditions of the platform [7] . After registration the user was manually approved or declined to prevent spam and a verification email was sent to the participant. After verifying the account the participant can start the experiment and does this by filling in a PHQ-9 and GAD-7 questionnaire to give an indication of their prior mental state. The participant can now connect to a wearable brand by pressing one of the wearable brand

6: The registration page is available at https://www.research-health.com/register/

7: The terms and conditions are available at https://research-health.com/terms_conditions

logos (see the right screen in Figure 2.4). After a connection is made, the experimental period of two weeks begins.

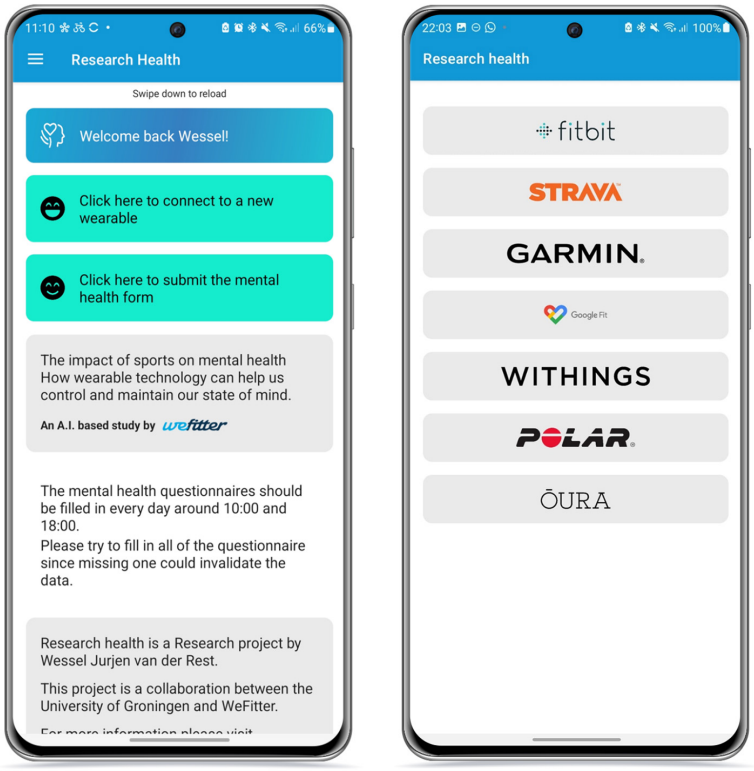


Figure 2.4: The Research-Health app.
Left: shows the home screen where a participant can connect to new wearables and fill in the questionnaire
Right: the wearable connections screen, on which the participant can connect to a new wearable.

Experiment Period

During the experimental period the participant is notified at 10:00 in the morning and 18:00 in the afternoon about filling in the positive mental well-being assessment. If the participant forgets to fill in the questionnaire it will be reminded hourly for the next three hours, either by app or by mail (the participant can set its own preference). If the participant follows the link in the notification it will be brought to the mental well-being assessment page where the participants has to fill in the questionnaire. The participant is also able to fill in the questionnaire earlier or later than when the notifications are sent, this functionality is added since the time slots may be inconvenient for some participants. Only allowing entries in certain time frames, for example between 10:00 and 10:30 will result into a huge loss of data. A participant cannot send in 2 questionnaires within 4 hours of submitting one. WeFitter pulls data from the wearables every ±10 minutes and stores this on their systems. All log files from the app and the platform are stored securely on the server and on Google Cloud Storage for Firebase. These log files together with the questionnaires were used to track progress and activity. In total 6 different participants showed some inactive behaviour by not submitting a questionnaire in at least 3 hours after the notification. Two of those participants had to be reminded more than once and contact was made to urge the participant to fill in the questionnaires or to check the connection with their wearable (1 participant just bought a new wearable 1 day into the experiment, so she had to reconnect).

Post experiment

When the participant has submitted 28 questionnaires, or is participating for 2 weeks, the notifications will stop and the participant has finished the experiment. After participation the participant receives an email with a link to a summary about their physical and mental health. The purpose of this email is that the participant can get a health check. Information is provided that may help them live a healthier life. For example if it is visible that they show some problems getting a good night of sleep, tips will be provided to try and prevent this. Their answers to the GAD-7 and PHQ-9 questionnaires and their answers to the positive mental well-being questionnaires are presented and discussed. Furthermore the amount of steps, the average sleep duration and the amount of workouts are compared to values from earlier studies and anonymously how they rank compared to other participants. The physical health data is downloaded from the WeFitter API and stored anonymously.

2.4 Pre-processing and data mapping

Each physical health data type is converted to values per questionnaire, and all data points measured between questionnaires are added to the following questionnaire. So for example the amount of workouts done after the questionnaire of 18:00 on Monday afternoon, will be added to the questionnaire of 10:00 on Tuesday morning.
Since the time of filling in the questionnaires is not exactly the same per participant, the questionnaires will be classified either as morning or afternoon submissions. The data is not yet directly connected to the questionnaires so some conversions have to be done to match data from the morning or afternoon time frames.

Daily Summary Mapping

The daily summaries show an activity summary of the whole day. This means that the daily information for steps taken, distance traveled and calories burned are returned as one value per day. To take into account the progress of activity over a day the total amount is estimated according to the average active time of the day. The active part of the day is based on the average sleep wake times found in previously existing WeFitter data. The average wake time is 7:02 in the morning and the average sleep time is 23:40 in the evening. This means that on average a person is 'active' for more than 16 hours a day.
From the daily summary objects two different variables are composed which should be correlated with mental health according to earlier studies [76][77][78]:

1. Daily activity intensity.
2. Daily activity frequency.

For the daily activity intensity only the amount of steps taken are used. Because steps, distance and calories showed to be highly correlated (Pearson correlation test showed p-values < 0.001 with correlation coefficients

between 0.66 and 0.90). The amount of steps taken were chosen because this is the easiest of the three measurements to be accurately recorded, and is also easiest to understand as a user.
The daily activity frequency is added since the infrequency in activity showed to have a negative relation with mental health. This variable is composed by getting the standard deviation of the daily steps over the last 3 days.

Workout Mapping

Workouts happen at an irregular moment in time and are also variable in duration. Because of this and the expectation from earlier studies [61][62][63][64][65] the workout data between the mentioned time frames is mapped to the questionnaires in the following fields:

4. Workout intensity.
5. Time since last workout.
6. Workout frequency.

The calculated workout intensity is based on the workout type, the workout duration and the total amount of calories burned during the workout. The workout types are ordered by WeFitter on their aerobic score where mindful has the lowest score and strength training has the highest score (see Table 2.4). The calculation of the workout intensity score can be seen in Equation 2.1, where n is the amount of workouts, $duration_i$ the duration of the i-th workout, $calories_i$ are the amount of burned calories of the i-th workout, and $score_i$ the aerobic score of the i-th workout.
The workout since last variable is the amount of minutes since the last workout was performed.
Similar to daily activity information, the workout frequency will be taken by looking at the last 3 days (not available for the first 2 days). The average amount of workouts in the last 3 days is used as workout frequency.

$$WorkoutIntensity = \sum_{i=1}^{n}(duration_i * calories_i * score_i) \tag{2.1}$$

Table 2.4: The workout types and their aerobic scores.
The workout type "Other" contains a wide range of workouts, based on all connections and all possibilities its aerobic score was scored between rowing and swimming.

Workout type	Aerobic score
Mindfulness	1
Yoga	2
Walking	3
Cycling	4
Running	5
Rowing	6
Other	7
Swimming	8
Crossfit	9
Strength Training	10

Sleep Mapping

At first, it seems to be possible to map sleep data directly to a questionnaire, but this would neglect smaller possible naps during the day. Furthermore this data is mapped to both the morning and afternoon questionnaire,

since a bad night of sleep might have a bigger effect at the end of the next day compared to the beginning of that day. The smaller possible naps during the day that occurred after the submission of the previous questionnaire are only mapped to the next questionnaire. According to expectations from earlier studies [44][50][51][18][45] the sleep data is mapped to the following fields:

7. Sleep duration.
8. Sleep depth.
9. Sleep onset frequency.
10. Sleep duration frequency.

The sleep duration is the total duration which the wearable classifies as sleeping, this may include laying awake in bed. The sleep depth is the percentage of time spent in NREM3 or REM stage.
The sleep onset frequency and the sleep duration frequency are composed by getting the standard deviation of the main sleep recording onset time and duration of the last three days (both not available for the first 2 days). The main sleep recording is selected by getting the longest duration of the sleep recordings in the specified interval. If there is no sleep data for all of the last three days, the frequency values will be nulled.

Stress Mapping

Earlier studies showed that the stress values are correlated with mental health disorders [87][88][89]. Mental health disorders such as depression and anxiety showed to be accompanied with higher stress scores. The stress data from Garmin devices is coming in every 3 minutes (if available), The mean of these values over the mentioned time interval is used as the last variable.

2.5 Data imputation

Out of the 32 participants, only 20 recorded sleep data, only 14 recorded sleep phase data and only 8 recorded stress data. Only 7 participants recorded both sleep duration, sleep phase and stress data, but not during all days of the experiment. If all the assessments are excluded where no stress or sleep data is recorded, only 55 out of the 727 assessments would be used. To still make use of all the assessments and the activity and workout data that accompany them, some data imputation was performed on the data set. Because of the small amount of participants the decision was made to solely do median imputation. This replaces all missing values with the median value of that datatype. So for stress data all the missing values were replaced by 22.25. For sleep duration the missing values were replaced by 492 minutes of sleep (8.2 hours) and for sleep depth the missing values were replaced by 33.9% of time in NREM3 and REM sleep phase. The use of data imputation means that the specificity in the data set becomes lower, while the general effects of difference in activity, workouts, sleep and stress data on positive mental well-being should still be visible.

2.6 Data Transformation and Scaling

To prepare the data set for analysis some transformations, grouping and scaling are performed.
First of all the data types are mostly skewed, if we use this skewed data in the analysis, the spectrum of values wont be used equally. To transform this skew in a more normally distributed data set the variables are transformed using the Yeo-Johnson power transformation [112].
After this transformation we want to make sure that all data types are on a comparable scale, currently the amount of steps taken is much higher than for example the sleep depth. To scale the variables, the pre-processing module from the sklearn [113] library is used. From this module the StandardScaler was used to transform all used features in the data set. This scaler standardizes the data set by removing the mean values and scaling it to unit variance. This is important for feature importance selection.

2.7 Analysis

To analyse the results and see whether the physical information from the wearables can accurately be used to show a correlation between physical health and mental well-being, statistical analyses were carried out using the statistical programming language R [114] in RStudio [115].
Research-Health has an export function that returns all physical and mental data per participant as a JSON file, all files from all participants are combined and transformed to a data set that has the information stored per submitted questionnaire. Only participants who actively tried to participate in the experiment and filled in multiple questionnaires for a duration of 2 weeks were included in the analysis. Some participants did not fill in all 28 questionnaires, this does not mean that the participant is completely excluded from participating.
In total 727 questionnaires were filled in from the expected 924 (78%). This unfortunately means that some of the physical health data went lost because it couldn't get mapped to mental health questionnaires. On average the questionnaire was submitted 61 minutes too late (52% of the questionnaires where submitted within the first hour of receiving the notification).
To analyse the possible correlations between the physical and mental health data Linear Mixed Effects models (LME) were created using the $lme4$ package [116]. Since multiple aspects about physical health are used in this study a correlation matrix is made comparing all the different variables described in the pre-processing section (see Appendix.2). This correlation matrix showed a strong correlation between the workout intensity and the time since the last workout. This correlation is confirmed using the Pearson correlation test, which showed a p-value of $4.355e^{-05}$ with a correlation coefficient of -0.15. Because of the lower number of entries with a known value for the time since the last workout, this variable was excluded from the models.
The frequency data will be evaluated in a separate linear mixed effects model since this data is based on longer term recordings. Since most previous studies about the correlation between mental and physical health are based on longer period recordings, the difference between the

two linear mixed effects models is important to be analyzed.
A composite variable positive mental well-being is composed by inverting the scores of the negatively formulated questions from the positively formulated questions. This score is an ambiguous term referencing to how positive or stable someone is mentally. The division of positive and negative questions was confirmed using K-Means clustering and factor analysis. Horn's (1965) "parallel" analysis [117] and the gap statistic [118] stated that 2 clusters could be the optimal number of clusters in the questionnaire data (see Appendix Figure.1 and Table.1 for the results of K-Means clustering and factor analysis). The 'question types' (Positive affective state, rumination, resilience, cognitive complaints, and depression) are all summed scores of the answers to the questions from Table 2.3, except for the depression score. Here the answer to the second question "I feel hopeful about my future" is inverted when used with the answer to the first question "I feel depressed".

Table 2.5: Table of how the question are classified in a positive and negative class. The left column are all positive formulated questions, the left column all negative formulated questions.

Positive	**Negative**
At this moment I feel energetic	At this moment I am focused on my feelings
At this moment I feel happy	At this moment I am worried about problems
At this moment I feel satisfied	At this moment I feel forgetful
At this moment I feel like I would be able to effectively cope with a negative life event	At this moment I feel inattentive
At this moment I feel hopeful about my future	At this moment I experience physical discomfort
	At this moment I feel depressed
	At his moment I feel stressed / I am too busy

The Satterthwaite Formula for Degrees of Freedom was used to calculate the p-values using the $lmerTest$ package [119]. Each participant and each date is used as a random variable, while the physical health parameters (and the initial mental health questionnaires) are used as fixed variables. These fixed variables are used as a predictor with which the composite variable positive mental well-being is predicted.
Maximum likelihood estimation and ANOVA were used to compare models and to create the optimal model which explained the most variance. The AIC values of the models were compared to see which model fits the best and a more complex model will only be preferred if its AIC value is at least 2 points lower.
Next to the model that is predicting the composite variable positive mental well-being, separate models were created where each of the question types are used as the predicted values.

