Classification of patients with depression and healthy controls based on behavioural patterns acquired from smartphone sensor data

Anna Hakala

School of Science

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

Espoo 29.7.2021

Thesis supervisor:

Prof. Aristides Gionis

Thesis advisor:

D.Sc. (Tech.) Talayeh Aledavood

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Anna Hakala

Title: Classification of patients with depression and healthy controls based on behavioural patterns acquired from smartphone sensor data

Date: 29.7.2021 Language: English Number of pages: 7+84

Department of Computer Science

Professorship: Theoretical Computer Science

Supervisor: Prof. Aristides Gionis

Advisor: D.Sc. (Tech.) Talayeh Aledavood

Mental health disorders comprise one of the highest burdens of disease in the world, and therefore it is important to use effective prevention methods and improve the treatment of patients. Digital phenotyping is a new field of study which uses data from wearable and consumer devices to find new behavioural markers and phenotypes.

This thesis uses the data from a digital phenotyping study with patients with depression and healthy controls to find behavioural markers for depression. Different methods such as, calculating the correlation between the mood of the subjects and the smartphone data, k-means clustering, classification using linear discriminant analysis and a decision tree classifier, and kh-segmentation, were used for this purpose.

A correlation between patient mood and behavioural changes was found. The classification of patients and controls ranged with an accuracy of 0.74-1.0 using the linear discriminant analysis and decision tree classifier, which is satisfactory considering the small sample size. The kh-segmentation method is a valid tool to be used in future research.

Keywords: Digital Phenotyping, Behavioural patterns, Machine learning, Mental health

AALTO-UNIVERSITETET
HÖGSKOLAN FÖR TEKNIKVETENSKAPER

SAMMANDRAG AV
DIPLOMARBETET

Författare: Anna Hakala

Titel: Klassificering av patienter med depression och frisk kontrollgrupp baserat på beteendemönster samlat från smarttelefonsensordata

Datum: 29.7.2021 Språk: Engelska Sidantal: 7+84

Institutionen för datavetenskap

Professur: Teoretisk datavetenskap

Övervakare: Prof. Aristides Gionis

Handledare: D.Sc. (Tech.) Talayeh Aledavood

Psykisk ohälsa utgör en av de största sjukdomsbördorna i världen. Därför är det viktigt att använda effektiva förebyggande metoder och förbättra behandling av patienter. Bestämning av digitalfenotyp är ett nytt studieområde där bärbara- och konsumentenheter används för att söka efter nya biomarkörer och fenotyper.

Denna avhandling använder data från en digitalfenotypstuide bestående av patienter med depression och friska kontroller för att hitta beteendemarkörer för depression. Olika metoder så som att, beräkna korrelationen mellan försökspersonernas humör och smarttelefondatan, k-medelkluster, klassificering med linjär diskriminantanalys och beslutsträdsinlärning, och kh-segmentering användes för detta ändamål.

Ett samband mellan patientens humör och beteendemässiga förändringar hittades. Klassificeringen av patienter och kontroller med hjälp av linjära diskriminantanalysen och beslutträdsinlärningen hade en noggrannhet på 0,74-1,0, vilket är tillfredsställande med tanke på den begränsade provstorleken. För framtida forskning är kh-segmentering ett giltigt verktyg.

Nyckelord: Bestämning av Digitalfenotyp, Beteendemönster, Maskininlärning, Mentalhälsa

Preface

It has been a couple of strange years. The Covid-19 pandemic hit, and everything changed. I find it ironic that I have learnt so much about mental health and then suffered from a burnout myself, which led me to take a long break from my thesis. Now when I am finally graduating, I have not only learnt a lot from my studies, but I have also learnt more about myself and that I should be more kind to myself. I would not have come this far without support, and I have spent much time on thinking about all the people that have affected my life in a positive way.

I would like to thank:

My parents and my brothers.

My grandparents, aunts, and uncle.

My friends that I got to know from living in Vaasa, Turku, and Espoo.

My online friends that kept me company so many evenings when there was nothing else to do than play games.

My remote study group and the Pomodoro timer.

The Complex Systems research group for being so inspiring.

My supervisor professor Aris Gionis for all guidance and suggestions.

My advisor Talayeh Aledavood not only for supporting me, but also for being such a good role model and not giving up on me.

Finally, I would like to thank Magnus, for believing in me when I did not and helping me move forward when I did not have the strength to.

The summer cottage, 16.7.2021

Anna L. I. Hakala

Contents

Abbreviations

Abbreviations

BIC	Bayesian information criteria
LDA	Linear Discriminant Analysis
PCA	Principal Component Analysis
PHQ-9	Patient Health Questionnaire

1 Introduction

In 2018, 80 % of Finns aged between 16 and 89 had a smartphone in personal use [1]. These small computers can sense the user's surroundings, as well as, monitor the actual usage of the smartphone [22]. Thus by analyzing the data, the researcher can get insight of the user's individual and social behaviour. For example, researchers can follow how the user makes calls, writes texts, and uses social media and other apps. Smartphones also contain the opportunity for more low-level sensor measures, for example, GPS, noise and light. If this data is collected, the smartphone can be used to monitor and analyse behavioural patterns in the user. Behaviour is something that has been tracked by, for example, advertising firms to suggest personalized ads. The smartphones gives us a new way to follow a person's life, by collecting and analysing data from it.

A promising new field of research is mobile mental health [18][29][64]. Mental health is an important part of everyone's life; good mental health is fundamental for a persons well-being, whereas, mental illnesses affects oneself and people around them negatively. Mental health disorders affect 18.8 % of Finnish citizens, according to the OECD report from 2018, which is the highest prevalence in the EU [44]. In Finland, mental health problems lead to a cost of 11 billion euros per year. Patients have traditionally had face-to-face meetings with a mental health professionals. In these meetings they discuss the patients symptoms and behaviour from a subjective view. The clinician then has to make a diagnosis based on what is said during the meetings. Mobile mental health technologies can provide the clinician and patient with objective data, gathered outside the meetings, that can be analysed. Analysis would bring forth a deeper understanding of the patients daily life, which further opens up options to treat a patient based on data gathered in the patients' own environment. The data gathered is not affected by recall biases [29]. Recall bias is the error that occurs due to that a person is not able to accurately remember an event or past experience [32]. In the case of mental health, a recall bias would be the patient having problems to remember what has happened between the meetings with the clinician, often resulting in inaccuracy. Analysing gathered data enables new options to recognize behaviour and mood changes and helps adjusting the treatment accordingly [18].

Mobile mental health is a broad concept. Including apps with active and passive data collection. Active data requires user actions, whereas passive data is automatically collected by a device [66]. As well as, apps that monitor the user for diagnostic purposes or apps that do active interventions. Being able to monitor the effect of daily behaviors on mood could further help to predict and classify the state of a patient.

This thesis aims to explore possible analysis methods for passive mobile data collected in Mobile Monitoring of Mood (MoMo-Mood) Pilot study [68]. Thus improving the evidence-based digital health research within mental health, by trying to standardized measurements and outcomes. One concrete goal is to classify patients and controls from the behavioural patterns found in the collected passive mobile data. A second goal is to segment the data to detect the changes in behaviour. The analysis

results could hopefully help the patient and the clinicians to better understand the changes in the patients mood, for example, has it improved or gotten worse. Giving the clinician and the patient the possibility to see and discuss when changes have occurred and why, could be a good tool to improve the treatment plan. Firstly, an exploration of the passive data is done. From there a search for correlation between the changes in behaviour and mood is performed. Secondly, machine learning is used to classify if a person is part of the control group or a patient. Lastly, another method was examined, the kh-segmentation, which is used to look for changes in behaviour. The calculations presented in this thesis are performed using computer resources within the Aalto University School of Science "Science-IT" project.

The rest of this thesis is organized as follows. In Section 2 behavioural patterns are described and how digital devices create these patterns in personal usage. The section also explains how these behavioural patterns can be used to detect mental health disorders. In Section 3 the data used i this thesis is described and the methods are presented. Section 4 describes the experimental results. The 5 Section does discusses the results and does conclusions. Section 6 is a summary.

2 Background

This chapter introduces the interdisciplinary fields included in this thesis. First, the potential use of behavioural patterns acquired from sensor data is presented in Section 2.1. Second, a introduction to mental health will be given in Section 2.2.

2.1 Behavioural patterns

Individuals usually exhibit persistent behavioral patterns [9], which can be monitored and discovered in several different ways [5]. The more traditional way is to gather information via self-report questionnaires and observation in an artificial environment setting [51]. Nowadays people interact with and carry digital devices that can monitor the individual seamlessly, this data can be seen as behavioural information [45]. This section will introduce how digital devices can be used to get an insight into individuals behavior.

2.1.1 Quantification of behavioural patterns using digital footprints

Device and user interactions can be collected as data. This type of data is called digital footprints [45][46]. A definition of digital footprints is "the digital traces that people leave while interacting with cyberphysical spaces" [69]. Gathering and combining the data from devices open up the possibility to explore individual-level behaviour [14]. For example, a smartphone can tell about the social interactions made via applications and phone calls, the GPS data can tell if the individual has been traveling and even how frequently the screen is being turned on and off can tell about behavioural patterns [7][8]. There are several studies showing how human activity usually appears in patters, for example, in bursts of social activity both on individual-level and inter-individual interactions [15][53][40]. Even though the data is gathered on an individual-level the data can be combined into a network of interactions between individuals, making analyzing group behaviour, social interaction and community dynamics, possible [15][53][40][14]. Depending on the device, the data gathered can be used to analyze, for example, social behaviour to human health to activity and sleep rhythms [9][10][38].

2.1.2 Digital phenotyping

The definition of digital phenotyping used in this thesis is "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices given" coined in [65]. This means that we aim to quantify human behavior in a natural environment, that is in their everyday life. By measuring human behaviour we can, for example, look for behavioural markers which indicate out health or disease.