2.8 Prediction model

To see if the collected data could be used to make a model capable of predicting ones "positive mental well-being" based on data originating from consumer wearables, multiple classification methods were tested and compared to each other. To stay impartial to the data set, eight of the most popular classification algorithms from the scikit-learn library[113] were

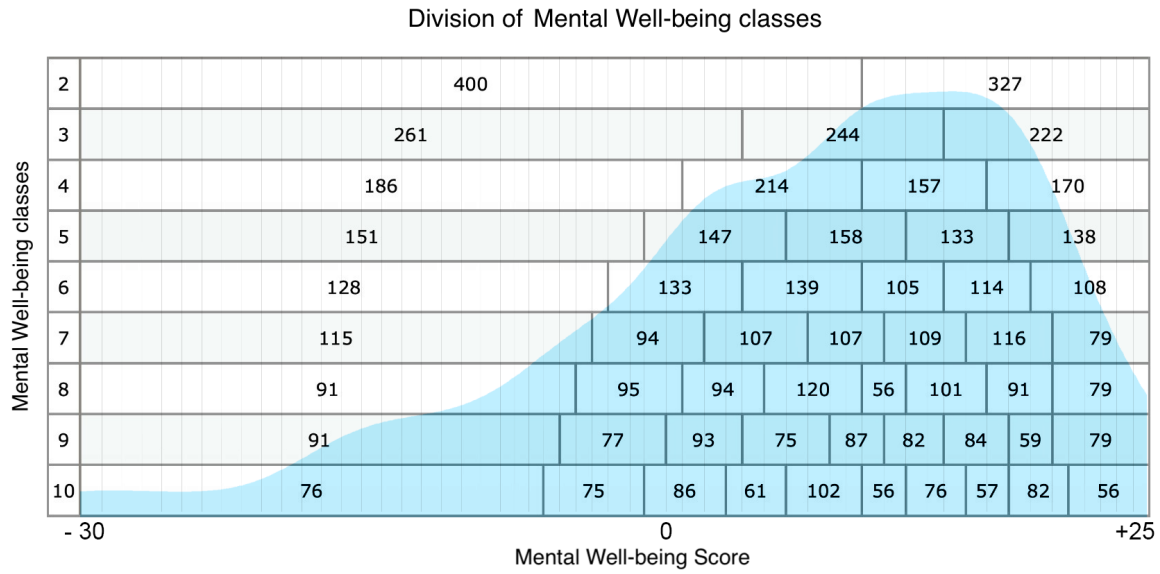


Figure 2.5: Division of the mental well-being scores. The numbers in the sections indicate the amount of questionnaires gathered within that mental well-being range. The curve in the background shows the distribution of available assessments.

used in the python programming language [120] to make a cross comparison on model performance; naive bayes, logistic regression, K-nearest neighbors, Support vector machine, decision trees, linear discriminant analysis and ensemble learning (random forests, and ADA-boost).

The scale of the positive mental well-being variable now ranges from -30 to +25, so a 55-point scale. It was decided to make this less specific and categorize this data. To see how specific our predictive model could classify the positive mental well-being, the data was split into 9 different variants. The least specific variant included 2 classes, so either a positive or a negative class of mental well-being. The most specific variant included 10 different classes, so that positive mental well-being could be scored on a 10 point scale. This split was done using pandas its [121] quantile-based discretization function, this function groups the values into similarly sized buckets. Figure 2.5 shows how the different classes of positive mental well-being are divided and how much recordings are available per class. Each horizontal line shows the number of classes, and each number in each 'box' shows the amount of recordings in that group. The shaded curve in the background shows the histogram of available recordings and it is visible that most recordings are skewed to the positive part of the spectrum. The Shapiro-Wilk normality test showed a p-value < 0.05, so the data is not normally distributed. Other forms of data normalisation did not show to improve the normality distribution.

For each algorithm model optimization was performed based on when the classification was done on 3 different mental well-being classes. This amount of classes was chosen because this is the amount that could still be usable (negative-, neutral- and positive- mental well-being) while still showing some promising accuracies. The individual model specifications and its parameter settings are visible in Appendix.1. Further optimization includes that the models were trained with 10-KFold cross validation to avoid the chance of over-fitting.

The variables that showed to have a significant effect on positive mental

well-being in the linear mixed effects models were used as features during model training (daily steps, workout intensity, mean stress score and workout frequency). The addition of the PHQ and GAD values is again explored by comparing the average model accuracies. These PHQ and GAD scores are categorized based on the values visible in Table 2.1 and Table 2.2 to be able to make a more generalized model.

Results 3

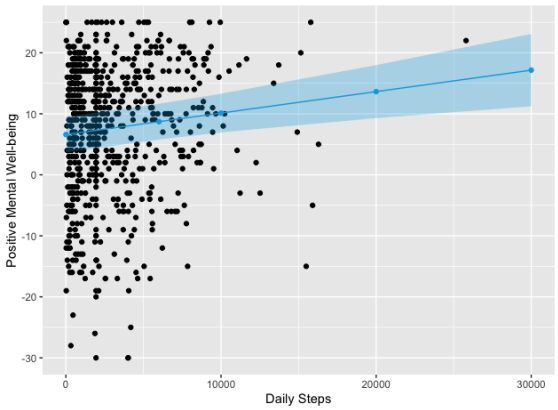
First the results of the linear mixed effects models will be presented, where the effects of the different data types on the positive mental well-being will be shown. The effects that are based on daily frequency or difference are modelled in a separate linear mixed effects model since this requires consecutive daily information which not all participants have (so less data is available). After this the predictive model will be presented in section 3.2.

3.1 Linear Mixed Effects

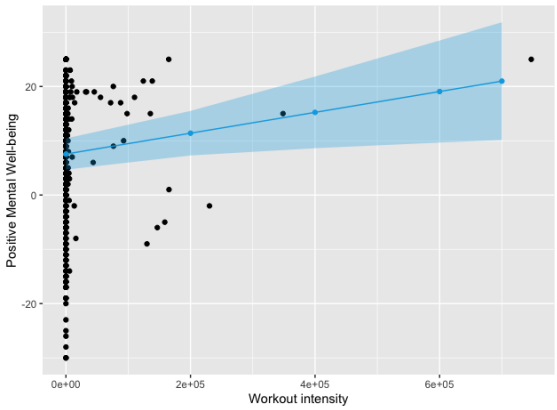
Daily information

To see if positive mental well-being is significantly affected by workout, sleep and stress information originating from consumer wearables, the initial model included the data types mentioned in the methods section. To specify, the created model includes the daily activity steps, the workout intensity, the sleep duration and depth, and the mean stress scores as fixed effects. The participants and date are used as random effects. The predicted value will be the composite variable positive mental well-being. In this model (see Table Appendix.3) it is visible that the higher amount of daily steps, the higher the mental well-being score (so the better a participant felt)(Est. = 3.517e-04, p-value < 0.001). The same significant relation can be seen between the workout intensity and the mental well-being score (Est. = 1.919e-05, p-value < 0.05). Additionally, the mean stress score showed to have a negative relation with mental well-being, the higher the mean stress score, the worse a participant felt (Est. = -1.230e-01, p-value < 0.05). This model showed a marginal R^2 score of 0.019 and a conditional R^2 score of 0.57.
The effect of the daily amount of steps taken on positive mental well-being are shown in Figure 3.1a. The shaded area in the figure shows that the score of positive mental well-being increases whenever more steps are taken. For every ±270 steps taken, the linear mixed effects model estimates that the positive mental well-being score improves by one point.
Figure 3.1b shows the effects of the workout intensity on the mental well-being of a person. The shaded area in the plot shows a positive effect from workout intensity on the mental well-being. This estimation shows that whenever the intensity of a workout is ±50.000 points higher, the mental well-being improves by one point. Using the workout intensity equation (see Equation 2.1) this is similar to a bike ride of 30 minutes where 450 kcal are burned.
The effects of stress on mental well-being are shown in Figure 3.1c. The shaded area shows a negative effect of stress on positive mental well-being. The models shows that the mental well-being score decreases with one point whenever the mean stress score increases by ±8 points (on a

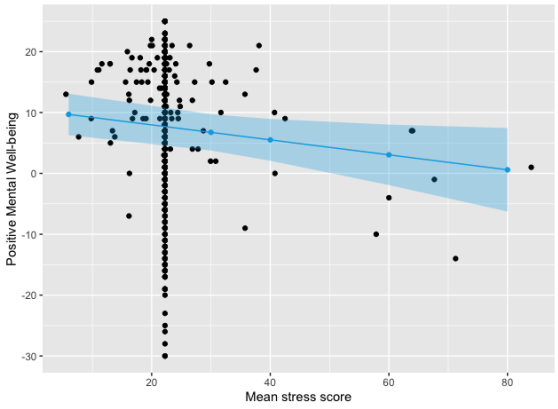
scale form 0 to 100).



(a) Effects of the amount of steps taken on the positive mental well-being in the linear mixed effects model on daily information.



(b) Effects of the workout intensity on positive mental well-being in the linear mixed effects model on daily information.



(c) Effects of mean stress scores on the positive mental well-being in the linear mixed effects model on daily information.

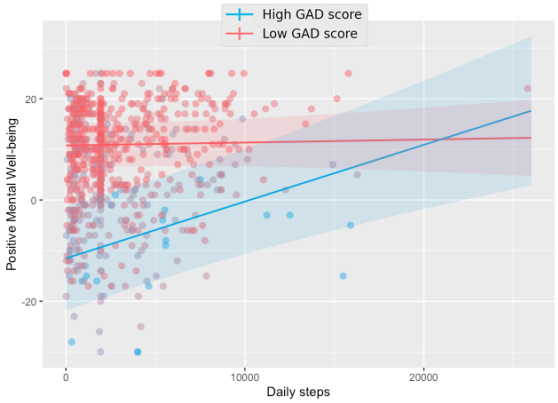
Figure 3.1: Plots of the linear mixed effects model with daily information. The shaded areas indicate 95% confidence intervals. The dots indicate the sample points. The results in these plots are converted back to the original values instead of showing the normalized data.

The Addition of PHQ and GAD scores

A new model was able to explain more variance whenever the PHQ and the GAD scores, and their interaction with some physical health features, were added ($\chi^2(8)$=59.514, p-value < 0.001, see Table Appendix.4). To specify, the interaction between the PHQ score and the sleep duration, and the interactions between the GAD score and the daily steps, workout duration, sleep duration and the PHQ score were added. The marginal R^2 score of this model is 0.33, while the conditional decreases a bit to 0.59.

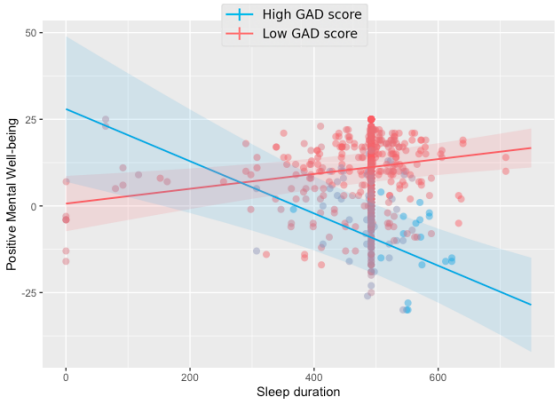
The significant effects in this model have changed, the PHQ score has a significant negative effect on mental well-being (Est. -2.24, p-value < 0.05) as does the mean stress score (Est. -1.20e-1, p-value < 0.05). Furthermore the interactions between the GAD score and the daily steps (Est. 5.06e-5, p-value < 0.01, see Figure 3.2a) and the workout intensity (Est. 3.50e-6, p-value < 0.01, see Figure 3.2b) have a significant positive effect on mental well-being, and the interaction between GAD score and the sleep duration

has a significant negative effect (Est. -4.97e-3, p-value < 0.05, see Figure 3.2c).
The plots in Figure 3.2 show the significant effect of the interactions between the GAD score and the amount of steps, workout intensity and sleep duration on positive mental well-being. For Figure 3.2a and Figure 3.2b it is visible that the effect of more steps and a higher workout intensity is strongly positive with individuals who suffer from a anxiety (a high GAD score of 21). For individuals without anxiety (a low GAD score of 0) the effects are only slightly positive, almost neutral.
The significant effect of the interaction between the GAD score and the sleep duration is visible in Figure 3.2c. This plot shows that this effect is positive for individuals with anxiety and negative for individuals without anxiety.



(a) Interaction of the Daily Steps and the GAD score on positive mental well-being. The red scores represent a low GAD score of 0, and the blue scores represent a high GAD score of 21.

(b) Interaction of the Workout Intensity and the GAD score on positive mental well-being. The red scores represent a low GAD score of 0, and the blue scores represent a high GAD score of 21.



(c) Interaction of the Sleep duration and the GAD score on positive mental well-being. The red scores represent a low GAD score of 0, and the blue scores represent a high GAD score of 21.

Figure 3.2: The significant interactions found in the linear regression model including the PHQ and GAD scores. The dots indicate the sample points. The results in these plots are converted back to the original values instead of showing the normalized data.

Frequency information

The frequency or daily difference of physical health data showed to be an important factor of mental health in previous research [57][71][72][78].

Since this data is based on the differences or similarities between multiple consecutive days and less participants filled in the questionnaire on consecutive days, the decision was made to make a separate model for these data types. To be more specific, a linear mixed effects model was created to see if any significant effects could be found between the positive mental well-being of an individual and the daily activity difference, workout frequency, sleep start frequency and sleep duration frequency of that individual. The last four mentioned variables are used as fixed effects, while the date and participant are again used as random effects. The predicted variable is once more the composite variable positive mental well-being.
Again, the addition of the PHQ and GAD scores and their interaction with some parameters showed to explain more variance to the model ($\chi^2(7)$=43.964, p-value < 0.001). To specify, the interaction between the PHQ score and the sleep onset frequency and the interactions between the GAD score and the daily activity frequency, sleep onset frequency and the PHQ score were added. The linear mixed effects model (see Appendix.5) shows that there is a significant positive effect of a higher workout frequency on positive mental well-being. Whenever the frequency over the last 3 days is on average 1 workout higher, the positive mental well-being increases by almost 2 points (Est. 1.98, p-value < 0.001, see Figure 3.3). The marginal R^2 score of this model is 0.33 and the conditional equals 0.58.

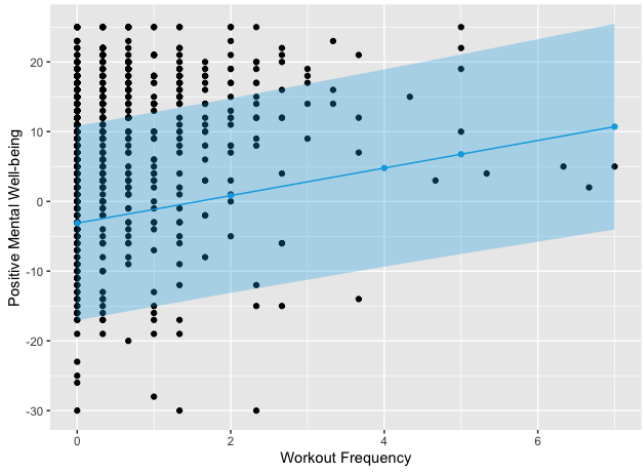


Figure 3.3: The effect of the workout frequency over the last 3 days on positive mental well-being. The shaded areas indicate the 95% confidence intervals. The dots indicate the sample points.

Effects on Question Types

To see the effects of the variables on each question type, separate models were created where the predicted variable was set to the scores for each question type. These models do not include the interactions between the PHQ or GAD scores and the physical health parameters. This decision was made to focus mostly on the effects of the physical health parameters on the individual question types. The estimations and significance from the effects found in these models are visible in Table 3.1. It is visible that when positive affective state is the predicted variable, a significant effect is found for the daily steps (Est. = 1.87, p-value < 0.001), workout intensity (Est. = 7.51, p-value < 0.05), sleep depth (Est. = -5.24, p-value < 0.05) and mean stress score (Est. = -4.98, p-value < 0.05). For rumination a significant effect was found on the GAD score (Est. = 2.71, p-value < 0.01),

and for resilience a significant effect was found for the mean stress score (Est. = -2.25, p-value < 0.01). The cognitive complaints are significantly effected by both the GAD (Est. = 1.85, p-value < 0.05) and PHQ scores (Est. = 2.60, p-value < 0.05) and depression by the daily steps (Est. = -7.27, p-value < 0.001) and the workout intensity (Est. = -3.06, p-value < 0.05). For the frequency information again only the workout frequency showed to have a significant effect on all question types except for rumination. The significant effects on positive affective state (Est. = 7.14, p-value < 0.001) and resilience (Est. = 1.99, p-value < 0.01) are positive. The effects on cognitive complaints (Est. = -4.76, p-value < 0.001) and depression (Est. = -2.30, p-value < 0.01) are negative.

Table 3.1: Table that shows the estimates and the significance of variables in different linear mixed effects models. The question types are visible on the horizontal axis of the table and represent the positive affective state (PAS), rumination (R), resilience (RES), cognitive complaints (COG), and depression (D) (see Table 2.3). Each question type was used as the predicted variable in different linear mixed effects models. The estimations of each variable are visible, the color and stars show the significance of the variable in the model.
Legend: * p < .05 ** p < .01 *** p < .001

	PAS	R	RES	COG	D
daily steps	1.865e-04***	-4.543e-05	2.455e-05	-2.624e-05	-7.269e-05***
workout intensity	7.511e-06*	-3.176e-06	1.156e-06	-4.459e-06	-3.063e-06*
sleep duration	-2.055e-05	1.639e-03	8.343e-04	-8.945e-04	-2.625e-04
sleep depth	-5.236e-02*	-7.962e-03	-1.378e-02	3.036e-02	1.951e-02
stress mean	-4.975e-02*	1.872e-02	-2.254e-02**	1.707e-02	1.501e-02
GAD	-1.117e-01	2.712e-01**	-6.813e-02	1.849e-01*	1.121e-01
PHQ	-2.597e-01	5.710e-02	-5.487e-02	2.600e-01*	6.723e-02
Workout Freq.	7.143e-01***	-2.456e-01	1.991e-01**	-4.759e-01***	-2.295e-01**

3.2 Prediction Model

It would be ideal if these correlations could be used to make a prediction model about mental well-being. This feature could be used to detect deteriorating trends early, and possibly act upon it.
The different algorithms only showed some slight differences in classification performance. Figure 3.4 shows the classification performances of each of the models through cross validation. It is visible that the average accuracy over all models is ±38% (see the blue boxplots). This figure also shows the balanced accuracy scores for these classification methods (see the red boxplots). These balanced accuracy scores remove the guessing chance from the model performance. It is visible that the mean balanced accuracy values of all classifiers are close to the chance level of 0. The differences between algorithms are quite small but the CART classifier shows to have the highest mean accuracy. This classifier is used to compare the effects of an increase in amount of mental well-being classes, the permutation importance of each of the used features and the predictive capabilities on separate question types. In some cases the CART classifier will be compared to other algorithms.
The difference in the amount of mental well-being classes showed to have a significant effect on the performance of the models. Figure 3.5 shows the model accuracy and balanced accuracy against the amount of to-be predicted classes. When the mental well-being score is divided into two different classes, the average accuracy of the model is 58%. This accuracy quickly drops when more classes are added. Whenever a classification is

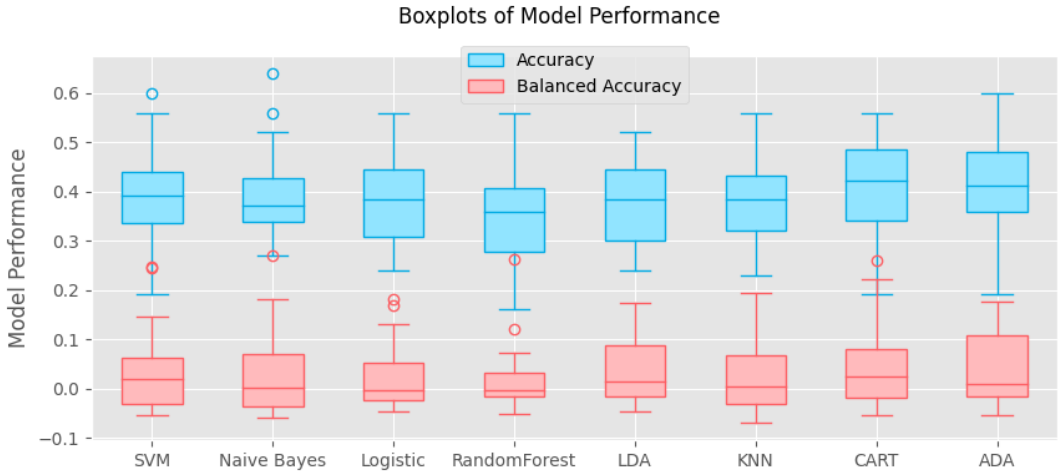


Figure 3.4: The cross validated accuracy and balanced accuracy scores of the different classification models when 3 classes of mental well-being are predicted.
The chance level for balanced accuracy is at a model performance of 0.
SVM = Support Vector Machine
Naive Bayes = Gaussian Naive Bayes
Logistic = Logistic Regression
RandomForest = Random Decision Forests (Ensemble Learning)
LDA = Linear Discriminant Analysis
KNN = K-Nearest Neighbors
CART = Classification and Regression Decision Tree
ADA = ADA Boost (Ensemble Learning)

done on 3 different classes of mental well-being, the accuracy already drops to 38%. The results of a binomial tests on these values show a p-value > 0.05 for all of these models, this shows that we cannot reject the null hypothesis. There is not sufficient evidence to accurately predict the mental well-being of an individual. The balanced accuracy also indicates this, which stays close to the chance level of 0 (see the red line in Figure 3.5). The addition or removal of physical health features did not seem to improve model performance. To see how much the model performance is dependant on certain features, the permutation importance of each

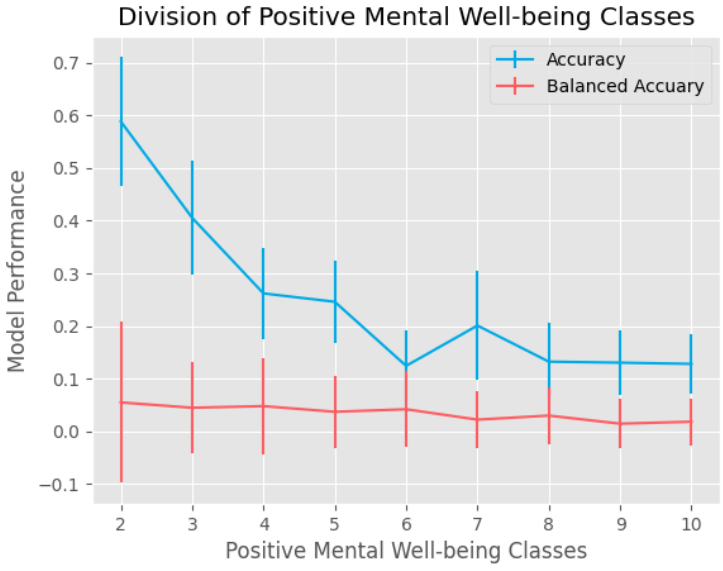
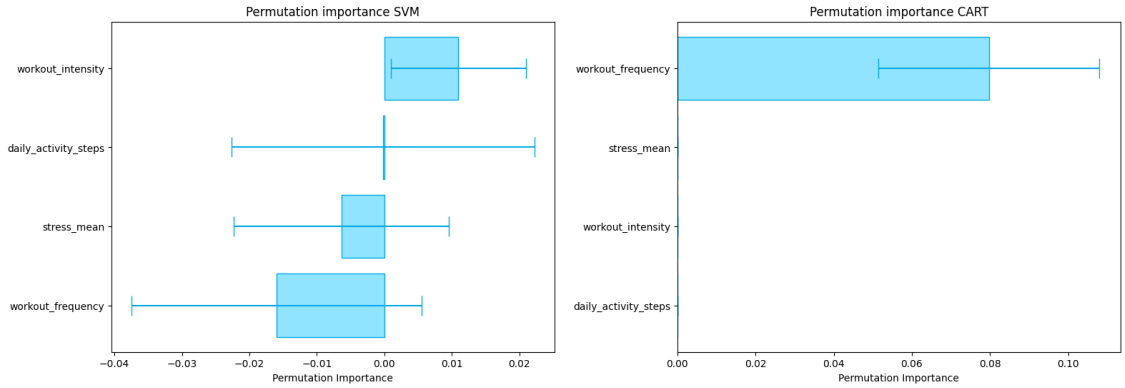


Figure 3.5: The accuracy and balanced accuracy of the CART classification per amount of to-be predicted mental well-being classes.
The chance level for balanced accuracy is at a model performance of 0.



(a) Permutation importance of the features in the Support Vector Machine classifier.

(b) Permutation importance of the features in the CART classifier.

Figure 3.6: Permutation importance plots for the prediction of 3 mental well-being classes using the CART classifier and the Support Vector Machine classifier.

feature for the CART classifier and the SVM classifier are presented in Figure 3.6. These two models are presented because of similar accuracies (see Figure 3.4), but big differences in permutation importances. The plot of the support vector machine model is visible in Figure 3.6b and Figure 3.6a shows the permutation importance for the CART classifier. It is visible that the $workout\_intensity$ feature shows to be the most important feature for the SVM model, while not being used at all in the CART classifier. The negative values mean that using a random value for that feature showed to increase the model performance in some cases.
This low classification accuracy is also visible is the confusion matrices. The CART classifier was used to create confusion matrices which are visible for respectively 2, 3 and 5 mental well-being classes in Table 3.2, Table 3.3 and Table 3.4. Especially in the second and third tables it is visible that the predicted classes do not correlate correctly with the actual classes. In the last table only 2 out 5 different classes are actually being predicted, where the negative classes are never predicted.
The predictive performance of the models is different when the individual question types are used as target variables. Table 3.5 shows the permutation importance of each feature when predicting each question type using the support vector machine classifier including the model accuracy. The support vector machine classifier is used since this classifier shows some permutation importance for all features (the model with the CART classifier is only dependant on the workout frequency). The question types are visible in the columns of Table 3.5. Each column is a separately trained SVM classification model, where the target variable is the score of the question type. The values for each question type are split into 3 classes using the quantile-based discretization function, the same method that was used for the positive mental well-being classes. Next to the permutation importance, this table shows the accuracy per column. This accuracy is equal to the amount of predicted classes being true to the actual classes. The accuracy for each question type is always higher than when predicting the composite variable positive mental well-being. The permutation importance is higher for the daily steps, workout frequency and mean stress when predicting the positive affective state questions and for the workout frequency when predicting the questions about depression.

Including the PHQ and GAD scores

The model performance increases whenever the model is trained including the PHQ and GAD scores as features. The average accuracy for the prediction of 3 classes of mental well-being increased to ±57% whenever both scores are added. The importance of these two features can again be seen from the permutation feature importance (see Figure 3.7). For all models the PHQ and GAD scores show to be the most important features, if not the only. The classification and regression decision tree and the support vector machine algorithms showed that the PHQ and/or GAD values are the only values with any relevant importance to the models. The Pearson correlation test between the composite variable positive mental well-being (PMW) and the PHQ and GAD scores showed p-values < 0.001 with correlation coefficients of ± -0.50. Which shows that there is some overlap between the PHQ/GAD values and the mental well-being scores. Using only these two features when training each model also showed to reach similar average accuracies than when using these values in combination with the physical health data originating from the wearables.
The final specifications of each model (see Appendix.1) are chosen because they showed to reach the highest possible accuracy while also minimizing any signs of over- and under-fitting. Figure 3.8 shows the averaged accuracy of all models when the PHQ and GAD values are added against the number of mental well-being classes. It is visible that the accuracies are much higher than in Figure 3.5. Table 3.6 again shows the permutation importance of each feature per question type and the averaged accuracies per prediction of each question type. The values for the question types are again split using the quantile-based discretization function, and each column is again a different SVM classification model. It is visible that the accuracy does not drastically change when predicting the question types. What is noticable is that the permutation importance of the PHQ scores are mostly affecting the positive affective state and cognitive complaints questions. Next to that, the GAD score is mostly affecting the cognitive complaints questions.

Table 3.2: Confusion matrix for the CART classifier when 2 Mental Well-being classes are being predicted.
Negative: PMW score < 10.
Positive: PMW score >= 10.
PMW = positive mental well-being.

		Predicted	
	n=219	Negative	Positive
Actual	Negative	77	44
	Positive	34	64

Table 3.3: Confusion matrix for the CART classifier when 3 Mental Well-being classes are begin predicted.
Negative: PMW score < 4.
Neutral: 4 <= PMW score < 14 .
Positive: PMW score >= 14.
PMW = positive mental well-being.