Digital phenotyping is a fairly new field as it has been made possible through the era of digitalization. We have not had the sensors nor the computational capacity to

gather and analyse data to a sufficient extent until recently. As digital components have become smaller, the computational capacity has also grown. We can now store huge amounts of data that can be processed using different methods to get results. This is all thanks to the invention and further development of the transistor, which is observed in the Moore's law [42].

As mentioned in Chapter 1, the smartphone has become a part of our daily lives and in 2018, 80 per cent of Finns aged between 16 and 89 had a smartphone in own use[1]. Smartphones is also often a very personal device mostly used by one person. The smartphone contains several digital sensors that enable the collection of different data, for instance, GPS, accelerometer, social interaction via calls and messaging, ambient noise [6][66]. Gathering this data when in a natural day-to-day environment gives us the opportunity to analyze complex longitudinal multivariate data. Digital sensors can also be found in other devices such as watches, bed sensors and even clothing. The most important properties of these devices is that they are small enough to conveniently collect data.

Data can also be gathered through active participation by the user, for example, by answering surveys [66]. The data collected in a way that demands subject active participation is called active data and data collected automatically without subject engagement for the purpose of the study is called passive data [66]. By collecting both active and passive data, it is possible to connect, for example, surveys about daily exercise with actual GPS and accelerometer data. Digital phenotyping could approach analysing active and passive data using statistical tools and machine learning to find biomedical and clinical insights[45]. To be able to quantify the results, the quality of the raw data, the features and methods used need to be chosen wisely. For example, using a supervised machine learning method instead of a complex deep learning method, makes it easier for the scientist to follow the steps of the method and understand how the result is shaped[26].

Digital phenotyping has become a token of interest especially in the field of mental health [45][27][30][17]. This thesis aims to find quantifiable observations within the passive data, which can be linked with the mental health condition of the patients and then further used as features in training the machine learning model. The models could, for example, predict future mood or classify the subject into a group. The latter one is one of the goals for this thesis. The model is intended to classify the subject either as a patient or as a healthy control. The methods are described in Chapter 3 and the results are presented in Chapter 4.

2.2 Mental health disorders

First, a brief introduction to mental health in general will be given. Then the potential use of behavioural patterns for mental healthcare is presented.

The WHO [47] defines mental health as "a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community".

The OECD [44] says that, mental health is an important part of everyone's life; good mental health is fundamental for a persons well-being. Mental disorders can affect all people, regardless of gender, age and background. Mental illness makes it hard to carry out our daily lives. In Europe, for example, tens of millions EU citizens endure one mental health problem at some point of their life [44]. There occurs tens of thousands cases of death directly due to mental health disorders or due to suicide [44]. Mental illness can impact a persons education, work and social activity. Finland is the EU country with highest prevalence of mental health disorders, with a rate of 18.8 %, according to the OECD health report from 2018 [44]. Mental health disorders comprise one of the highest burdens of disease in Finland and therefore it is important to use prevention methods and to improve the treatment of the patients.

The field of medicine is undergoing a transformation due to big data and the methods available, such as machine learning, which transform the data into significant clinical knowledge. We already have and are to see improvements in prognosis, displacement of work and improvement in diagnostic accuracy [43]. There is also opportunities within psychiatry, which is the medical specialty devoted to the diagnosis, prevention, and treatment of mental disorders.

There have been other studies using mobile phone sensors to create context-aware systems. For example, [18], conducted in 2011 is one of the first ecological momentary interventions for unipolar depression, as well as, attempts to use context sensing to identify mental health-related state. The results showed the feasibility of the system. Another study [25], compared a smartphone based mood state and mood state change detection system to self-assessment questionnaires. Using data from 9 patients and a total of 800 days, they showed that the automatic detection is much closer to the objective psychiatric diagnosis.

The pilot study data used in this thesis is from major depressive disorder patients and healthy controls collected in Mobile Monitoring of Mood (MoMo-Mood) Pilot study [68]. The main study, which is currently being conducted, data consists of patients with Major Depressive Disorder (MDD), Bipolar Syndrome (BD), Borderline Personality disorder (BPD) and healthy controls. The pilot study is described more in Section 3.1.4.

2.2.1 Patient Health Questionnaire, PHQ-9

This section discusses the importance of the Patient Health Questionnaire-9 (PHQ-9) and how the results of the questionnaires can be useful to track behavioural changes.

PHQ-9 score	Depression severity
0-4	None
5-9	Mild
10-14	Moderate
15-19	Moderately severe
20+	Severe

Table 1: PHQ-9 score and depression severity.

The PHQ-9 is the 9-item depression module from the full Patient Health Questionnaire (PHQ) [59][35]. It is used for screening, establishing depressive disorder diagnoses, monitoring and also measuring the severity of depression. The PHQ-9 asks questions about the severity of depression symptoms for the past 2 weeks. Each question asks how often have you felt this symptom and is scored between 0 to 3, "not at all" to "nearly every day". The PHQ-9 total score ranges from 0 to 27. The PHQ-9 scores of 5, 10, 15, and 20 represent the limit for mild, moderate, moderately severe, and severe depression, as shown in table 1. The presence of 5 or more symptoms and a score of more than 10 is criteria for major depression.

PHQ-9 is a useful tool, but is dependent of the patients retrospective recall [62]. The possibility to monitor the subjects behaviour via the smartphone could reduce the recall bias. A study [67] investigating major depressive disorder and use of smartphones to track the PHQ-9 score, showed that the smartphone collected scores both had a high adherence and correlated with the pen and paper PHQ-9 scores. It further says that the scores were higher for the smartphone collected scores and more subjects answered that they had suicidal thoughts.

3 Research material and methods

The previous chapter presented the problems and opportunities of using behavioural data acquired via sensors for mental healthcare. This chapter describes data and methods implemented in the thesis. Sections 3.1, 3.2 and 3.2.1 describe the data and the methods implemented for preprocessing of the data. Section 3.3 provides the methods for choosing the features to be implemented using machine learning. Sections 3.4 and 3.5 present the machine learning methods for classifying patients and for evaluating the results.

3.1 The Data

The data used in this thesis is from the MoMo-Mood pilot [68] MoMo-Mood study run at HYKS, the Helsinki University Central Hospital. The data is gathered from smartphone apps, actigraph devices and bed sensors used by patients with mental disorders as well as healthy controls. The data is of a time series type, which means that the data points are indexed in time order. The data can be divided into two categories active data and passive data. Active data requires active participation from the subject, for example, the subject answers a questionnaire about how he/she is feeling. Passive data is generated automatically through sensors in the smartphone, such as, communication or GPS.

One of the strengths of the data is that it contains subjects with data from a long period of time. For example, there exists subjects that have attended the study for up to eleven and fourteen months. Therefore, it is possible to study more long term changes in behaviour. The time stamps for several of the features are also recorded with fairly short intervals, providing us with detailed data. However, a downside of the data is that there are only thirty-seven subjects in the pilot study, which could be seen as too few for many modelling methods. Nevertheless, there are methods that manage to create working models from a small sample size.

3.1.1 The data collection platform

The data used in this thesis was collected with a digital platform called Koota [6], which is a research data collection system based on the Non-Intrusive Individual Monitoring Architecture, Niima, which is designed especially for mental health studies [6]. Koota can be used in any kind of study requiring multi-sensor and/or multi-device data collection from human participants. The three main use cases are individuals actively collecting their own data, collecting data from many subjects and a hybrid of these were the individual collects own data and donates it to research to be reused. There are a lot of advantages in collecting data on an individual's daily life and behavior, for example, the data is not affected by recall biases. However, as the data can be seen as very private, this opens up the possibility of privacy breaches. The paper, [6], addresses the problem of collecting data for only one study and also

presents a platform which makes it possible to reuse data in other studies while keeping the privacy of the participants. The platform has three key design features, flexibility of access control, flexibility of data sources and first-order privacy protection.

Flexibility of access control makes it possible to conduct several studies with the same data without compromising the privacy.

Flexibility of data sources permits easy linking of the sources collected from different studies.

First-order privacy protection is protection of the subjects who participate in a study, but also prevents breaches of privacy regulations by the researchers. Meaning that the researchers will not have enough access to breach the privacy of the subjects.

As mentioned in Section 3.1, this thesis uses passive data gathered from mobile phones. This is possible due to Koota including the AWARE framework in the application of the system [2]. The AWARE framework is a Android instrumentation framework for logging, sharing and reusing smartphone content. It is open-source and community supported and maintained by the University of Oulu in Finland. The AWARE framework enforces privacy in their design. It does not log personal information, such as phone numbers or contacts information. AWARE collects smartphone sensor and plugin data for the study being conducted. This data is uploaded to the study servers. As the AWARE API is used for logging the smartphone data it also restricts which data can be collected. The Section 3.1.2 will describe the sensors and their features further.

Koota also collects active data in the form of surveys and questionnaires.
Another safety measure is that the platform performs hashing on the data first on the device and a second time on the server. The hashing is done by adding a secret salt [34].

The data collected on the server is not accessed directly by the researchers, but must be pushed through converters [3]. These converters are study specific giving the researchers only the data they need for the study they are conducting. These converters may apply random transformations to the data for privacy and they can also add higher aggregation levels before the data is extracted from the server.

As the data is longitudinal data containing timestamps for each action, it is a privacy threat. The privacy threat is that if actions and timestamps are kept the same, they can be more easily linked to the subject. For example, if someone knows at which times a subject made calls, the person can easily find that subjects call data and therefore also find other data in the database. The timestamps are therefore also altered to prevent identification of subjects.