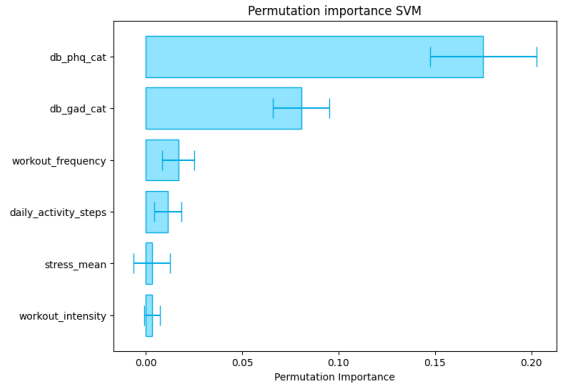
		Predicted		
	n=219	Negative	Neutral	Positive
Actual	Negative	0	55	28
	Neutral	0	37	30
	Positive	0	19	50

n=219		Predicted				
		VNeg	Neg	Neu	Pos	VPos
Actual	Very Negative	0	0	33	0	18
	Negative	0	0	28	0	15
	Neutral	0	0	21	0	16
	Positive	0	0	19	0	25
	Very Positive	0	0	10	0	34

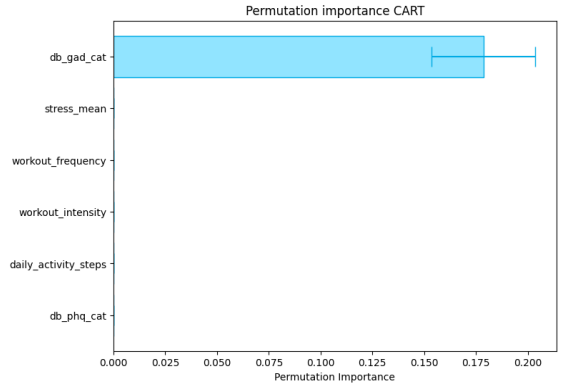
Table 3.4: Confusion matrix for the CART classifier when 5 Mental Well-being classes are begin predicted.
Very Negative(VNeg): PMW score < -1.
Negative(Neg): -1 <= PMW score < 6.
Neutral(Neu): 6 <= PMW score < 12 .
Positive(Pos): 12 <= PMW score < 17.
Very Positive(VPos): PMW score >= 17.
PMW = positive mental well-being.

	PAS	R	RES	COG	D	PMW
Daily Steps	0.020	0.0002	0.005	0.009	0.005	-0.0005
Work. Int.	0.007	0.0002	0.001	0.0003	0.001	0.009
Work. Freq.	0.049	0.0008	0.006	0.001	0.017	-0.019
Mean stress	-0.0004	0.007	-0.001	0.007	0.003	-0.005
Model performance						
Accuracy	0.45	0.44	0.45	0.39	0.47	0.38
Bal.Acc.	0.016	0.0	0.0	-0.001	-0.001	0.011

Table 3.5: The Permutation importance of the Support Vector Machine classifier for the prediction of 3 classes of each of the question types and the positive mental well-being.
Bal.Acc. is an abbreviation of balanced accuracy.
To make the values comparable, the predicted variables are again split into 3 classes using the quantile-based discretization function.
The question types are visible on the horizontal axis of the table and represent the positive affective state (PAS), rumination (R), resilience (RES), cognitive complaints (COG), and depression (D) (see Table 2.3).
The most right value is the composite variable, Positive Mental Well-being (PMW).



(a) Permutation importance of the features in the SVM model.



(b) Permutation importance of the features in the CART classification model.

Figure 3.7: Permutation importance plots for the prediction of 3 mental well-being classes using the CART classifier and the Support Vector Machine classifier including PHQ and GAD values.
db_phq_cat and db_gad_cat are the categorical values of the answers to the PHQ and GAD questionnaires (see Table 2.1 and Table 2.2).

	PAS	R	RES	COG	D	PMW
PHQ Score	0.108	1.0e-02	0.047	0.168	0.078	0.157
GAD Score	0.025	9.4-02	0.086	0.103	0.034	0.075
Daily Steps	0.005	7.4e-04	0.006	0.006	-0.002	0.005
Workout Int.	0.003	-2.2e-04	0.004	0.0003	0.006	0.003
Workout Freq.	0.006	8.5e-05	0.008	0.005	0.010	0.005
Mean stress	0.005	1.7e-03	0.002	-0.0002	0.004	0.003
Model performance						
Accuracy	0.58	0.53	0.56	0.58	0.55	0.57
Bal.Acc.	0.19	0.07	0.09	0.17	0.18	0.21

Table 3.6: The Permutation importance of the Support Vector Machine classifier including the PHQ and GAD scores for the prediction of 3 classes of each of the question types and the positive mental well-being.
Bal.Acc. is an abbreviation of balanced accuracy.
To make the values comparable, the predicted variables are again split into 3 classes using the quantile-based discretization function.
The question types are visible on the horizontal axis of the table and represent the positive affective state (PAS), rumination (R), resilience (RES), cognitive complaints (COG), and depression (D) (see Table 2.3).
The most right value is the composite variable, Positive Mental Well-being (PMW).

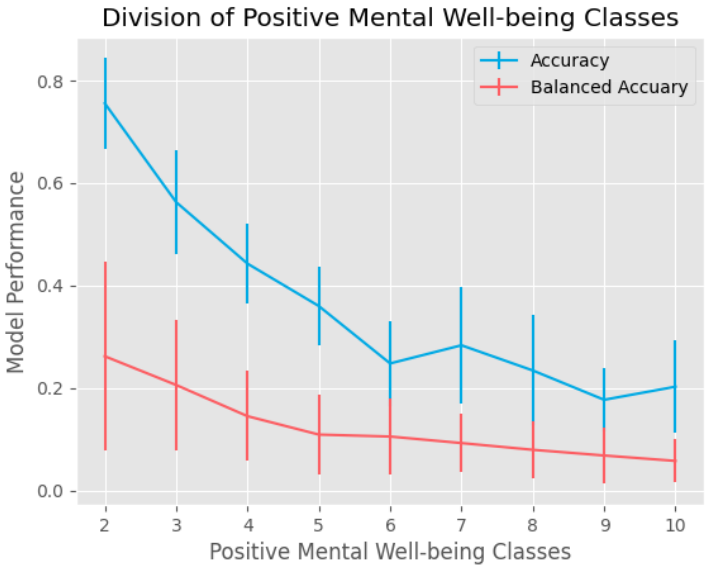


Figure 3.8: The accuracy and balanced accuracy for the CART classifier per amount of to-be predicted mental well-being classes. The included models are trained on the activity data from the wearables and the PHQ and GAD questionnaire scores.
The chance level for balanced accuracy is at a model performance of 0.

Discussion 4

The goal of this study was to use wearable technology in combination with mental state assessments to see if any correlation could be found and if this correlation could be used for a predictive model. Two research questions were formulated:

- *Can we use physical health data originating from consumer wearable technology to note a correlation between mental and physical health?*
- *How can the data from these wearables be used to create a model which is capable of predicting the mental state of an individual according to physical activity information?*

According to earlier studies based on self-reflection questionnaires or medical grade devices, a correlation should be found between physical data such as sleep, workout and heart information. These previous studies also showed the importance of the current mental state of a person, which we record using the PHQ and GAD questionnaires. We hypothesized that more intense and more frequent activity, longer sleep durations and a higher percentage in NREM3 and REM stages, and lower stress levels would correlate with an improved mental well-being.
The results partially confirm the hypotheses. The participants reported a more positive mental well-being whenever they were more active and their stress scores were lower. The amount of daily steps, the workout intensity and the mean stress scores showed to have a significant effect, while none of the sleep variables showed to have a significant effect on the mental well-being. The marginal and conditional R^2 values showed to differ a lot, which suggests that the random effects (participant and date) could explain more variance to the model. This shows that the physical health data originating from the wearables is not the only factor in the correlation to mental state. The date could have an effect because of different weather on each date, as mentioned in the introduction [41]. This is also visible in that more variance to the model could be explained whenever the PHQ and GAD scores were included. Ideally this data should not have been included to the model since we want to see if we can only use physical health data to state the correlation with mental well-being. But it shows that prior knowledge about mental health helps with explaining the correlation. From all frequency/periodical data only workout frequency has shown to have a significant effect on mental well-being . All these different data types will be explained more in detail. After this the predictive model will be discussed which did not show to be able to predict mental well-being with a high accuracy whenever more mental well-being classes are categorized. Even when only 2 classes could be predicted, the highest accuracy was 75%.

4.1 Review of the different data types

Daily Activity Steps

Previous literature showed a correlation between daily steps, daily activity intensity and mental health [76][77]. In this study the variable "Daily activity intensity" is used, in which only the amount of steps are actually used. This decision was made because the amount of steps taken, distance traveled and calories burned showed to be highly correlated.
To get a better estimation of the activity intensity a variable called VO2-Max could be used. The VO2-Max value represents the maximum amount of oxygen that your body is capable of using while performing an activity and this variable has shown to be correlated with mental health [122]. Recently wearable brands started to offer a VO2-Max value over a day, unfortunately this was not yet available in the WeFitter platform during the experimental period.
Since the daily information was only recorded per date, we did not have access to more specific data concerning daily activity. This means information like sedentary time and specific minute per minute data could not be used, while both showed to be correlated with mental health [123]. Because of this unavailability only step information (estimated on different timestamps over a day) were used. The amount of steps taken before submitting the forms was estimated based on the average sleep wake times of previously available WeFitter data. This estimation does not accurately represent the division of activity of a participant over a day, it could be the case that the participant is only highly active for one hour in the day, while spending the rest of the day sitting down. It could be argued that because of this estimation, the significant effect found for the daily activity steps is unreliable. By estimating the amount of steps taken based on previously available data, this value is heavily affected by the time of submission of the questionnaire. However, neither the time of submission nor the fact that the questionnaire was submitted in the morning or the afternoon showed to have a significant effect on mental well-being. This shows that the found significant effect is not only caused by the submission time of the questionnaire but actually also by the total amount of daily steps taken.
Previous research also showed an effect of daily activity frequency on mental health issues [78]. This data did not show to have a significant effect on mental well-being in this study. One possible explanation for this is that the previous research mostly saw the correlation between daily activity frequency and state of mental health with elderly individuals (65+ Years old). This study did not have a lot of older participants, and the oldest participant was 62 years old.

Workouts

Previous research showed that workout intensity, duration and frequency showed to have a significant effect on mental well-being. In this study the workout intensity was calculated using the workout duration and workout type. The linear mixed effects model confirmed that this variable has a significant positive effect on mental well-being. Figure 3.1b shows the effect of the workout intensity on mental well-being. This plot shows

that there are not much workouts on the higher spectrum of intensity. The skew in distribution is also exaggerated by the equation used for the workout intensity. This equation was used since the duration showed to have a significant effect on mental health in previous studies, and a longer duration of strength training is more aerobic intensive than a longer duration of meditation. This problem again could be solved by using the VO2-Max value during the workout, but this parameter was also not available during the experimental period in the WeFitter platform.
Most workouts are around an intensity of 13672, this is comparable to a bike ride of 45 minutes (using Equation 2.1). Only 52 of the 727 questionnaires were accompanied with a workout intensity score higher than 0. This means that a lot of the participants in this experiment did not perform or record their workout sessions. Since the previous research found a correlation between high intensity workouts and mental health, our data might be too heavily skewed to reliably annotate a correlation.
Two of the participants informed me that they did not wear their wearable every time they performed a workout. The first participant thought that the wearable was annoying to wear during boxing classes and the second did not completely trust the water resistance of their wearable for a swimming session. Both of these workouts are high intensity workouts, so this data is important to show the correlation between high intensity workouts and mental well-being.
Most other participants informed me that they did wear the wearable during workouts, although they thought it was uncomfortable to do so.

Sleep

Although previous research showed a significant effect of sleep duration and percentage of time in NREM3 and REM sleep, this study did not confirm these correlations when using consumer wearable devices. There are two important shortcomings that could be the reason for this lack of significance.
The first possible cause of the lack of significance is that only 316 of the total 727 amount of questionnaires were provided with sleep data. The data imputation should have been able to fix this issue, but because 57% of the data had to be imputated, the actual correlation estimation is done over mostly the same values, where the actual real values became the minority. This problem is even worse for the sleep phase data, only 185 different questionnaires were provided with sleep phase data (75% had to be imputated).
The second possible cause is that the sleep recordings that were collected did not really vary that much from each other. Our mean sleep duration is just below 8 hours, and 60% of our data set is between 7 and 9 hours of sleep. The data set also includes sleep recording for only some couple of hours, this data is questionable because the possibility of someone actually having only 1 hour of sleep is low (none of the participants suffer from insomnia). Excluding this data did not improve any significance levels. The previous research showed the correlation between sleep and mental health for people who sleep way too much of way too little.

Stress

Although the stress levels were only recorded for a small amount of individuals (8 participants). The mean stress score still showed to have a significant effect on mental well-being. Because of the low number of participants this data is not completely reliable.
Figure 3.1c clearly shows the imputated data points. 87% of the questionnaires had to be imputated with stress data. This again means that the actual data becomes the minority data in the set. Although the correlation between lower and higher than average stress values could still be mentioned, the specificity from the data is now much lower.
From some participants I got feedback that they thought that the stress score was almost never accurate to how they actually claimed to feel. The Garmin watch notified them about high stress levels while the participants felt really relaxed. This problem was also found in the low correlation scores between the mean stress score and the answers to the stress question in the mental well-being assessments. The Pearson correlation test shows a p-value lower than 0.05 for a correlation value of 0.125, this indicates that there is indeed a correlation, but this correlation is low. The source of this problem can be many different things, wearing the watch too loose, wearing the watch only during workouts, a bad skill of self-reflection or just a low level of specificity in the watch.
The fact that the stress score now only comes from Garmin variables does not help with the generelizability of the correlation. It is possible that the method which Garmin uses to calculate the stress score from the heart rate variability values is different than how a different brand would do it. During the experimental period the raw HRV data was not yet available in the WeFitter platform, this data would be more reliable to use since the equation that Garmin is using to calculate the stress score is not publicly available.

The Importance of Prior Mental Health Data

The model were the PHQ and GAD scores were included showed to explain more variance to the model. Figure 3.2 shows that the beneficial effects of more steps, higher workout intensities or longer sleep durations are significantly different between individuals with high and low GAD scores. For the daily steps and workout intensity, it shows that the mental well-being of an individual who suffers from a severe anxiety disorder is stronger affected by these two parameters than individuals who do not have this disorder. For sleep duration, individuals with severe anxiety feel worse when sleep is longer and individuals without severe anxiety feel better when sleep is longer. This is interesting since next to being less active [124][125], individuals who suffer from anxiety typically show signs of insufficient sleep [34]. The fact that individuals with extreme anxiety show an improved state of mental well-being when sleep is shorter defies expectations based on earlier literature. A high score of mental well-being is not expected to be correlated with high anxiety scores (also since the PHQ and GAD values showed to be negatively correlated with the mental well-being scores). The small sample size of individuals with signs of severe anxiety (3 participants) explains the need for more research on this topic.
The Pearson correlation test showed no significant correlation between

high GAD scores and shorter sleep sessions. This might indicate that the beneficial factor of these physical health parameters on mental well-being is truly different between individuals with and without anxiety. Especially with the sleep duration this shows that prior knowledge about mental health is important to correctly note a correlation between physical health information originating from consumer wearables and the mental well-being of an individual.

Effects on Question Types

The correlation effects between each separate question type and the wearable data showed some interesting results. For the linear mixed effects model where positive affective state was the to-be predicted value, the GAD and PHQ scores did not seem to have a significant effect. The sleep depth did however show to have a significant negative effect. The negativity of this effect seems questionable since earlier research showed that an increase of time in deeper sleep phases showed to be beneficial for mental health [51][55]. One possible explanation might be sleep disruptions. Studies found that sleep disruption might be worse for mood than problematic sleep duration [126]. Especially waking from deeper sleep might results into worse mood, so if someone woke up to an alarm while in deep sleep, this chance of a worsened mood is higher. Unfortunately this measurement was not included in this experiment, which might be doable with access to the alarm of the participant or ask a question about how they woke up.
The significant effects for rumination and resilience are as expected. A higher GAD score results into higher scores on the rumination questions, and a higher stress score results into lower resilience scores (see Section 1.3).
The questions about cognitive complaints are the only questions where both the GAD and the PHQ scores have a significant effect. It was expected that both scores would have a significant effect on all question types because all question types are a possible indication of depression or anxiety. That the cognitive complaints scores are the only scores significantly affected by both PHQ and GAD values might indicate that participants who score higher on both these tests are more prone to problems with cognition. Since one of the questions about cognitive complaints was about physical discomfort, it could've been expected that workout intensity and so muscle ache would have a significant effect but this was not the case. An explanation for this might be that workout intensity does not directly correlate with muscle ache, that the muscle ache was never bad enough or that the question is only one third of the score of the question type, so it does not influence it that massively. Whenever the workout frequency is higher, the cognitive complaints showed to decrease. As mentioned in Section 1.3, this complies with what earlier studies showed.
The depression questions showed to be significantly negatively affected by the daily steps, the workout intensity and the workout frequency. This is as expected according to earlier studies (see Section 1.3).

4.2 Predictive model

Eight different classification methods were used and compared, and for each method the used features were the same. It was decided to only use the variables that showed to have a significant effect on the mental well-being in the linear mixed effects model. This decision was made because including the other data types decreased the model accuracy and increased the amount of over-fitting.

Classifiers

The different classification methods showed to be similar in classification performance while some showed to be able to reach a better fitting model than others. Appendix.3 shows the accuracy of the learning curves for all used methods. It is visible that Naive Bayes, Random Forests, K-neighbours and ADA Boost learning still show (slight) signs of over-fitting. To decrease over-fitting, the model specifications are set so that the classification threshold is more linear. This is best seen in the amount of neighbours in the K-neighbours classifier (120 neighbours are used for classification). This again suggests that the specificity of the prediction models decreased.
The logic of ensemble learning is that weak aspects of one classifier can be combined with strong aspects from another classifier. Although the model performance of the ADA boost algorithm was close to the highest performing model of the CART classifier (see Figure 3.4), it showed signs of over-fitting. The random forests algorithm showed the lowest mean accuracy scores and also showed signs of over-fitting. The balanced accuracy of all models showed to stay close to a model performance of 0, which suggests that none of these models have any 'strong' aspects. This could be the reason why the ensemble models do not show any increased model performances. Other methods such as training a neural network were explored, but this did not show to improve model performance.

Mental Well-Being Classes

The final models were optimized for 3 classes of mental well-being, because this showed to be the most promising one (2 classes would not have been useful, and more showed to decrease accuracy). This means that all parameter and model values are ideal for the 3 classes, while it could be not ideal for more classes. This decision was made because otherwise the different amount of mental well-being classes would not have been comparable. when trying to optimize the settings for more well-being classes, the accuracy did not improve.
The divisions of the classes are currently done using pandas' quantile-based discretization function. This is optimal for the training of the model since it groups the data set in similar amount of buckets, but this might not be ideal for the segregation of the classes. Currently in the 3 mental well-being classification, everything with a score between 4 and 14 is classified as a neutral class of mental well-being. Since the score is a composite variable of the used questionnaires, we cannot say that this is comparable to an actual 'neutral' class of mental well-being. Only

when this is compared to other participants in this study, this would be classified as neutral.

Model performance

The final results of the predictive model only showed promising scores when the GAD and PHQ scores were added and when we wanted to classify mental well-being on either a positive or a negative score (accuracy of ±75%). When we add one more class the accuracy already drops to ±57%. When the GAD and PHQ scores were not used during training the model performance for 2 classes was ±58% and for 3 classes this already dropped to ±38%. This 38% is just a bit higher than the guessing chance for three classes. This indicates that the physical information features are not really useful for the model, and that the PHQ and GAD questionnaires are necessary to have a higher accuracy for predicting any ones mental state.
Classifying mental well-being on only two different classes is not really useful. This could mean that someone who is just not having an ideal day is classified in the same group as someone who is experiencing a really bad day. This also means that the prediction of 3 separate classes would not be really useful. Ideally the model would be able to make at least a prediction to 5 or more classes, strongly negative, slightly negative, neutral, slightly positive, and strongly positive. Table 3.4 shows that a predictive model for 5 classes is not working correctly. The predictive class is mostly the neutral class. This selective classification is also visible when more classes are added (see the confusion matrix of values when 10 classes are predicted with the SVM classifier in Appendix Table.2). For a 10 class classification it is clear that there are only 4 different classes which are most often predicted. A reason for this could be that previous research was mostly performed between mentally healthy participants and participants who are suffering from a mental disorder. It could mean that the effects visible from activity, sleep and stress data are mostly effecting individuals who would be on the more extreme side of mental health. According to the answers to the PHQ and GAD questionnaires this would mean that most predictions would go to the neutral class, since we do not have a lot of participants in the high classes of the GAD questionnaire, and none in the high classes of the PHQ questionnaire. This is not the case so this could mean that the model is wrongly predicting a lot of individuals to be in the extremely negative mental well-being class.

Permutation Importance

The permutation importance graphs and tables show that the importance of the features originating from wearable devices (so not the PHQ and GAD scores) are just a small fraction of the total model score. This means that the model performance is quite independent from the used features. The standard deviation for each value shows to be a big value compared to the importance. The permutation importance of the PHQ and GAD scores showed to be a larger fraction of the total model scores. This indicated that the general mental well-being variable could be predicted using the PHQ and GAD values only. Doing so resulted into similar

accuracy scores. This changes when the predictive variable was only one of the question types.

Predicting Individual Question types

The prediction of the individual question types seemed to be promising when looking at the difference it had on the significance in the linear mixed effects models. For the models that did not include the PHQ and GAD values as features, the prediction accuracy of the individual question types outperformed the prediction accuracy of the composite variable mental well-being (see Table 3.5). This could mean that this mental well-being score is not optimally composed. Some tests were done where some of the question types where excluded from this composition but this did not show to improve model accuracy. The permutation importance of the features did show to increase in some cases when predicting the individual question types. Although the permutation importance of daily steps, workout frequency and stress showed to increase when predicting positive affective state, cognitive complaints or depression, the importance is still very low. This shows that the model accuracy is not really dependant on the used features.
When the PHQ and GAD values were added the difference in predicting mental well-being or predicting the individual question types did not show any big differences (see Table 3.6). This again confirms that the model performance is not dependant on the features originating from the wearables.

4.3 Limitations

The main limitation of this study was to get people invested enough to participate correctly and to wear their wearable as often as possible. Most people do not wear their wearable the whole day nor do they always wear it during sleep. It was already hard to find participants who would want to join the study even though they would be compensated with information that could coach them on how to live a healthier life. Previous research showed that all used data types could be important to note some correlation between mental and physical health so if a person does not wear their wearable during sleep or does not fill in all questionnaires, this can already compromise their individual results.
The willingness to participate also differed a lot. Some participants mentioned at the end of their experimental period that they would like to continue doing the questionnaires. They liked to have a short reflective questionnaire at the end of the day. This method of self-reflection at the end of the day could be used as a antidepressant, although it could also lead to more rumination, which in turn could function as an depressant [127][34]. The majority of the participants were however mostly annoyed by the questionnaires, and did not like the frequency of them. Presenting the questionnaires less often to the participants or changing the questionnaires could help with engagement. The expectation of an actual coaching platform with life coaches is that all participants want to be helped. This might also result into better participation in their prognosis trajectory and thus might result into

more reliable data.
The addition of the location question in the daily questionnaires was done since the location of a person could heavily influence the mental state of a person. Because of the low number of participants this question was not added to the models. The choice of using different wearable brands was made with the intention to reach more participants. Although it helped with the amount of participants, it clustered the small set of participants across the different brands. The results for not constant available data such as sleep and stress information are suffering from this limitation. Since not all participants have a wearable that records stress or sleep (phase) data, a lot of important data went missing.
To still use all the daily activity and workout information a lot of data for the sleep and stress scores was imputated. Without imputation the used data set would be small. With this imputated data a correlation could still be found between low and higher values in the data type, but the specificity of this data type decreased. This is mostly affecting the accuracy of the predictive models. This was visible in that the model showed to work optimally when a more linear decision boundary was used.
The WeFitter platform was missing some important data types during the experimental period. Currently WeFitter added the functionality of intraday step, calories, distance, and heart rate information. This information shows a better picture on how active a person is over a day than just summarized information. Another addition to the platform is more clinical biometric data such as blood pressure, body fat, glucose levels, temperature, blood oxygen levels, VO2-max levels, and more compatibility for heart-rate variability recordings. The WeFitter platform is now also capable of recording individual sleep phase start and end times. This is more informative about whether the participant has an adequate ultradian rhythm since a mentally healthy individual enters REM sleep on average 4 or 5 times a night with ±90 minute intervals [128].
Even though this study was introduced by improving on existing studies since wearables are less intrusive to wear than medical grade devices, a lot of participants informed me that they did not wear their wearable when they are more relaxed. Some participants only wore their wearable whenever they went out of the house or whenever they started a workout and took them off them whenever they came home or were done with the workout. This means that a lot of 'normal life' data is not recorded.
The current battery life of the wearables is another mayor limitation. Most wearables are still only capable of functioning for 1 or 2 days. This means that the wearable should be charged almost every day, which is something that could be forgotten. Some participants notified me about an empty battery in their wearable before the middle of the day (this happened on 2 occasions). This again means that unfortunately a lot of important information is not recorded.
The last limitation of this study is that it took place during the COVID-19 pandemic. 1 participant notified me about having COVID in the last 2 days of the experimental period and 2 participants received the vaccine during the study. This made them feel less positive, less active, and sometimes also really sick. In two cases this was only a day or two, but one participant was sick for 10 days in which he did not leave his bed or couch. This participant was excluded from the study since this does not

show normal activity, for the other two participants the days that they notified to be sick were excluded form the experiment.
As mentioned in the section about stress, the raw HRV values could not be used in this study, only the related stress scores are used. The scientific background of the correlation of mental health and stress was based on raw HRV values. Garmin is stating that their stress scores are based on HRV values, but do not disclose the equations used to get these stress scores. Although earlier research also used the Garmin stress scores instead of raw HRV values [129][130][131], this unknown factor can be problematic.

4.4 Future work

Since one of the biggest limitations of this study was the lack of hundreds or thousands of participants, a possible follow up study might be to reproduce this experiment with much more participants. If this follow up study could also make use of the new data types such as intraday data, advanced biometric measurements, proper heart rate variability recordings, more detailed sleep recordings, and a wider variety of wearable brands that can be connected, the results would be more informative.
The experiment design and goal should be revised. The current setup to have a questionnaire twice a day made the participants feel annoyed and less engaged.
Another variant of this study could be implemented by looking at cognitive abilities by for example letting the participant play a small game on their mobile phone. The ease of getting data through the wearables (through the WeFitter platform) makes it possible to apply this data to a wide variety of studies. This research has shown an indication that wearables could function as a easy source of physical health tracking.
A way of improving engagement with the study should be explored. Maybe if the participant was presented with their performances during the experimental period this would help with engagement. This could however also influence the data massively, so this could only reliably be done over a longer time period. Then it could also be tested what effect certain feedback has on the physical and mental health of an individual and how this feedback is ideally presented to the participant. This longer time period might also be informative to look at. Mental disorders do not just develop over a time span of 2 weeks. This can take multiple months to properly assess after which multiple months are necessary for rehabilitation [132]. It might therefore be much more interesting to follow participants in a long term study with periodic (professional) checkups. Eventually the goal of trying to use wearables to capture early signs of a deteriorating mental state might be more useful than being able to show signs of their current mental state.

Conclusion 5

The aim of this study was to see if wearables could be used to find a correlation between mental health and physical health, and if this correlation could effectively be used to create a model capable of predicting the mental state of a person. Although a correlation was found in some data types, not all expected data types showed to have a significant effect and not all data types could be interpreted to have reliable results.
A higher amount of daily steps, a higher workout intensity and frequency and a lower mean stress score showed to significantly correlate with an improved mental well-being. The correlation of the mean stress score on mental well-being was based on only 8 different participants, which makes the correlation of this data type less reliable. The different sleep data types did not show to have a significant effect on mental well-being, this could be because of the low number of participants or because of a low accuracy of measurements in the wearable devices.
The maximum accuracy of the most optimal predictive model showed to be 75% for only predicting either a positive or a negative class of mental well-being. To reach this accuracy the scores of the PHQ and GAD questionnaires had to be added. The scope of this study was to see if physical health data originating from wearable devices could be used to predict the mental state. The inclusion of the PHQ and GAD scores means that prior information about mental health is necessary to make a prediction about the current mental state of an individual.
This study showed to be more than a progressive observational study about the correlation between physical health data originating from consumer wearable devices and mental health data. It also showed to be a feasibility study on using wearable devices in such a short diary study with this amount of participants, these mental well-being assessments and the use of multiple different wearable brands.
The correlation found between mental state and the daily steps, workout data and heart rate information shows that the use of consumer wearables in similar studies is promising. Big intrusive medical devices or long questionnaires show to become less necessary to get simple physical health information from an individual. The absence of the significant correlation between sleep data and mental well-being shows that more research is necessary, where more participants but also more detailed data types should be used. The low prediction performance shows that the current setup of using physical health information from wearables to predict the current mental well-being of an individual is not a reliable/usable method.