One last safety measure is that, there is not one single person who can have

access to both the identities of the participants and the data about them. Data from multiple sources coming from each participant is securely linked to their account on the Koota platform and can later be anonymously accessed by the researchers.

Finally, researchers can access the data using niimpy [4], which is a Python package for managing individual-level data developed in the Niima project. The niimpy package opens the databases, provides a querying shortcut for basic operations, as well as, a few more high-level operations. The further pre-processing and actual analysis of the subject data is done by the researchers.

3.1.2 The sensors and their features

There were several phone sensors used to collect data. This section will go through them and explain what they are. The converters applied on the raw data will also be described.

The data collection platform, described in Section 3.1.1, uses the AWARE framework for passive smartphone sensing [22][2]. The research codes converters apply a thin wrapper over the data collected via the AWARE framework. As mentioned in Section 3.1.1, niimpy is a python package used to access the individual-level data from the servers. The package contains functions that opens subject data into tables and maps the raw data into better assembled data. These functions will be described for the sensors that they apply. Niimpy also has functions for missing data and subject selection, but these will be discussed in Chapter 3.2, about preprocessing.

The sensor data used in this thesis includes applications, battery, communication, location, ambient noise and screen.

The applications sensor logs the usage of applications as well as the notifications of applications, see table 3. It logs when the user turns on, changes or turns off the application. The application name is logged, but the purpose of the application is not given by AWARE. The definition and grouping of the applications is left for the researcher analysing the data. Using niimpy, one can apply a mapping from the apps data to groups, returning a dataframe with duration and count of application usage for each group. The application groups are as follows; Sports, Games, Communication, Social Media, News, Travel, Shop, Entertainment, Work/study, Transportation, and Other.

The battery sensor logs phone power related data, including, when it is turned on, turned off, rebooting, the battery level, if the phone is charging or discharging and the battery health. The battery plays an important part in identifying missing data, and further selection of data and subjects. This is described in Section 3.2.1.

The communication sensor logs calls and messages to and from the smartphone. The AWARE framework saves the contacts with a unique ID which is encrypted. This makes it possible for the researcher to see how much the subject communicates with

Sensor	Description
Battery	The Battery sensor monitors battery information and monitors power related events (e.g., phone shutting down, rebooting). This sensor provides user-driven contexts, such as initiating a charge and unplugging the device.
Screen	The screen sensor monitors the screen statuses, such as turning on and off, locked and unlocked.
Ambient noise	This plugin measures the ambient noise (Hz, dB) as noisy or silent moments. It adds the daily noise exposure on the stream, showing the average dB and Hz per hour throughout the day.
Locations	The locations sensor provides the best location estimate for the users' current location, automatically. We have built-in an algorithm that provides the user's location with a minimum battery impact. However, we offer the flexibility to researchers to change how frequently the location gets updated, the minimum accuracy and others. In our endurance tests, we got a full day of location updates (8h and higher, depending on device usage) from the user with the default parameters..
Communication	The Communication sensor logs communication events such as calls and messages, performed by or received by the user. This sensor does not record personal information, such as phone numbers or contact information. Instead, an unique ID is assigned that is irreversible (SHA-1 encryption) but it is always the same for the same source. It also provides higher level context on the users' calling availability and actions. Android does not officially support messages content providers and messages sensor functionality might break at some point. If it does, contact us as we will try our best to fix it..
Applications	The Applications sensor logs application and notifications usage on the device. It captures every time the user changes from an application and keeps track of what is running in the background. It may also monitor any new application's notifications and capture when an application has crashed.

Table 2: Sensors as described by the AWARE framework documentation on their website [2].

the same contact, but neither the server nor the researcher gets access to personal information, such as, phone numbers. Niimpy gets the five most frequent contacts for the chosen period of time and calculates the duration and amount of times calls

Table field	Field type	Description
_id	INTEGER	Primary key, auto incremented
timestamp	REAL	Unixtime milliseconds since 1970
device_id	TEXT	AWARE device UUID
package_name	TEXT	Application's package name
application_name	TEXT	Application's localized name
is_system_app	BOOLEAN	Device's pre-installed application

Table 3: Application sensor log fields as described by the AWARE framework documentation on their website [2].

Table field	Field type	Description
_id	INTEGER	primary key, auto incremented
timestamp	REAL	unixtime milliseconds since 1970
device_id	TEXT	AWARE device UUID
battery_status	INTEGER	one of the Android's battery status, phone shutdown (-1) or rebooted (-2)
battery_level	INTEGER	the battery level, between 0 and SCALE
battery_scale	INTEGER	the maximum battery level
battery_voltage	INTEGER	the current battery voltage
battery_temperature	INTEGER	the current battery temperature
battery_adaptor	INTEGER	one of the Android's battery plugged values
battery_health	INTEGER	one of the Android's battery health values
battery_technology	TEXT	the battery chemical technology (e.g., Li-Ion, etc.)

Table 4: Battery sensor log fields as described by the AWARE framework documentation on their website [2].

have been made.

Table field	Field type	Description
_id	INTEGER	primary key, auto incremented
timestamp	REAL	unixtime milliseconds since 1970
device_id	TEXT	AWARE device UUID
call_type	INTEGER	one of the Android's call types (1 – incoming, 2 – outgoing, 3 – missed)
call_duration	INTEGER	length of the call session
trace	TEXT	SHA-1 one-way source/target of the call

Table 5: Communication calls log fields as described by the AWARE framework documentation on their website [2].

The locations sensor logs the GPS location estimate of the smartphone, see table 7. This data goes through transformations so that the privacy can be secured. The data the researcher accesses is an aggregation for each day, where a day goes from

Table field	Field type	Description
_id	INTEGER	primary key, auto incremented
timestamp	REAL	unixtime milliseconds since 1970
device_id	TEXT	AWARE device UUID
message_type	INTEGER	message type (1 – received, 2 – sent)
trace	TEXT	SHA-1 one-way source/target of the message

Table 6: Communication message log fields as described by the AWARE framework documentation on their website [2].

04:00 to 04:00 the next day. The aggregation consists of daily movement in the of total distance, the radius mean and location variance. The documentation however defines the variables as follows; locstd is the radius of gyration of locations, after the binning (meters), and radius_mean isn't exactly a radius, but the longest distance between any point and the mean location (both mean location and other points after binning). The documentation also says that the radius_mean should be compared to locstd to make sense.

Table field	Field type	Description
_id	INTEGER	primary key, auto incremented
timestamp	REAL	unixtime milliseconds since 1970
device_id	TEXT	AWARE device UUID
double_latitude	REAL	the location's latitude, in degrees
double_longitude	REAL	the location's longitude, in degrees
double_bearing	REAL	the location's bearing, in degrees
double_speed	REAL	the speed if available, in meters/second over ground
double_altitude	REAL	the altitude if available, in meters above sea level
provider	TEXT	gps or network
accuracy	INTEGER	the estimated location accuracy
label	TEXT	Customizable label. Useful for data calibration or traceability

Table 7: Location log fields as described by the AWARE framework documentation on their website [2].

The Ambient Noise is a plugin which records audio heard via the phone to measures the decibel levels. The plugin does not log the audio recorded. The plugin is set to record the surroundings every 30 minutes. The converter used for noise creates the fields; is_silent, double_decibels, double_silence_threshold, double_rms, double_frequency, and blob_raw. This is similar to the table 8. The double_decibels is the loudness of noise recorded in the surroundings. The double_frequency is the sound frequency in Hz. A dB threshold is set to determine if the surroundings are silent or noisy at the recorded moment, this is shown in is_silent. To classify silent

or not silent, the double_rms is calculated. The double_rms is the root mean square error and is calculated using both double_frequency and double_decibels. From here the preprocessing starts, see Section 3.2.

Table field	Field type	Description
_id	INTEGER	primary key auto-incremented
timestamp	REAL	unix timestamp in milliseconds of sample
device_id	TEXT	AWARE device ID
double_frequency	REAL	sound frequency in Hz
double_decibels	REAL	sound decibels in dB
double_RMS	REAL	sound RMS
is_silent	INTEGER	0 = not silent 1 = is silent
double_silence_threshold	REAL	the used threshold when classifying between silent vs not silent
blob_raw	BLOB	the audio snippet raw data collected

Table 8: Noise log fields as described by the AWARE framework documentation on their website [2].

The screen sensor logs the status of the screen, for example, when the phone screen is turned on and off, or is locked and unlocked. The fields can be seen in table 9. Note that screen flashes also return the on or off status when, for example, a notification pops or an call is incoming. This is simple data, represented by 0,1,2 and 3, but when preprocessed and analysed behavioural patterns can be found.

Table field	Field type	Description
_id	INTEGER	primary key, auto incremented
timestamp	REAL	unixtime milliseconds since 1970
device_id	TEXT	AWARE device UUID
screen_status	INTEGER	screen status, one of the following: 0=off, 1=on, 2=locked, 3=unlocked

Table 9: Screen log fields as described by the AWARE framework documentation on their website [2].

3.1.3 Questionnaires and surveys

Questionnaires and surveys give researchers a way to track changes in, for example, mood, behaviour or opinion. This thesis used PHQ-9 questionnaire answers, where the subject has answered questions reminiscing the past 14 days. PHQ-9 is described in Section 2.2.1. This type of data is seen as active data, as the subjects has to actively answer the questions about their mood. These questionnaires were conducted through Web surveys via Koota. The PHQ-9 questionnaire was answered in the beginning of the study and every 2 weeks after that. In this thesis the answers given

by the subjects were seen as ground truth of their mood and used in the segmentation part of the thesis, see Section 3.4.5.

3.1.4 The Pilot study

The pilot study [68] had fourteen patients with major depressive disorder and twenty-three healthy controls, see figure **??**. The participants were given devices that monitored their sleep, activity and factors such as heart rate and respiration. The participants also installed the smartphone app which gathers data on the smartphone usage and the surroundings. The application also pushed questionnaires which the subjects could answer.

Table 10 shows the data gathered via the mobile application for each patient and table 11 shows the data gathered for each control. The days with battery data column shows how many days of possible gathered data there is. If the mobile phone is on the app will gather battery data. The first goal of the thesis was to be able to classify between patients and controls. For this enough location, screen, noise or app data is needed. It is seen in table 10 that five patients have NaN values in their location, screen, noise and app data, whereas one patient has one day of location data and NaN in screen and noise. In this table NaN stands for not being able to calculate the amount of gaps in the data, which means that there is no data. In comparison 0 gaps mean that there are no gaps found. This is reinforced by the days with battery data column, where the NaN is followed by no battery data. This is further discussed in Section 3.2.1 under preprocessing. Table 11, has similarly five controls with NaN values in location, screen, noise and app data. In conclusion there seems to be possibly eight patients and eighteen controls available for the classification part of the thesis.

The second goal for this thesis was to analyse how the mood changes with respect to the behaviour patterns acquired from the mobile sensor data. In order to do this, enough of PHQ-9 questionnaires has to be answered for each subject. To be able to analyse change in mood the minimum requirement is two different PHQ-9 questionnaire scores, with adhering mobile sensor data. As previously discussed in Section 3.1.3, the first PHQ-9 questionnaire is answered in the beginning of the active phase, meaning that there is no passive data gathered yet to connect with the first PHQ-9. Table 10, shows that eight patient have only answered the first PHQ-9 questionnaire and two patients have answered two PHQ-9 questionnaires. Table 11 has one control that has only the first PHQ-9 questionnaire answered and five controls that have answered two PHQ-9 questionnaires. The same constriction for location, screen, noise and app data applies for acquiring behaviour patterns, there has to exist data to be able to analyse it. There seems to be possibly four patients and fourteen controls available for analysing changes in the mood with regards to behaviour patterns found in mobile data.

In the Results chapter 4 it is shown that the data amount is insufficient for proper

verification for reaching both goals of the thesis. Therefor, the main study is used for verification of the methods used on the pilot study.

	number of PHQ9	days with battery data	days with location data	screen gaps	noise gaps	app gaps
patient 1	1	0	NaN	NaN	NaN	NaN
patient 2	1	0	NaN	NaN	NaN	NaN
patient 3	1	0	NaN	NaN	NaN	NaN
patient 4	1	0	NaN	NaN	NaN	NaN
patient 5	1	0	NaN	NaN	NaN	NaN
patient 6	1	20	17	9	0	0
patient 7	1	20	21	1	16	6
patient 8	1	585	394	181	3	350
patient 9	2	0	1	NaN	NaN	0
patient 10	2	16	2	2	0	3
patient 11	3	79	30	16	1	2
patient 12	4	104	106	9	0	1
patient 13	6	241	121	6	7	2
patient 14	7	225	227	55	0	0

Table 10: Pilot study patient data statistics. A gap is no data for at least 6 hours. NaN means that gaps could not be calculated because of no data.

3.2 Preprocessing

In this thesis missing data is verified with battery data, which is described in Section 3.2.1 and the selection of subjects is described in Section 3.2.2. Different aggregations are performed on the data depending on the method. This preprocessing is partly done with niimpy. The niimpy part is described in Section 3.2.3.

3.2.1 The use of battery data to find missing data

Missing data is a known problem in all data analysis. Since this thesis aims to detect important features in subject behaviour, it is important to know why the data is missing, as it can be part of the subjects behaviour. For example, if a sensor is out of order or the data has not been correctly added to the database, it is missing data as a result of a technical error, whereas, if the mobile phone was simply turned off by the user it can be a manifestation of symptoms, taking form as a result of lack of activity. In order to classify the gaps in the total data, the battery data was used as a reference. Given that the battery sensor acquires data in regular intervals as long as the phone is on, It can be assumed that the battery sensor is the most reliable

	number of PHQ9	days with battery data	days with location data	screen gaps	noise gaps	app gaps
control 1	1	0	NaN	NaN	NaN	NaN
control 2	2	0	NaN	NaN	NaN	NaN
control 3	2	14	11	1	0	1
control 4	2	16	17	3	0	0
control 5	2	335	221	43	24	4
control 6	2	345	346	2	5	29
control 7	3	0	NaN	NaN	NaN	NaN
control 8	3	19	18	5	0	9
control 9	3	34	36	2	0	0
control 10	3	63	62	1	0	0
control 11	4	0	NaN	NaN	NaN	NaN
control 12	4	0	NaN	NaN	NaN	NaN
control 13	4	81	82	4	0	80
control 14	5	66	27	27	0	17
control 15	5	75	66	23	1	97
control 16	6	92	26	15	13	5
control 17	7	440	134	74	241	38
control 18	8	145	135	29	0	14
control 19	8	161	123	58	0	7
control 20	8	223	117	6	1	22
control 21	12	253	252	54	42	118
control 22	15	365	220	59	16	86
control 23	16	364	53	10	0	0

Table 11: Pilot study control data statistics. A gap is no data for at least 6 hours. NaN means that gaps could not be calculated because of no data.

sensor. Meaning that if there should be any data, there should also be battery data and if there is no battery data, the phone could not have been on. The data gaps were classified into real gaps, non-battery gaps and battery gaps. A real gap has a gap in both the sensor and the battery data, meaning the phone most probably has been turned off and therefore the data is missing. A non-battery gap has no gap in the battery data and a gap in the sensor data, meaning that the sensor has not collected any data, but it can be seen that the phone has been on. A battery gap is a gap in the battery data, but there can be found sensory data, which would indicate that battery data has somehow not been collected. There were no battery gaps found in the whole data set, which further demonstrates the usability of battery data as a missing data indicator. If missing battery data occurs, it would be assumed that it is missing due to an error in the database. However, if this is not an error in the database, it should be an error in the battery, which would be indicated in the

battery health output. The battery health output is discussed further bellow.

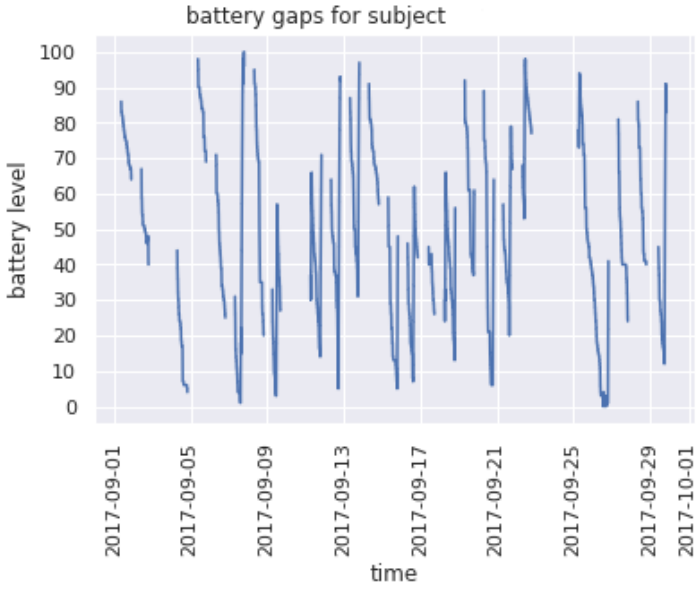

Figure 1: Example of battery data. The plot shows how the battery level changes between 0-100. Real battery gaps, missing battery data and sensor data, are seen where the line is cut off.

Another indicator for missing data is a fast decrease of the battery level. A fast decrease of the battery level indicates that the phone is being used and therefore other sensory data should be available. As the data used in this work has been acquired in Finland, it is important to take the possibility of cold weather into account. Cold weather discharges the battery faster and can cause problems with the devices [13]. Moreover, heat may cause a problem as a battery that is exposed to heat for a prolonged time will start to degrade the battery performance [13]. However, the battery sensor outputs the battery health, so these kinds of battery statuses can be noticed, if needed.

The last, but not least, aspect to consider with battery is the battery level. Was the phone simply turned off due to an empty battery and was there something significant draining the battery. It is interesting to see what the user was doing before the battery was drained. The phone can also have been shut down by the user. Is the turning off of the phone a significant behaviour? One scenario could be that the phone was turned off with low battery and turned on with full battery, indicating that the user wanted to charge the phone faster by turning it off. Another scenario is that the user wants to take a distance from the phone.

One of the biggest decisions regarding the data and subject selection is deciding how long can the smartphone be turned off until the data gap affects the analysis

Gap	Battery	Sensor	Description
Real gap	X	X	A gap in both the sensor and the battery data. The phone has probably been turned off and therefore the missing data.
Non-battery gap	✓	X	No gap in the battery data, but a gap in the sensor data. No collected sensor data, while the phone has been on.
Battery gap	X	✓	There can be found data for other sensors, but not for the battery. There were no cases of this in the data.

Table 12: A description of how the battery data is used to mark missing data.

and results. Consider the case where the subject turns off his or her phone each night when going to bed. This would be seen as a huge gap that usually starts in the evening and ends the following morning. This could be classified as a normal behavioural data gap where the subject is probably staying in bed. Now consider the case where the smartphone is turned off during the day for several hours. This will cause a problem if the subject is active, for example, moves from one location to another or talks to people, as this will not be seen in the sensor data collected by the phone. This leaves out useful behavioural data, may skew the data and may force the day to be discarded from the analysis. Decisions that have to be considered are thus, how big gaps are allowed and are they allowed at anytime. These raise further questions such as, is acceptable to turn off the phone for the night and how do we know if a person is sleeping during the night.

For this thesis real gaps, where both battery and sensor data was missing, that discard data were chosen as 6 hours or longer. This was considered as a suitable time limit for the data and the methods used. For prediction of future PHQ-9 scores data gaps could be a bigger issue.