Appendix

Clustering of Positive Mental Well-Being Assessments

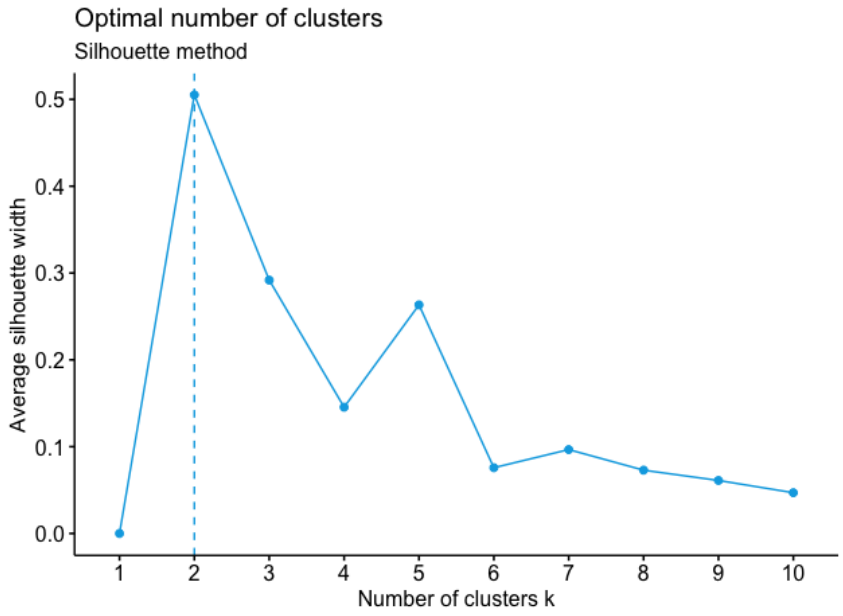


Figure .1: Visualization of the results of K-Means clustering using the silhouette method.

Table .1: Results of the Factor analysis on the mental well-being submissions

Factor Analysis using method = pa					
Call: fa(r = dat, nfactors = 2, rotate = "varimax", max.iter = 100, fm = "pa")					
Standardized loadings (pattern matrix) based upon correlation matrix					
	PA1	PA2	h2	u2	com
ds_energetic	0.70	-0.31	0.59	0.41	1.4
ds_happy	0.81	-0.30	0.74	0.26	1.3
ds_satisfied	0.79	-0.26	0.70	0.30	1.2
ds_focused_on_feelings	0.06	0.42	0.18	0.82	1.0
ds_worried_problems	-0.39	0.58	0.48	0.52	1.8
ds_cope_with_negative_life_event	0.72	-0.21	0.56	0.44	1.2
ds_forgetful	-0.23	0.55	0.36	0.64	1.3
ds_inattentive	-0.31	0.60	0.45	0.55	1.5
ds_psysical_discomfort	-0.29	0.46	0.29	0.71	1.7
ds_depressed	-0.40	0.52	0.44	0.56	1.9
ds_hopeful	0.80	-0.15	0.66	0.34	1.1
ds_stressed	-0.25	0.52	0.34	0.66	1.5

Correlation matrix

Table .2: Correlation of all physical health data types used.

	daily_activity_intensity	workout_intensity	workout_since_last
daily_activity_intensity	1.000	-0.152	0.056
workout_intensity	-0.152	1.000	-0.492
workout_since_last	0.056	-0.492	1.000
sleep_duration	0.0341	0.027	-0.075
sleep_depth	0.021	0.035	-0.072
stress_mean	-0.011	-0.016	0.055

	sleep_duration	sleep_depth	stress_mean
daily_activity_intensity	0.034	0.021	-0.011
workout_intensity	0.027	0.035	-0.016
workout_since_last	-0.075	-0.072	0.055
sleep_duration	1.000	0.033	-0.066
sleep_depth	0.033	1.000	-0.040
stress_mean	-0.066	-0.040	1.000

Predictive model

Code 1 Classification model specifications

```
"SVM": SVC(
    C=0.3,
    kernel='rbf',
    gamma='auto',
),
"Naive Bayes": GaussianNB(),
"Logistic": LogisticRegression(
    C=0.55
),
"RandomForest": RandomForestClassifier(
    n_estimators=5,
    max_depth=2,
    ),
'LDA': LinearDiscriminantAnalysis(),
'KNN': KNeighborsClassifier(
    n_neighbors = 120,
    leaf_size = 10,
    p=1,
    n_jobs=-1
),
'CART': DecisionTreeClassifier(
    min_weight_fraction_leaf = 0.4
),
'ADA': AdaBoostClassifier(
    n_estimators = 100,
    learning_rate = 0.001
)
```

Linear Mixed Effects

Daily information

Table .3: Summary of a linear mixed effects model of the effects of the daily activity, workout, sleep and stress data. The daily activity, workout intensity and mean stress score have a p-value < 0.05

Model: mental well-being ~ daily_activity_steps + workout_intensity + sleep_duration+ sleep_depth + stress_mean+ (1 | participant) + (1 | date)						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	1.280e+01	3.502e+00	4.898e+02	3.656	0.000284	***
daily_activity_steps	3.517e-04	9.783e-05	6.914e+02	3.595	0.000347	***
workout_intensity	1.919e-05	7.663e-06	6.912e+02	2.504	0.012518	*
sleep_duration	4.701e-04	4.136e-03	7.079e+02	0.114	0.909533	
sleep_depth	-1.093e-01	6.116e-02	6.949e+02	-1.788	0.074281	.
stress_mean	-1.230e-01	5.535e-02	6.948e+02	-2.223	0.026566	*

* p < .05 ** p < .01 *** p < .001

Table .4: Summary of a linear mixed effects model of the effects of the daily activity, workout, sleep and stress data. This model included the answers of the PHQ and GAD questionnaires. The PHQ value, mean stress score, and the interactions between GAD score and daily stpes, workout intensity and sleep duration have a p-value < 0.05

Model: mental well-being ~ (daily_activity_steps + workout_intensity + sleep_duration * phq) * gad + sleep_depth + stress_mean + (1 | participant) + (1 | date)						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	2.264e+01	7.283e+00	6.152e+02	3.109	0.00196	**
daily_activity_steps	5.812e-05	1.367e-04	6.870e+02	0.425	0.67093	
workout_intensity	5.900e-06	8.697e-06	6.845e+02	0.678	0.49779	
sleep_duration	1.757e-03	1.333e-02	6.908e+02	0.132	0.89515	
phq	-2.244e+00	9.531e-01	3.613e+02	-2.354	0.01909	*
gad	7.908e-01	1.329e+00	3.934e+02	0.595	0.55223	
sleep_depth	-1.098e-01	5.968e-02	6.993e+02	-1.840	0.06619	.
stress_mean	-1.199e-01	5.416e-02	6.932e+02	-2.214	0.02717	*
sleep_duration:phq	2.806e-03	1.746e-03	7.035e+02	1.607	0.10850	
daily_activity_steps:gad	5.060e-05	1.614e-05	6.931e+02	3.135	0.00179	**
workout_intensity:gad	3.499e-06	1.143e-06	6.862e+02	3.060	0.00230	**
sleep_duration:gad	-4.967e-03	2.440e-03	6.921e+02	-2.035	0.04221	*
phq:gad	4.893e-02	1.444e-01	4.779e+02	0.339	0.73488	
sleep_duration:phq:gad	5.130e-05	2.717e-04	6.966e+02	0.189	0.85031	

* p < .05 ** p < .01 *** p < .001

Frequency data

Table .5: Summary of a linear mixed effects model of the effects of the frequency information. This model includes the difference in daily activity, the workout frequency, the sleep start frequency and sleep duration frequency over windows of 3 days. The workout frequency has a p-value < 0.001.

Model: mental well-being ~ (daily_activity_frequency + sleep_start_frequency * phq) * gad + workout_frequency + sleep_duration_frequency +(1 | participant) + (1 | date)						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	2.531e+03	2.284e+03	6.792e+02	1.108	0.2681	
daily_activity_frequency	3.755e-04	2.978e-04	7.032e+02	1.261	0.2077	
sleep_start_frequency	-1.069e+02	9.710e+01	6.792e+02	-1.101	0.2714	
phq	-2.748e+02	2.495e+02	6.792e+02	-1.101	0.2711	
gad	-1.861e+02	1.664e+02	6.792e+02	-1.118	0.2638	
workout_frequency	1.977e+00	4.100e-01	6.977e+02	4.822	1.75e-06	***
sleep_duration_frequency	-2.836e-02	3.138e-02	6.889e+02	-0.904	0.3664	
sleep_start_frequency:phq	1.164e+01	1.061e+01	6.792e+02	1.098	0.2728	
daily_activity_frequency:gad	-6.034e-05	3.652e-05	7.030e+02	-1.652	0.0989	.
sleep_start_frequency:gad	7.840e+00	7.076e+00	6.792e+02	1.108	0.2683	
phq:gad	1.958e+01	1.765e+01	6.792e+02	1.110	0.2675	
sleep_start_frequency:phq:gad	-8.283e-01	7.504e-01	6.792e+02	-1.104	0.2701	

* p < .05 ** p < .01 *** p < .001

Confusion matrix When Predicting 10 Mental Well-being Classes

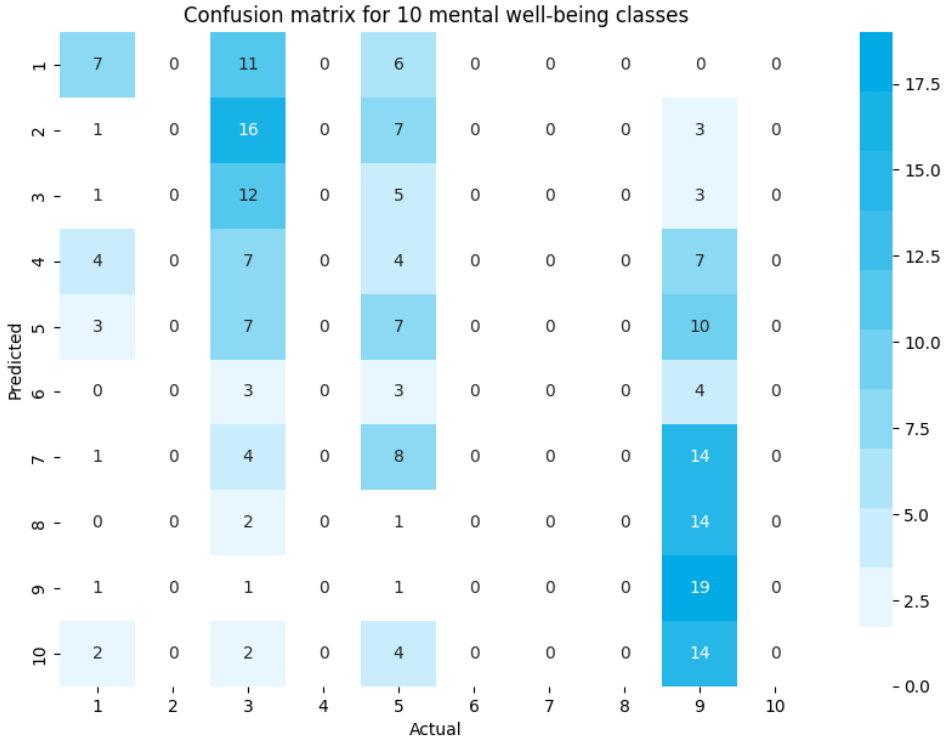
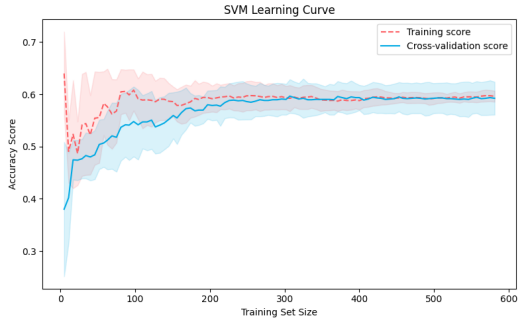


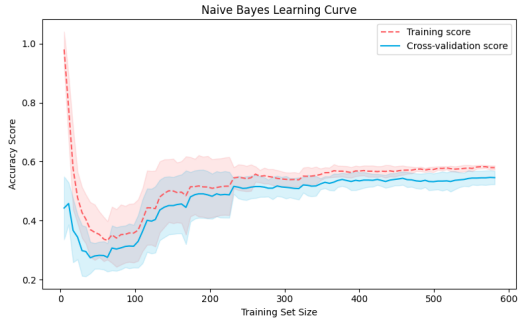
Figure .2: Confusion matrix of the test cases when 10 mental well-being classes are being predicted with SVM classifier including the PHQ and GAD values.

Prediction Model Performances

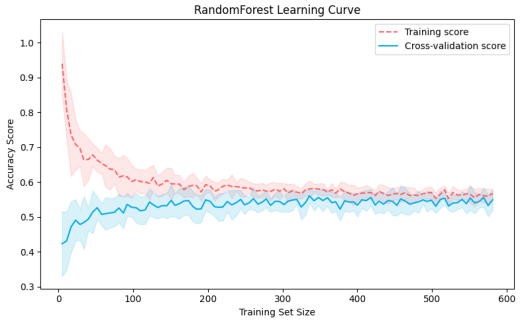


(a) Learning curve of the support vector machine classifier with 3 mental well-being classes.

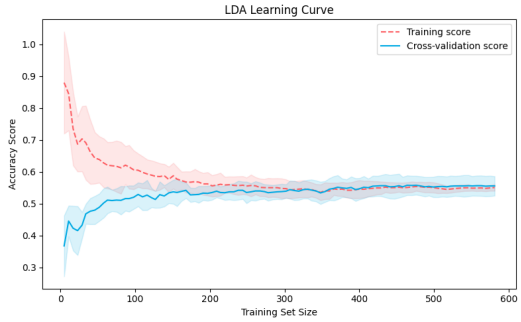
(b) Learning curve of the logistic regression classifier with 3 mental well-being classes.