3.2.2 The selection of subjects

When choosing subjects for the analysis, the main criteria are that there is enough data over a time span and that there is not too much missing data during this time period. The participants in the studies were supposed to have the application installed on their phones for 6 months to one year. However, many of the participants kept it installed for a longer time.

For the methods used in this thesis, the subjects selected needed to have answered at least three PHQ-9 surveys. When analysing how the PHQ-9 answers have changed between each taken survey, it is optimal that there exist data for each day

fourteen days before the taken test. This is due to that the questionnaire asks the participant to reflect over the past two weeks. However, analysis of changes can be done, even though some days do not have sufficient amounts of data. A maximum gap between taken questionnaires was considered, but not seen necessary, as the the aim is analysing the behaviour connected to the PHQ-9 score.

To know if a subject has enough days of data, the days of battery data is first checked. A day that has a gap in battery data which is larger than six hours is discarded. This gaps in battery data occurs most often due to that the phone is turned off. An often occurring behaviour for turning of the phone could be turning the phone off for the night or turning the phone off when it is left in a locker. The biggest issue when discarding data due to too much missing battery data, is that not using the phone can be a behavioural pattern. Analysing this is however deemed out of scope of this thesis.

As mentioned in Section 3.2.1, the battery data can be used to find gaps in other data. However this can not be done for the location data, as it is given as an aggregation for one day. Due to this location data is simply seen as days with and without data. The Ambient noise sensor is activated every 30 minute, so if the recording is not performed for a gap of six hours the data is discarded. The other sensors create timestamps when an action is triggered, for example, screen is turned on, a call is answered or an application is being used. For these sensors, discarding is applied if the gaps in the data are larger than six hours.

In this thesis one of the aims was to analyse which sensors and features are good indicators for classifying and showing behavioural changes in mental health patients. This means that a participant that has, for example, sufficient screen data, but insufficient location data will be accepted when analysing only screen data. This applies for all sensor data that are not discarded.

3.2.3 Aggregation of data

Koota provides the data in the form of sqlite databases. The niimpy package contains functions that open the sqlite databases and do basic querying. It also has more high-level functions, such as basic preprocessing/aggregation, visualizing data quality, and other transformations. The rest is done by the researcher.

The sensor data used in this thesis includes applications, battery, communication, location, ambient noise and screen. First the high level niimpy functions will be described for the sensors that apply, and then the different aggregations are discussed.

The niimpy Python package processes individual-level data, meaning the code analyses one subject at a time. The high-level functions used for preprocessing in this thesis will be described bellow.

The app duration function in niimpy returns a table with the usage duration and amount of times an app group was used during a day. The apps are grouped as follows;

Method	Interval	Study data	Groups	Aggregation
Correlation between sensors and mood	2 weeks	pilot	compares correlation of individual subjects, patients, controls, all subjects	mean, median, standard deviation, minimum, maximum
k-means clustering	1 week	pilot	the method creates own clusters, where a subject is assigned to a group	mean, median, standard deviation
Linear discriminant analysis	monthly	pilot	classifies subject into patient or control	mean, median, standard deviation
kh-segmentation	daily for 2 weeks	pilot, (main)	individual subjects	sum of day

Table 13: Short summary of the methods and the preprocessing.

Sports, Games, Communication, Social Media, News, Travel, Shop, Entertainment, Work/study, Transportation, and Other.

As mentioned earlier in Section 3.1.2, the battery plays an important part in identifying missing data, and further selection of data and subjects. This is described in Section 3.2.1.

For communication niimpy gets the five most frequent contacts for the chosen period of time and calculates the duration and amount of times calls have been made. It also returns the duration used and count on different communication events.

The location converter preprocesses the data enough for analysis, see Section 3.1.2.

The noise function in niimpy returns a table with the daily values for the average decibels, average frequency, the number of times when there was noise in the day, number of times when there was a loud noise during the day (defined as higher than 70dB), and number of times when noise matched the speech noise level and frequency (where frequency is between 65Hz and 255Hz, and dB higher than 50).

The screen is grouped into transitions between the states on to off, off to on, off to in use, and irrelevant. The duration and number of these events during a day are returned by the preprocessing function in niimpy.

The data was aggregated into groups of individual subjects, controls, patients and all subjects. The data was also split into intervals of one day, fourteen days and all data for each subject. The data was further aggregated by calculating the mean, standard deviation, minimum value and maximum value for each subject. The aggregation differs depending on the method, as seen in table 13. The main methods were search for correlation between PHQ-9 score and sensors, k-means clustering, linear discriminant analysis, and kh-segmentation.

3.3 Searching for correlations between the PHQ-9 score and the sensor data

One of the main objectives of this thesis is to find quantifiable observations within the passive data, which can be linked with the health condition of the patients and then further used as features in training the machine learning model. This is a reason for why the correlations between the PHQ-9 questionnaire score and passive data was calculated. It gives an insight in selecting features for the machine learning part.

The Pearson product-moment correlation was first developed in 1895 [48]. Correlations can indicate a predictive relationship between variables [36], which is applicable to finding passive data patterns that decrease or increase the PHQ-9 score. The result of correlation is called a correlation coefficient and varies between -1 and 1, where a correlation coefficient close to -1 means that as one variable gets larger the other one gets smaller, a correlation coefficient close to 1 means that as one variable gets larger the other one gets larger and a correlation coefficient close to 0 means that there is no dependency between the variables. In the case of PHQ-9 score and passive data, the interesting ones are the negative or positive correlations, as they indicate a change in the PHQ-9 score of the subject due to change in the passive data. A change in passive data imply a change in behaviour.

Correlation can be calculated using different methods, resulting in different correlation coefficients. The Pearson correlation coefficient was used for calculating the correlation between the PHQ-9 score and the sensor data. The Pearson correlation coefficient measures the linear correlation between two variables. Similarly, it takes a value between -1 and 1. A correlation coefficient close to -1 indicates a positive linear correlation, a correlation coefficient close to 1 indicates a negative correlation and a correlation coefficient close to 0 means that there is no linear correlation.

$$r = \frac{N \sum XY - (\sum X \sum Y)}{\sqrt{[N \sum x^2 - (\sum x)^2][N \sum y^2 - (\sum y)^2]}} \tag{1}$$

The search for dependencies was done using the pilot study data. As mentioned in Section 2.2.1, the PHQ-9 asks questions about the past two weeks. Therefore, passive data from only the past two weeks is used to calculate the correlations. As previously discussed in Section 3.1.4, the first PHQ-9 questionnaire is answered in the beginning of the active phase, meaning that there is no passive data to connect with the first PHQ-9. This restricts the amount of subjects. There are four patients and eighteen controls that have answered enough PHQ-9 questionnaires. For these subjects the quality of the passive data also varies. If the subject has gotten the same score, for a question or the summed score, for all taken questionnaires, the correlation can not be calculated for the separate question or the summed score.

Since each sensor outputs different features of data, the correlations were calculated and analysed separately for each sensor.

The amount of timestamps in two weeks of temporal data varies from sensor to sensor and person to person. This big quantity of data is at a low conceptual level,

by generalizing the data it can be presented at a higher conceptual level and thereby be easier understood.

As the usage of a smartphone varies from person to person, it is not straight forward to compare changes in behaviour. An example of this could be that person A uses the smartphone for 3 hours per day on a normal basis, whilst person B only uses the smartphone 1 hour per day. Something changes in person B's behaviour and the smartphone usage increases to 3 hours per day. This is an increase of 300%, for person B, but normal for person A. As an attempt to better compare changes in behaviour to the changes in PHQ-9, a normalization of the data on subject level was computed.

First, the two weeks of data was aggregated for each day, calculating the average for each output of the sensor. This was normalized looking at all available 14 day intervals for that subject. After the daily aggregation, the mean, standard deviation, minimum and maximum is calculated for the mean of days, resulting in a summary of the two weeks. Calculating the mean of each day and then calculating the mean, median, standard deviation, max and min of the mean days, is a simple and in this case sufficient way of comparing each day to another. This method allows us to approximate without the problem of one day having more timestamps than the other. It is a sufficient approach as it is used for searching for possible features to be used in the classification.

Next, the summary is joined with the corresponding PHQ-9 score. This procedure is done for each PHQ-9 total score. Finally, correlations between the different variables can be calculated.

The correlation coefficients were calculated for all subjects as a group, for separate groups as patients and controls, and for each individual subject. The results are presented in Section 4.1.

3.4 Machine learning methods

Section about Machine Learning and the methods used in this thesis. The results are presented in Chapter 4.

This thesis aims to learn from data and make classifications based on this knowledge. Machine Learning follows a principle of learning from the data given to predict an outcome [12][33][52]. Machine learning algorithms are an effective way of modelling the complex structures of large data sets. The desired result, when using machine learning, is to learn the links between the data and the labels, and to create a model that is able to predict or classify based on the data given to the model [12][33][52]. This mapping of patterns in data to an output is known as pattern recognition, for which regression and classification are examples of in machine learning [12][33][52].

In machine learning literature, observations are often called instances or data points and the variables describing these data points are called features [33]. These features are grouped into feature vectors, consisting of the features that best describe the data points. The chosen machine learning method depends on the dataset, for example, is the dataset small or big, labeled or unlabeled.

Machine Learning problems are often divided into four areas; classification, regression, clustering and dimensionality reduction [12][33][52]. In classification the task is to determine to which category an instance belongs to, for example, is the animal a dog or a cat [12][33][52]. The predicted categories are called classes. Regression is used for predicting and forecasting values of a quantitative type, for example, the hours spent working on different tasks of a course may have a relationship with the course grade. In clustering, the problem is to divide the datapoints into a chosen number of groups where the groups consist of similar datapoints [12][33][52]. The group of similar datapoints is called a cluster. In dimensionality reducation, the task is to reduce the number of variables in an observation to the variables that explain the observations the best, finding the best feature vectors [12][33][52]. Dimensionality reduction consists of feature selection and feature extraction.

There are three basic machine learning paradigms; supervised learning, unsupervised learning and reinforcement learning [12][33][52]. In supervised training the input data and output data is labeled. This means that we can try to create a mapping function between input and output data, that could also be generalized for unlabeled input. Classification and regression are typical problems solved with supervised learning. In unsupervised training the data is unlabeled and the data is instead explored for patterns to learn from. Examples of unsupervised machine learning methods are clustering analysis and principal component analysis. These methods extract useful features from the unlabeled data and build a model that describes the data structure. An example of clustering analysis is K-means clustering which was used in this thesis, see Section 3.4.1. Reinforcement learning does not need labeled input and output. It uses the knowledge it has to update the model and searches for more knowledge by further exploration. No reinforcement methods are used in this thesis.

One of the often occurring problems in machine learning is overfitting. Overfitting means that a trained model performs well on the original training data, but the performance drops when the model is presented new data [12][33][52]. The model does not generalize. An overfit model has learned specific data irregularities from the training data and basing predictions on specific oddities does not perform well. This can be prevented by having only a few features and by regularizing the data. The opposite problem to overfitting is underfitting. In underfitting the model is too simple and has a low variance [12][33][52]. Underfitted models tend to predict the wrong outcome, by having a bias towards some predictions. Identifying overfitting can be done by splitting the data into a training set and a test set. The model is built with the training set and the test set is used to test the real accuracy of the model. Cross-validation can be used for tuning the model, without peeking at the test set. In cross-validation the training set is further split into training and test sets, where iteratively one split is used as test set and the rest is the training data. Another way of reducing overfitting is more data. More data often prevents overfitting due to that the training and test set becomes bigger, making it more likely to find a better fit. Removing features also helps preventing overfitting [41]. The rule of thumb for features versus samples is 1 to 10. When reducing the features, it is important to remove redundant and irrelevant features. Redundant features are,

for example, features that overlap with each-other. For further feature reduction, the earlier mentioned dimensionality reduction methods can be used, either feature selection or feature extraction.

Deep learning is a type of neural network consisting of multiple layers between the input and output layer [54]. These layers calculate the probabilities of outputs when given an input. The more layers a deep neural network (DNN) has, the deeper it is. DNNs have become popular machine learning methods as they can model complex non-linear relationships, however, it is not suitable for all kinds of modelling. Deep learning often demands a big evenly distributed dataset, so that it does not overfit. Deep learning also demands higher computational power, for each added layer. Disadvantages with deep learning are that the resulting model is complex and not easily interpreted, and information security breaches can also be introduced, for example, via a model that leaks information about the training dataset [58].

Choosing a machine learning method is not simple. The selected model has to be supported by the problem and, the quantity and quality of the data. Further, the purpose of the model has to be well defined, as well as, the criteria to measure its performance. Depending on the purpose of the model, a prediction error may be crucial. An example of a devastating prediction would be not finding a cancer tumor when a person has cancer.

In [31] it is stated that linear regression is not appropriate in the case of a qualitative response. If the response variable's values did take on a natural ordering, such as mild, moderate, and severe, and if the gap between mild and moderate was similar to the gap between moderate and severe, then a 1, 2, 3 coding would be reasonable. Unfortunately, in general there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression. [31]. The PHQ-9 score could be seen as a scale between mild-major depression. However the scale is only indicative and it can be problematic to transform it to a quantitative response.

In this thesis the main problem is to classify subjects into patients and controls. The data points used in the learning are smartphone data and PHQ-9 answers. The data is labeled, it is known if the subject is a patient or a control. The biggest restriction in the data is that the sample size is small versus the number of features available.

3.4.1 k-means clustering

To analyse the subjects and their associated data the k-means clustering method was used. k-means clustering finds clusters in unlabeled data [26]. In data labels, patient and control, are removed and an exploration of the data is done by trying to separate the unlabeled data into groups. These groups are called clusters. The clusters are made up by close-by data points, meaning the data points with the least distance to the cluster center are part of the same cluster. The distance to the cluster center can be calculated using different definitions, but a popular definition is the Euclidean distance.

k-means is a hard clustering method, where the amount of clusters is k and a cluster center is represented by the cluster mean [26]. The number of clusters can be two or up to the amount of data points. A data point is appointed to the cluster with the shortest distance from the cluster center to the data point. After each appointed data point, the cluster center is updated. When all data points are sorted to a cluster, the algorithm has finished. It should be noted that, the end result is not always the same. The end result depends on the sorting order of the data points, meaning that some data points may be sorted to another cluster depending on what the cluster center is at the time of the sorting. The end result of the clustering can be evaluated using a criterion called inertia. Inertia is the sum of squared distances of data points to their closest cluster mean. By minimizing the inertia the clustering error can lowered.

An optimal result would classify each data point to the correct cluster/group, but for smartphone data there will probably be some overlap and miss-classification. The method will still show how close the two groups data is to each other and more importantly if some features overlap so much that they should be excluded.

The scikit-learn clustering package function KMeans [49], was used for the k-means clustering. The function clusters data by separating the samples into groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. The function uses either Lloyd's or Elkan's algorithm.

k-means clustering was used to see how well the different sensors clustered the subjects according to the labels, patients and controls, and to see which subjects resemble each other, which is seen by them clustering together. Similar smartphone behaviour for subjects group them together. The results are presented in Section 4.2.

3.4.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) finds a linear combination of features which can be used to separate two or more classes from each other. The found linear combination can be used for classification, either as a linear classifier or for dimensionality reduction before classification [39]. In this thesis the latter is used in combination with decision tree classification, which is introduced in Section 3.4.3.

LDA is a supervised feature reduction technique which reduces the dimensionality of the feature space [57]. Principal component analysis (PCA) is a similar technique, however, it is unsupervised and does not perform as well as LDA in classification tasks [57]. LDA is a suitable method if there are two or more classes. LDA can be used to overcome small sample size problems [57][31] and is therefore feasible for the pilot study data set.

LDA finds a matrix W that transforms the feature vectors to a lower dimensional feature space in a way that maximizes the separation between the classes [31]. The highest possible dimensionality is the number of classes reduced by one. In the case of two classes, patients and controls, the dimensionality is reduced to one [31]. This is seen as a strong reduction. It should also be pointed out that if the classes are well-separated, then logistic regression models become unstable, whereas, LDA does not run into the same problem [31].

In [26] it is said that, for LDA the parameters of the distribution are not known, so they are estimated using the training data. They further say that, LDA tries to fit decision boundaries on the data by minimizing the training error. As this method is linear, visualized the boundaries would be just straight lines dividing the data optimally. However quadratic decision boundaries can be found using LDA in a higher -dimensional space. Nearly similar results can be obtained using Quadratic discriminant functions (QDA).

In implementing LDA in this thesis, the LDA components are used as features in a decision tree. The LinearDiscrminantAnalysis classifier in the scikit-learn package was used [49]. The class conditional densities are fitted to the data and the Bayes' rules are used to create the classifier. The classifier fits for each class a Gaussian density. The classifier can be used to classify or for dimensionality reduction before classification. As stated before, the model is used for dimensionality reduction and then transformed data is used to train the decision tree, which does the predictions, see Section 3.4.3.

As LDA is suitable for more than two classes, it can be used for future research topics for classifying between different cohorts.

3.4.3 Decision Tree Classification

Decision tree learning is a predictive model in the form of a decision tree [50][12][33][52]. The tree's branches leads from where an internal nodes divides based on an observation to the next node and the leaves are the terminal nodes showing the resulting prediction, when following the different branches of observations [50][12][33][52]. It should be pointed that the tree is often visualised upside-down, with the root starting in the top of the image and the leaves ending on the bottom of the image, see example figure 2. A tree that predicts a value in a discrete set of values, for example a class, is called a classification tree. In this thesis the leaves are either patient or control and the observations are the smartphone data features.

The biggest advantages with decision trees are that they are easy to interpret, display graphically and the decisions taken by the tree are easy to explain [50][12][33][52]. On the other hand a tree model is not robust and the model will change as the training data changes, meaning the branches and the decisions will look different.

A decision trees branches are built by splitting the predictor space. These splits produce regions where certain observations are true. This means that before the first split all observations are part of the same region. For each split of the predictor space two branches occur, meaning one region is split into two regions.

More splits ends in better accuracy for the training set, which often means overfitting and bad accuracy for the test set. This can be prevented by having a smaller tree with fewer branches.

For a classification tree, a split is based on the most common class in a region [50][12][33][52]. An internal node can hold different classes, but at the terminal node only one class. The classification error rate can be used as a split criterion, but it is often not sensitive enough. Two commonly used criteria are the Gini Index and entropy. The Gini Index is the default for scikitlearn's DecisionTreeClassifier and

is used in the implementation of the machine learning model. The Gini index is a measure of total variance across the K classes. The Gini Index is also to describe or denote a node as pure or impure. A pure node contains only observations from a single class, whereas a impure node contains observations from more than one class.

As mentioned before in Section 3.4.2, the decision tree implemented is trained with the transformed data from the LDA dimensionality reduction. The results of the LDA and decision tree classification results are presented in Section 5.3.