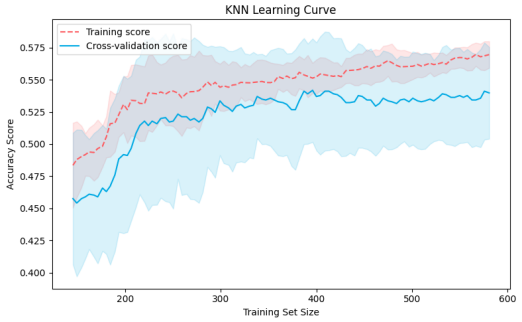
(c) Learning curve of the naive Bayes classifier with 3 mental well-being classes.



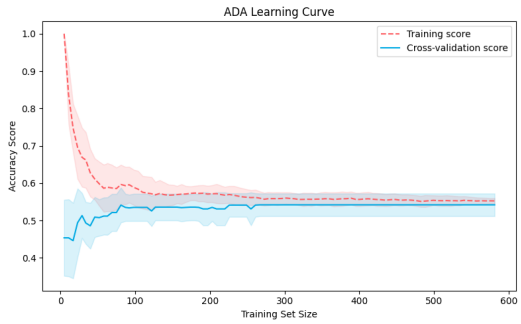
(d) Learning curve of the random forests classifier with 3 mental well-being classes.



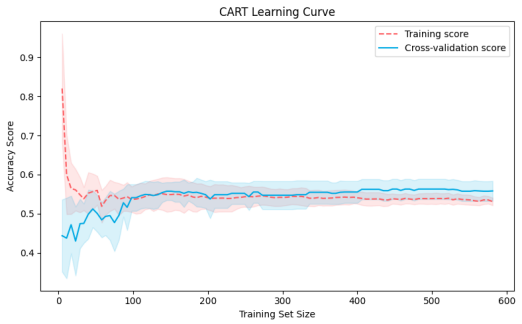
(e) Learning curve of the Linear Discriminant Analysis classifier with 3 mental well-being classes.



(f) Learning curve of the K-Neighbors classifier with 3 mental well-being classes.



(g) Learning curve of the ADA Boost classifier with 3 mental well-being classes.



(h) Learning curve of the classification and regression decision tree classifier with 3 mental well-being classes.

Figure .3: Learning curves of all trained models including the PHQ and GAD scores. The blue values represent the cross-validation score and the red values the training scores. These plots do not include the balanced accuracy scores because of visualization clarity. The KNN learning curve (Figure .3f) has a different start of the X-axis because 120 neighbours are used for the classification method.

Bibliography

The references in citation order:

[1] Lukasz Piwek et al. 'The rise of consumer health wearables: promises and barriers'. In: *PLoS medicine* 13.2 (2016), e1001953 (cited on pages 1, 6).

[2] Hugh Hunkin, Daniel L King, and Ian T Zajac. 'Perceived acceptability of wearable devices for the treatment of mental health problems'. In: *Journal of Clinical Psychology* 76.6 (2020), pp. 987–1003 (cited on pages 1, 6).

[3] Benjamin W Nelson and Nicholas B Allen. 'Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study'. In: *JMIR mHealth and uHealth* 7.3 (2019), e10828 (cited on pages 1, 6).

[4] George E Vaillant. 'Natural history of male psychologic health: Effects of mental health on physical health'. In: *New England Journal of Medicine* 301.23 (1979), pp. 1249–1254 (cited on page 1).

[5] Michael Phelan, Linda Stradins, and Sue Morrison. *Physical health of people with severe mental illness: can be improved if primary care and mental health professionals pay attention to it*. 2001 (cited on page 1).

[6] Monika Guszkowska. 'Effects of exercise on anxiety, depression and mood'. In: *Psychiatria polska* 38.4 (2004), pp. 611–620 (cited on page 1).

[7] Kavitha Kolappa, David C Henderson, and Sandeep P Kishore. *No physical health without mental health: lessons unlearned?* 2013 (cited on page 1).

[8] Julius Ohrnberger, Eleonora Fichera, and Matt Sutton. 'The relationship between physical and mental health: A mediation analysis'. In: *Social science & medicine* 195 (2017), pp. 42–49 (cited on page 1).

[9] Steven Marwaha et al. 'Mood instability and psychosis: analyses of British national survey data'. In: *Schizophrenia bulletin* 40.2 (2014), pp. 269–277 (cited on page 1).

[10] June Gruber et al. 'Happiness is best kept stable: positive emotion variability is associated with poorer psychological health.' In: *Emotion* 13.1 (2013), p. 1 (cited on page 1).

[11] Nicole CM Korten et al. 'Early and late onset depression in young and middle aged adults: differential symptomatology, characteristics and risk factors?' In: *Journal of affective disorders* 138.3 (2012), pp. 259–267 (cited on page 1).

[12] Timothy A Brown, Bruce F Chorpita, and David H Barlow. 'Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal.' In: *Journal of abnormal psychology* 107.2 (1998), p. 179 (cited on page 1).

[13] Jules Angst. 'The bipolar spectrum'. In: *The British Journal of Psychiatry* 190.3 (2007), pp. 189–191 (cited on page 1).

[14] Jules Angst and Giovanni Cassano. 'The mood spectrum: improving the diagnosis of bipolar disorder'. In: *Bipolar disorders* 7 (2005), pp. 4–12 (cited on page 1).

[15] Alissa Knight and Niranjan Bidargaddi. 'Commonly available activity tracker apps and wearables as a mental health outcome indicator: a prospective observational cohort study among young adults with psychological distress'. In: *Journal of Affective Disorders* 236 (2018), pp. 31–36 (cited on pages 1, 6).

[16] Kathleen Ries Merikangas et al. 'Real-time mobile monitoring of the dynamic associations among motor activity, energy, mood, and sleep in adults with bipolar disorder'. In: *JAMA psychiatry* 76.2 (2019), pp. 190–198 (cited on pages 1, 6).

[17] Katja Siefken, Astrid Junge, and Lena Laemmle. 'How does sport affect mental health? An investigation into the relationship of leisure-time physical activity with depression and anxiety'. In: *Human Movement* 20.1 (2019), pp. 62–74 (cited on pages 1–3, 6).

[18] Noah Lorenz et al. 'Temporal associations of daily changes in sleep and depression core symptoms in patients suffering from major depressive disorder: idiographic time-series analysis'. In: *JMIR Mental Health* 7.4 (2020), e17071 (cited on pages 1–4, 6, 17).

[19] Nicholas C Jacobson and Yeon Joo Chung. 'Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones'. In: *Sensors* 20.12 (2020), p. 3572 (cited on pages 1, 6).

[20] World Health Organization. 'The World Health Report 2001: Mental health: new understanding, new hope'. In: (2001) (cited on page 2).

[21] Graham Thornicroft. 'Most people with mental illness are not treated.' In: *The Lancet* (2007) (cited on page 2).

[22] JC Meyer, M Matlala, and AK Chigome. 'Mental health care-a public health priority in South Africa'. In: *South African Family Practice* (2019), pp. 25–30 (cited on page 2).

[23] Caroline Mitchell, Brian McMillan, and Teresa Hagan. *Mental health help-seeking behaviours in young adults*. 2017 (cited on page 2).

[24] Laurie Hare Duke. 'The importance of social ties in mental health'. In: *Mental Health and Social Inclusion* (2017) (cited on page 2).

[25] Margalida Gili et al. 'Mental disorders as risk factors for suicidal behavior in young people: A meta-analysis and systematic review of longitudinal studies'. In: *Journal of affective disorders* 245 (2019), pp. 152–162 (cited on page 2).

[26] Benjamin L Hankin et al. 'Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study.' In: *Journal of abnormal psychology* 107.1 (1998), p. 128 (cited on page 2).

[27] BH Esbjørn et al. 'The development of anxiety disorders: Considering the contributions of attachment and emotion regulation'. In: *Clinical child and family psychology review* 15.2 (2012), pp. 129–143 (cited on page 2).

[28] B Sandin et al. 'The PANAS scales of positive and negative affect: Factor analytic validation and cross-cultural convergence.' In: *Psicothema* 11.1 (1999), pp. 37–51 (cited on page 3).

[29] John F Greden. 'Physical symptoms of depression: unmet needs'. In: *Journal of Clinical Psychiatry* 64 (2003), pp. 5–11 (cited on pages 3, 13).

[30] Anya Topiwala et al. 'Subjective cognitive complaints given in questionnaire: relationship with brain structure, cognitive performance and self-reported depressive symptoms in a 25-year retrospective cohort study'. In: *The American Journal of Geriatric Psychiatry* 29.3 (2021), pp. 217–226 (cited on pages 3, 13).

[31] Alexander S Young et al. 'Persistent depression and anxiety in the United States: prevalence and quality of care'. In: *Psychiatric Services* 59.12 (2008), pp. 1391–1398 (cited on page 3).

[32] Paul E Jose, Holly Wilkins, and Jason S Spendelow. 'Does social anxiety predict rumination and co-rumination among adolescents?' In: *Journal of Clinical Child & Adolescent Psychology* 41.1 (2012), pp. 86–91 (cited on page 3).

[33] Odin Hjemdal et al. 'The relationship between resilience and levels of anxiety, depression, and obsessive–compulsive symptoms in adolescents'. In: *Clinical psychology & psychotherapy* 18.4 (2011), pp. 314–321 (cited on page 3).

[34] Susan Nolen-Hoeksema and Jannay Morrow. 'A prospective study of depression and posttraumatic stress symptoms after a natural disaster: the 1989 Loma Prieta Earthquake.' In: *Journal of personality and social psychology* 61.1 (1991), p. 115 (cited on pages 3, 4, 36, 40).

[35] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 'The PHQ-9: validity of a brief depression severity measure'. In: *Journal of general internal medicine* 16.9 (2001), pp. 606–613 (cited on page 3).

[36] Isobel M Cameron et al. 'Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care'. In: *British Journal of General Practice* 58.546 (2008), pp. 32–36 (cited on page 3).

[37] Richard P Swinson. 'The GAD-7 scale was accurate for diagnosing generalised anxiety disorder'. In: *Evidence-based medicine* 11.6 (2006), p. 184 (cited on page 3).

[38] Robert L Spitzer et al. 'A brief measure for assessing generalized anxiety disorder: the GAD-7'. In: *Archives of internal medicine* 166.10 (2006), pp. 1092–1097 (cited on page 3).

[39] Corey LM Keyes, John M Myers, and Kenneth S Kendler. 'The structure of the genetic and environmental influences on mental well-being'. In: *American Journal of Public Health* 100.12 (2010), pp. 2379–2384 (cited on page 4).

[40] Kenneth S Kendler, John M Myers, and Corey LM Keyes. 'The relationship between the genetic and environmental influences on common externalizing psychopathology and mental wellbeing'. In: *Twin Research and Human Genetics* 14.6 (2011), pp. 516–523 (cited on page 4).

[41] Han Yu et al. 'Personalized wellbeing prediction using behavioral, physiological and weather data'. In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE. 2019, pp. 1–4 (cited on pages 4, 33).

[42] Ruoyu Wang et al. 'Cross-sectional associations between long-term exposure to particulate matter and depression in China: the mediating effects of sunlight, physical activity, and neighborly reciprocity'. In: *Journal of affective disorders* 249 (2019), pp. 8–14 (cited on page 4).

[43] Michael F Holick. 'A perspective on the beneficial effects of moderate exposure to sunlight: bone health, cancer prevention, mental health and well being'. In: *Comprehensive Series in Photosciences*. Vol. 3. Elsevier, 2001, pp. 11–37 (cited on page 4).

[44] Markus Jansson-Fröjmark and Karin Lindblom. 'A bidirectional relationship between anxiety and depression, and insomnia? A prospective study in the general population'. In: *Journal of psychosomatic research* 64.4 (2008), pp. 443–449 (cited on pages 4, 17).

[45] Pasquale K Alvaro, Rachel M Roberts, and Jodie K Harris. 'A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression'. In: *Sleep* 36.7 (2013), pp. 1059–1068 (cited on pages 4, 17).

[46] Mohammed A Al-Abri. 'Sleep Deprivation and Depression: A bi-directional association'. In: *Sultan Qaboos University Medical Journal* 15.1 (2015), e4 (cited on page 4).

[47] Milton Erman and Sonia Ancoli-Israel. 'Sleep and Sleep–Wake Disorders'. In: *Psychiatry* (2008), pp. 1626–1657 (cited on page 4).

[48] Michael E Thase. 'Depression and sleep: pathophysiology and treatment'. In: *Dialogues in clinical neuroscience* (2022) (cited on page 4).

[49] Meir H Kryger, Thomas Roth, and William C Dement. *Principles and Practice of Sleep Medicine*. Elsevier Health Sciences, 2010 (cited on page 4).

[50] D Jones et al. 'Sleep and depression'. In: *Psychopathology* 20.Suppl. 1 (1987), pp. 20–31 (cited on pages 4, 17).

[51] Chiara Baglioni et al. 'Sleep and mental disorders: A meta-analysis of polysomnographic research.' In: *Psychological bulletin* 142.9 (2016), p. 969 (cited on pages 4, 17, 37).

[52] James A Betts et al. 'Growth-hormone responses to consecutive exercise bouts with ingestion of carbohydrate plus protein'. In: *International journal of sport nutrition and exercise metabolism* 23.3 (2013), pp. 259–270 (cited on page 4).

[53] Raman K Malhotra. 'Sleep, recovery, and performance in sports'. In: *Neurologic clinics* 35.3 (2017), pp. 547–557 (cited on page 4).

[54] Michael Kellmann et al. 'Recovery and performance in sport: consensus statement'. In: *International journal of sports physiology and performance* 13.2 (2018), pp. 240–245 (cited on page 4).

[55] Michael Deuschle et al. 'Serum brain-derived neurotrophic factor (BDNF) in sleep-disordered patients: relation to sleep stage N3 and rapid eye movement (REM) sleep across diagnostic entities'. In: *Journal of sleep research* 27.1 (2018), pp. 73–77 (cited on pages 4, 37).

[56] Sarah Laxhmi Chellappa and John Fontenele Araújo. 'Excessive daytime sleepiness in patients with depressive disorder'. In: *Brazilian Journal of Psychiatry* 28 (2006), pp. 126–129 (cited on page 4).

[57] Kirstie N Anderson and Andrew J Bradley. 'Sleep disturbance in mental health problems and neurodegenerative disease'. In: *Nature and science of sleep* 5 (2013), p. 61 (cited on pages 4, 25).

[58] Thomas Stephens. 'Physical activity and mental health in the United States and Canada: evidence from four population surveys'. In: *Preventive medicine* 17.1 (1988), pp. 35–47 (cited on page 5).

[59] Scott A Paluska and Thomas L Schwenk. 'Physical activity and mental health'. In: *Sports medicine* 29.3 (2000), pp. 167–180 (cited on page 5).

[60] Robert Stanton, Brenda Happell, and Peter Reaburn. 'The mental health benefits of regular physical activity, and its role in preventing future depressive illness'. In: *Nursing: Research and Reviews* 4 (2014), pp. 45–53 (cited on page 5).

[61] T Christian North et al. 'Effect of exercise on depression.' In: (2008) (cited on pages 5, 16).

[62] Yeliz Dogru. 'The Effect of 8-Week Crossfit Training on Social Physical Anxiety Levels.' In: *African Educational Research Journal* 8 (2020), pp. 157–160 (cited on pages 5, 16).

[63] Michael S Bahrke and William P Morgan. 'Anxiety reduction following exercise and meditation'. In: *Cognitive therapy and research* 2.4 (1978), pp. 323–333 (cited on pages 5, 16).

[64] Marleen HM De Moor et al. 'Regular exercise, anxiety, depression and personality: a population-based study'. In: *Preventive medicine* 42.4 (2006), pp. 273–279 (cited on pages 5, 16).

[65] Andreas Ströhle. 'Physical activity, exercise, depression and anxiety disorders'. In: *Journal of neural transmission* 116.6 (2009), pp. 777–784 (cited on pages 5, 16).

[66] Egil W Martinsen, A Medhus, and L Sandvik. 'Effects of aerobic exercise on depression: a controlled study.' In: *British medical journal (Clinical research ed.)* 291.6488 (1985), p. 109 (cited on page 5).

[67] Andrea Camaz Deslandes. 'Exercise and mental health: what did we learn in the last 20 years?' In: *Frontiers in psychiatry* 5 (2014), p. 66 (cited on page 5).

[68] Kathleen Mikkelsen et al. 'Exercise and mental health'. In: *Maturitas* 106 (2017), pp. 48–56 (cited on page 5).

[69] Howard A Wenger and Gordon J Bell. 'The interactions of intensity, frequency and duration of exercise training in altering cardiorespiratory fitness'. In: *Sports medicine* 3.5 (1986), pp. 346–356 (cited on page 5).

[70] Mats Hallgren et al. 'Associations of exercise frequency and cardiorespiratory fitness with symptoms of depression and anxiety-a cross-sectional study of 36,595 adults'. In: *Mental Health and Physical Activity* 19 (2020), p. 100351 (cited on page 5).

[71] Sammi R Chekroud et al. 'Association between physical exercise and mental health in 1· 2 million individuals in the USA between 2011 and 2015: a cross-sectional study'. In: *The Lancet Psychiatry* 5.9 (2018), pp. 739–746 (cited on pages 5, 25).

[72] Arthur F Kramer and Kirk I Erickson. 'Effects of physical activity on cognition, well-being, and brain: Human interventions'. In: *Alzheimer's & Dementia* 3.2 (2007), S45–S51 (cited on pages 5, 25).

[73] Andrea Di Blasio et al. 'Acute and delayed effects of high intensity interval resistance training organization on cortisol and testosterone production.' In: *The Journal of sports medicine and physical fitness* 56.3 (2014), pp. 192–199 (cited on page 5).

[74] Thomas Reilly and Bjorn Ekblom. 'The use of recovery methods post-exercise'. In: *Journal of sports sciences* 23.6 (2005), pp. 619–627 (cited on page 5).

[75] John L Ivy. 'Muscle glycogen synthesis before and after exercise'. In: *Sports Medicine* 11.1 (1991), pp. 6–19 (cited on page 5).

[76] Jeremy N Morris and Adrianne E Hardman. 'Walking to health'. In: *Sports medicine* 23.5 (1997), pp. 306–332 (cited on pages 5, 12, 15, 34).

[77] Areum Han, Junhyoung Kim, and Jaehyun Kim. 'A study of leisure walking intensity levels on mental health and health perception of older adults'. In: *Gerontology and Geriatric Medicine* 7 (2021), p. 2333721421999316 (cited on pages 5, 12, 15, 34).

[78] Thomas R Prohaska et al. 'Walking and the preservation of cognitive function in older populations'. In: *The Gerontologist* 49.S1 (2009), S86–S93 (cited on pages 5, 15, 25, 34).

[79] Gary G Berntson and John T Cacioppo. 'Heart rate variability: Stress and psychiatric conditions'. In: *Dynamic electrocardiography* 41.2 (2004), pp. 57–64 (cited on page 5).

[80] Joachim Taelman et al. 'Influence of mental stress on heart rate and heart rate variability'. In: *4th European conference of the international federation for medical and biological engineering*. Springer. 2009, pp. 1366–1369 (cited on page 5).

[81] Giampaolo Perna et al. 'Heart rate variability: Can it serve as a marker of mental health resilience?: Special Section on "Translational and Neuroscience Studies in Affective Disorders" Section Editor, Maria Nobile MD, PhD'. In: *Journal of Affective Disorders* 263 (2020), pp. 754–761 (cited on page 5).

[82] Hye-Geum Kim et al. 'Stress and heart rate variability: A meta-analysis and review of the literature'. In: *Psychiatry investigation* 15.3 (2018), p. 235 (cited on page 5).

[83] E Ron De Kloet, Melly S Oitzl, and Marian Joëls. 'Stress and cognition: are corticosteroids good or bad guys?' In: *Trends in neurosciences* 22.10 (1999), pp. 422–426 (cited on page 5).

[84] Katrin Starcke and Matthias Brand. 'Decision making under stress: a selective review'. In: *Neuroscience & Biobehavioral Reviews* 36.4 (2012), pp. 1228–1248 (cited on page 5).

[85] Andrew H Kemp and Daniel S Quintana. 'The relationship between mental and physical health: insights from the study of heart rate variability'. In: *International journal of Psychophysiology* 89.3 (2013), pp. 288–296 (cited on page 5).

[86] John A Chalmers et al. 'Anxiety disorders are associated with reduced heart rate variability: a meta-analysis'. In: *Frontiers in psychiatry* 5 (2014), p. 80 (cited on page 5).

[87] Leonard I Pearlin et al. 'The stress process'. In: *Journal of Health and Social behavior* (1981), pp. 337–356 (cited on pages 5, 17).

[88] HM Van Praag. 'Can stress cause depression?' In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 28.5 (2004), pp. 891–907 (cited on pages 5, 17).

[89] Bruce S McEwen et al. 'Stress and anxiety: structural plasticity and epigenetic regulation as a consequence of stress'. In: *Neuropharmacology* 62.1 (2012), pp. 3–12 (cited on pages 5, 17).

[90] Anthony David. 'Self-reflection in illness and health: literal and metaphorical?' In: *Palgrave Communications* 3.1 (2017), pp. 1–6 (cited on page 6).

[91] Bens Pardamean et al. 'Quantified self-using consumer wearable device: predicting physical and mental health'. In: *Healthcare informatics research* 26.2 (2020), pp. 83–92 (cited on page 6).

[92] Eric N Smith et al. 'Integrating wearables in stress management interventions: Promising evidence from a randomized trial.' In: *International Journal of Stress Management* 27.2 (2020), p. 172 (cited on page 6).

[93] Louise V Coutts et al. 'Deep learning with wearable based heart rate variability for prediction of mental and general health'. In: *Journal of Biomedical Informatics* 112 (2020), p. 103610 (cited on page 6).

[94] Mijeong Kang and Kyunghwan Chai. 'Wearable sensing systems for monitoring mental health'. In: *Sensors* 22.3 (2022), p. 994 (cited on page 6).

[95] Albert K Liau et al. 'A Quasi-experimental study of a fitbit-based self-regulation intervention to improve physical activity, well-being, and mental health'. In: *Cyberpsychology, Behavior, and Social Networking* 21.11 (2018), pp. 727–734 (cited on page 6).

[96] Olaf Binsch, Thymen Wabeke, and Pierre Valk. 'Comparison of three different physiological wristband sensor systems and their applicability for resilience-and work load monitoring'. In: *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE. 2016, pp. 272–276 (cited on page 6).

[97] Anna Shcherbina et al. 'Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort'. In: *Journal of personalized medicine* 7.2 (2017), p. 3 (cited on page 6).

[98] Kayla J Nuss et al. 'Assessment of accuracy of overall energy expenditure measurements for the Fitbit Charge HR 2 and Apple Watch'. In: *American journal of health behavior* 43.3 (2019), pp. 498–505 (cited on page 6).

[99] Sirinthip Roomkham et al. 'Sleep monitoring with the apple watch: comparison to a clinically validated actigraph'. In: *F1000Research* 8 (2019), Article–number (cited on page 6).

[100] *The PHP framework for web artisans*. URL: https://laravel.com/ (cited on page 10).

[101] *Firebase*. URL: https://firebase.google.com/ (cited on page 10).

[102] *Kotlin programming language*. URL: https://kotlinlang.org/ (cited on page 10).

[103] *WeFitter: Health & Fitness Gamification API*. URL: https://www.wefitter.com/ (cited on page 11).

[104] Andrew G Kubala et al. 'Field-based measurement of sleep: agreement between six commercial activity monitors and a validated accelerometer'. In: *Behavioral sleep medicine* 18.5 (2020), pp. 637–652 (cited on page 12).

[105] Illia Fedorin et al. 'Sleep stages classification in a healthy people based on optical plethysmography and accelerometer signals via wearable devices'. In: *2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. IEEE. 2019, pp. 1201–1204 (cited on page 12).

[106] Milad Asgari Mehrabadi et al. 'Sleep tracking of a commercially available smart ring and smartwatch against medical-grade actigraphy in everyday settings: instrument validation study'. In: *JMIR mHealth and uHealth* 8.11 (2020), e20465 (cited on page 12).

[107] David Watson and Lee Anna Clark. 'The PANAS-X: Manual for the positive and negative affect schedule-expanded form'. In: (1994) (cited on page 13).

[108] Gregory J Boyle et al. 'Measures of affect dimensions'. In: *Measures of personality and social psychological constructs*. Elsevier, 2015, pp. 190–224 (cited on page 13).

[109] Ineke Demeyer et al. 'Rumination mediates the relationship between impaired cognitive control for emotional information and depressive symptoms: A prospective study in remitted depressed adults'. In: *Behaviour research and therapy* 50.5 (2012), pp. 292–297 (cited on page 13).

[110] Justin Thomas and Belkeis Altareb. 'Cognitive vulnerability to depression: an exploration of dysfunctional attitudes and ruminative response styles in the United Arab Emirates'. In: *Psychology and Psychotherapy: Theory, Research and Practice* 85.1 (2012), pp. 117–121 (cited on page 13).

[111] Kristof Hoorelbeke et al. 'The interplay between cognitive risk and resilience factors in remitted depression: a network analysis'. In: *Journal of Affective Disorders* 195 (2016), pp. 96–104 (cited on page 13).

[112] In-Kwon Yeo and Richard A Johnson. 'A new family of power transformations to improve normality or symmetry'. In: *Biometrika* 87.4 (2000), pp. 954–959 (cited on page 18).

[113] F. Pedregosa et al. 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cited on pages 18, 19).

[114] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017 (cited on page 18).

[115] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA, 2020 (cited on page 18).

[116] Douglas Bates et al. 'Fitting Linear Mixed-Effects Models Using lme4'. In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: 10.18637/jss.v067.i01 (cited on page 18).

[117] John L Horn. 'A rationale and test for the number of factors in factor analysis'. In: *Psychometrika* 30.2 (1965), pp. 179–185 (cited on page 19).

[118] Robert Tibshirani, Guenther Walther, and Trevor Hastie. 'Estimating the number of clusters in a data set via the gap statistic'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423 (cited on page 19).

[119] Alexandra Kuznetsova, Per B Brockhoff, and Rune HB Christensen. 'lmerTest package: tests in linear mixed effects models'. In: *Journal of statistical software* 82 (2017), pp. 1–26 (cited on page 19).

[120] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009 (cited on page 20).

[121] Wes McKinney et al. 'Data structures for statistical computing in python'. In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. 1. Austin, TX. 2010, pp. 51–56 (cited on page 20).

[122] Wesley T Blumenburg et al. 'Physical fitness, but not physical activity, is associated with mental health in apparently healthy young adults.' In: *The Journal of Sports Medicine and Physical Fitness* (2021) (cited on page 34).

[123] Erin Hoare et al. 'The associations between sedentary behaviour and mental health among adolescents: a systematic review'. In: *International journal of behavioral nutrition and physical activity* 13.1 (2016), pp. 1–22 (cited on page 34).

[124] Andreas Broocks et al. 'Exercise avoidance and impaired endurance capacity in patients with panic disorder'. In: *Neuropsychobiology* 36.4 (1997), pp. 182–187 (cited on page 36).

[125] Adrian Wells and Robert L Leahy. *Cognitive therapy of anxiety disorders: A practice manual and conceptual guide*. 1998 (cited on page 36).

[126] Patrick H Finan, Phillip J Quartana, and Michael T Smith. 'The effects of sleep continuity disruption on positive mood and sleep architecture in healthy adults'. In: *Sleep* 38.11 (2015), pp. 1735–1742 (cited on page 37).

[127] Keisuke Takano and Yoshihiko Tanno. 'Self-rumination, self-reflection, and depression: Self-rumination counteracts the adaptive effect of self-reflection'. In: *Behaviour research and therapy* 47.3 (2009), pp. 260–264 (cited on page 40).

[128] Charles A Czeisler et al. 'Bright light resets the human circadian pacemaker independent of the timing of the sleep-wake cycle'. In: *Science* 233.4764 (1986), pp. 667–671 (cited on page 41).

[129] Miriam I Hehlmann et al. 'The use of digitally assessed stress levels to model change processes in CBT-a feasibility study on seven case examples'. In: *Frontiers in Psychiatry* 12 (2021), p. 613085 (cited on page 42).

[130] Merav Mofaz et al. 'Self-Reported and Physiologic Reactions to Third BNT162b2 mRNA COVID-19 (Booster) Vaccine Dose'. In: *Emerging Infectious Diseases* 28.7 (2022), p. 1375 (cited on page 42).

[131] Emre Ertin et al. 'AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field'. In: *Proceedings of the 9th ACM conference on embedded networked sensor systems*. 2011, pp. 274–287 (cited on page 42).

[132] Philip S Wang et al. 'Delay and failure in treatment seeking after first onset of mental disorders in the World Health Organization's World Mental Health Survey Initiative'. In: *World psychiatry* 6.3 (2007), p. 177 (cited on page 42).