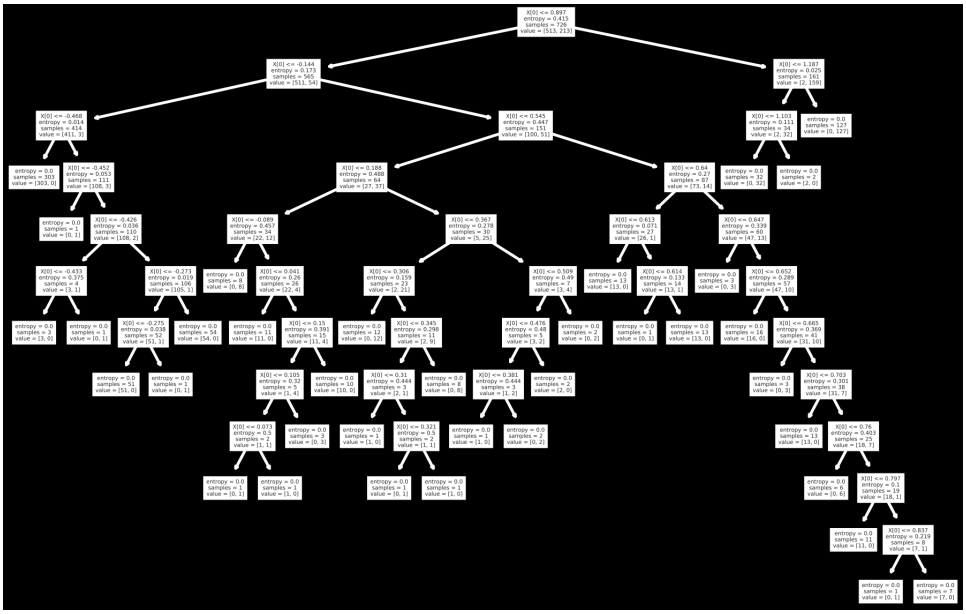

Figure 2: An example of a decision tree. The number of branches and leaves should be adapted according to the dataset and feature size.

3.4.4 Validation of the classifier

As a way to calculate the accuracy for the classification models an implementation using cross-validation was used. In cross-validation the partition of samples into the training and test set is done several times, resulting in different subsets each round [11][60][61]. This will give an average of the prediction accuracy for the different models [56].

The cross-validation was implemented by randomly picking controls and patients to the training and test sets. As each subjects data was split into smaller lengths depending to determine the best sequence lengths, it is important to not mix a subjects data between the different sets. When looking at the the different models accuracies separately one can see which models may have problems with overfitting or selection bias [21].

For the data used in this study there were eight patients and eighteen controls. When applying cross-validation on this dataset, it would result in subgroups with, for example, a training set containing 6 patients and 13 controls and a test set of

2 patients and 5 controls. If run with all different combinations, this would result in 28 different combinations of patient subgroups and 8568 control groups. One could further try to balance the training set by choosing equal amount of controls as patients. This would change the choosing of controls and result in 18564 different combinations of control groups. In this thesis the cross-validation was done by varying the patient and control size in the training and test set. With the patient set size ranging from two to five and control set size ranging from four to ten.

3.4.5 kh-segmentation

In time-series analysis a practical way of representing the data is to make it compressed and concise, and present it efficiently and understandably [63].

One interesting way of analysing sequence data is using kh-segmentation. The kh-segmentation is a modification to the original segmentation problem, where the optimal segmentation with k-segments can be found with a dynamic program [24][16]. The problem is to segment an n-element sequence into k-segments, where each segment is homogeneous with regards to an error measure. An example of a sequence is the heart rate during the the day. A persons heart rate usually has a resting state and changes in the heart rate due to, for example, physical exercise, sleep or stress. In this case the segments would consist of the these different heart rate zones. For example, the the sequence begins with sleep and changes to an awake resting state, here sleep and awake are different segments.

In kh-segmentation a sequence is divided into segments and these segments are further distributed into states, that are related to each other and coming from a number of sources. The problem of finding the best ways to segment an n-element sequence into k-segments, and further deciding the different h-sources they stem from, has been presented in [24]. Continuing with the heart rate example, the h-sources could represent the different heart rate zones, where the k-segments that are the same are marked as the same, for example, the resting states are marked as the same h and sleep is marked as sleep. If h is set equal to k, then it is the same as the k-segmentation. The amount of sources is the same as the amount of segments.

By choosing different k and h the level of granularity changes. A coarse low-level granularity does not see the small features and tries to look at a bigger picture, whereas the fine high-level granularity describes the small features and details. For the heart rate example, coarse granularity would be the difference between being awake and asleep, whilst fine granularity would even notice the different kinds of exercise heart rate training zones recovery, aerobic, anaerobic.

The problem gets even more dimensions when extending it to a multivariate setting, meaning adding more variables than just one sequences of n-elements. This gives us the possibility to look at several features when segmenting, not only one. For the heart rate example, it could mean adding an actigraph that measures the movement of the person.

When choosing k and h, the number of sources does not have to be the correct amount of states, more sources does not lower the accuracy and less sources does not lower it that much [24]. This is good if one does not want to have a set amount

of sources, for example, the states asleep and awake. Using a more exploring way of setting the amount of sources makes it possible to discover hidden states that one had not thought about. How to find the most optimal k and h is explored using two different criteria explained later in this section.

The problem of finding a good way of segmenting an n-element sequence into k-segments, with h different sources has thankfully different solutions, [24][37][19][63][28]. In this thesis the solution and code presented in the paper [24] was used and modified. The paper, presents three approximation algorithms; Segments2Levels, ClusterSegments and Iterative. In this thesis the ClusterSegments algorithm was modified for use.

Algorithm ClusterSegments: The algorithm solves the k-segmentation problem ending in a segmentation S. Each segment is a mean of the elements. Finally,the k segments are clustered into h clusters.

The paper provides a score for the chosen k and h values by applying the Baysian information criterion (BIC)[55][37]. A lower BIC value means a better segmentation.

In the case of mental health and passive smartphone data, the relationship between the segmentation of the passive data and the changes in the PHQ-9 score is interesting. The hypothesis is that if there are changes in the PHQ-9 score there should be changes in the behaviour, which further should show as segmentation in the passive data.

As mentioned, the PHQ-9 asks the subject questions about the last 14 days. The data for each feature for the last 14 days is thereby acquired, normalized and aggregated for each day. This is done for each PHQ-9 that a subject has submitted. These sequences are combined one after another.

First, the segmentation behaviour was explored through using different k and h.

Secondly, the optimal values for k and h were searched for by maximizing the correlation between the 14 days of data and the PHQ-9 score.

Thirdly, the optimal values for k or h were searched for by maximizing the correlation between the 14 days of data and the PHQ-9 score.

Lastly, the optimal value for k and h were searched for by minimizing the BIC value.

The results using the different criteria were verified by calculating the correlation between the different segmentations and the PHQ-9 score. The different correlations were compared with correlation between the average for a sequence, where k is set so the length of the sequence, and the PHQ-9 score. If the correlation is higher it is seen as better than average.

The results for each experiment can found in Section 4.4.

4 Results

This chapter presents and evaluates the results. The last section, Section 5, is dedicated to discussing the results and drawing conclusions.

4.1 Correlation results

The correlations between the sensor features and PHQ-9 scores were calculated separately for each sensor as shown in Section 3.3.

The PHQ-9 score reminisces the past two weeks, therefor the past two weeks of data from answering the PHQ-9 questionnaire is interesting. The data was normalized for each subject so that they could be compared to each other. First, the two weeks of data was aggregated for each day. Second, the mean, standard deviation, minimum and maximum was calculated for the two weeks, resulting in a summary of the two weeks. Next, the summary was joined with the PHQ-9 score. This procedure was done for each PHQ-9 score available for a subject. Finally, correlations between the different variables was calculated.

First, the correlation was calculated looking at all subjects and the total PHQ-9 score. Looking at the correlations shown in figures 3, 4 and 5 only slight correlation could be seen. Since the patients and controls are expected to behave differently, reasonably the correlations can also be different for the patient and control groups. It also to be noted that there are more controls than patients, making the controls affect the correlation results for all subjects more. Thus, the correlations were also calculated separately for each group.

Second, the correlation between patient sensor data and PHQ-9 score was calculated. The results show correlation between changes in the sensor data and changes in the PHQ-9 score. This is seen in figures 6, 7 and 8. For example, for noise data the correlation between the mean of noise and the PHQ-9 has a negative correlation, indicating that high mean noise correlates with a lower PHQ-9 score. This could further indicate that being surrounded by noise is a sign of improved mood. For example, leaving a quite apartment or being surrounded by people.

Looking at the screen data correlation results in figure 7 there is a positive correlation between the duration of having the phone screen off and the PHQ-9 score. Meaning that having the phone screen off for a shorter duration correlates with a lower PHQ-9 score. Meanwhile a high use count has a negative correlation with the PHQ-9 score. This would indicate that having a high use count correlates with a lower PHQ-9 score. Activating the phone screen and keeping it on would hence indicate an improved mood among patients. This is similar to the results in this study [20]

According to the same study, [20], it was expected that there would be correlation

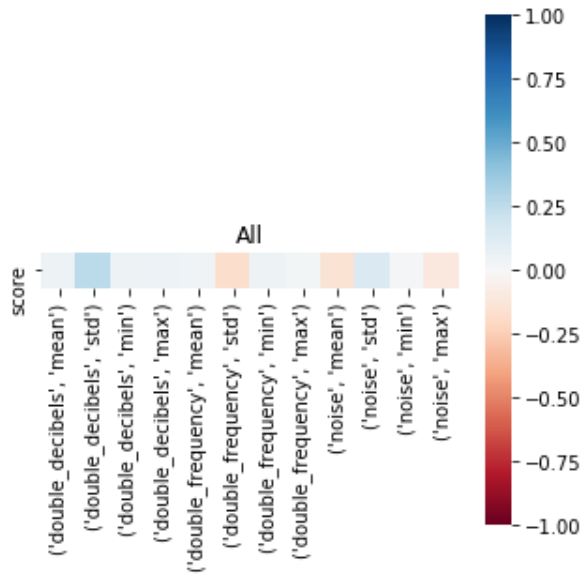

Figure 3: Correlation between PHQ-9 score and noise data for all subjects.

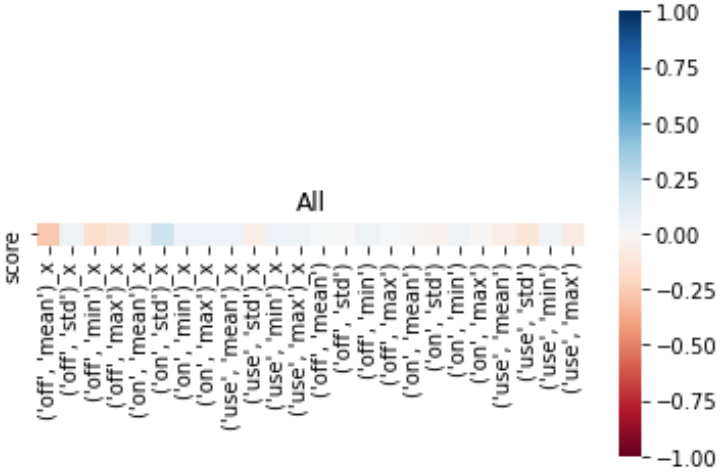

Figure 4: Correlation between PHQ-9 score and screen data for all subjects.

seen between the location data and the PHQ-9 score. There was however nothing marked seen in figure 8.

Last the correlation was calculated between control sensory data and PHQ-9 score. Compared to the patients, the control groups PHQ-9 score does not correlate that much with the different sensor data. This is seen in figures 9, 10 and 11. This can however be seen as expected, the control groups mood is not expected to vary much and it is not expected to change based on behavioural changes in the data gathered from the smartphone.

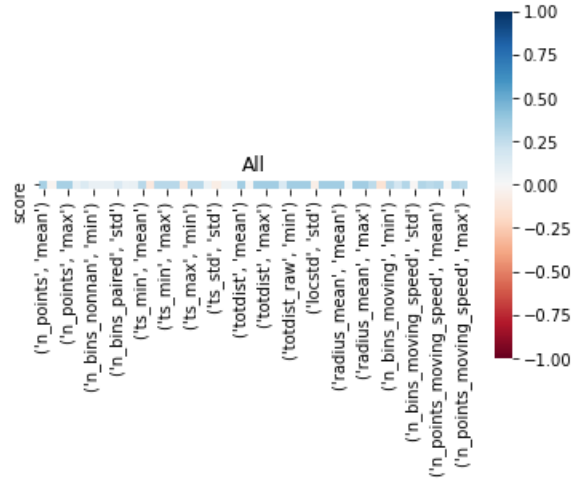

Figure 5: Correlation between PHQ-9 score and location data for all subjects.

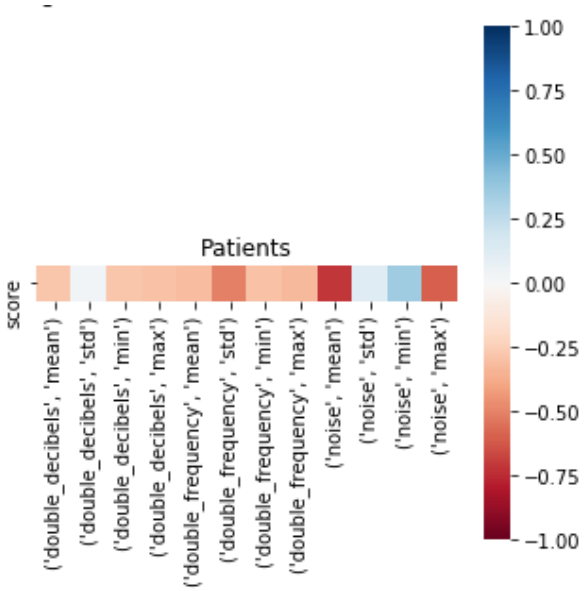

Figure 6: Correlation between PHQ-9 score and noise data for patients.

Considering that a difference could be seen in the correlations between the groups, it is further interesting to look at the individual correlations. When looking at the individual correlations, it is interesting to see that all subjects show a heterogeneous pattern while comparing them to each other. It seems like the subjects belonging to the control group, are more heterogeneous compared to the subjects belonging to the patient group.

Figures 13, 14, 15 and 16, show dark areas which indicate when there is no change either in the PHQ-9 score or in the sensor data. When there is no change the

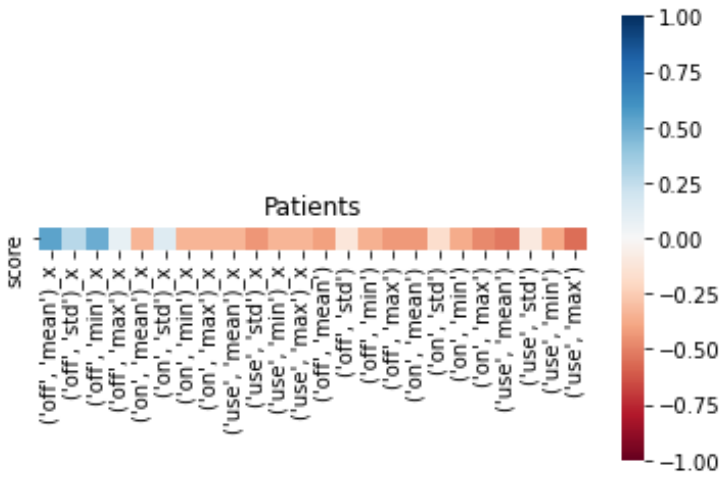

Figure 7: Correlation between PHQ-9 score and screen data for patients. The x indicates the duration and the data without x is count.

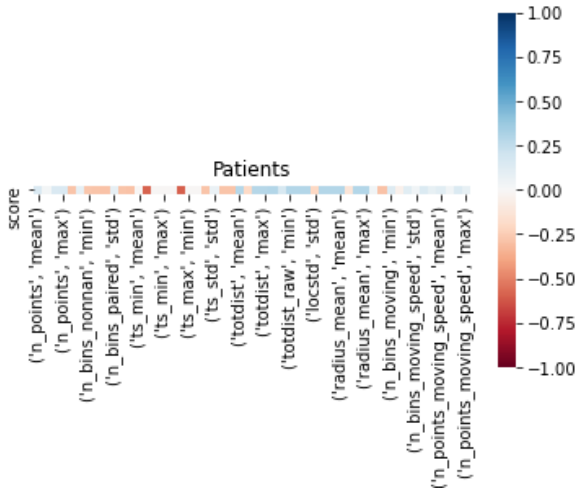

Figure 8: Correlation between PHQ-9 score and location data for patients.

correlation can not be calculated. It is good to remember that this also affects the correlations when calculating for groups or all subjects.

Due to the small sample size, it was not valid to calculate the correlation for the separate PHQ-9 questions. This could however be interesting to do when a bigger dataset is available.

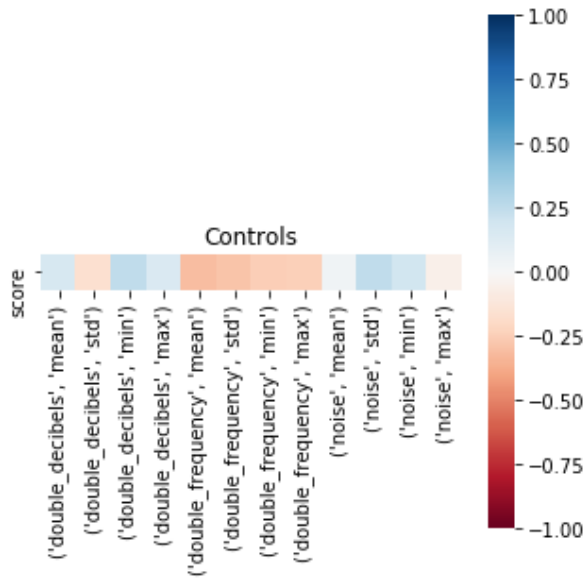

Figure 9: Correlation between PHQ-9 score and noise data for controls.

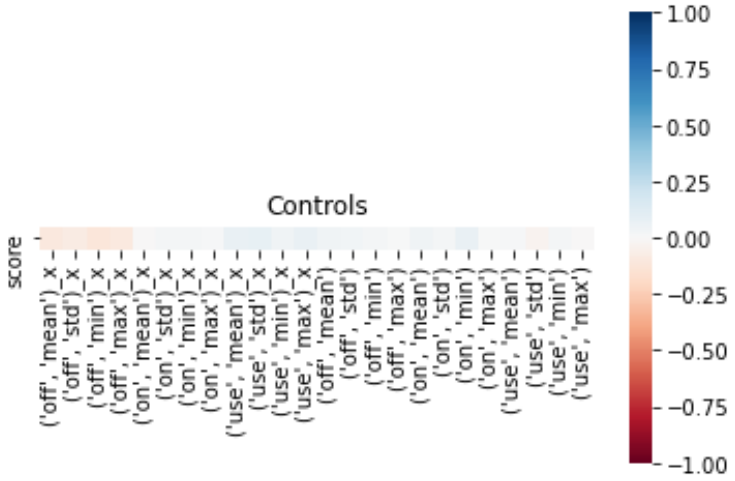

Figure 10: Correlation between PHQ-9 score and screen data for controls.

4.2 k-means clustering results

The results and conclusions of the the k-means clustering. The k-means clustering was calculated as shown in Section 3.4.1. The samples were clustered into 2 groups to see if the clustering would be similar to the actual labels.

For the noise data it seems like there are zero subjects in group 1, see figure 18. However, when looking at the pair plots in appendix A three subjects are seen in the other group.

It is seen in figure 20 that the screen data is more grouped. When comparing the

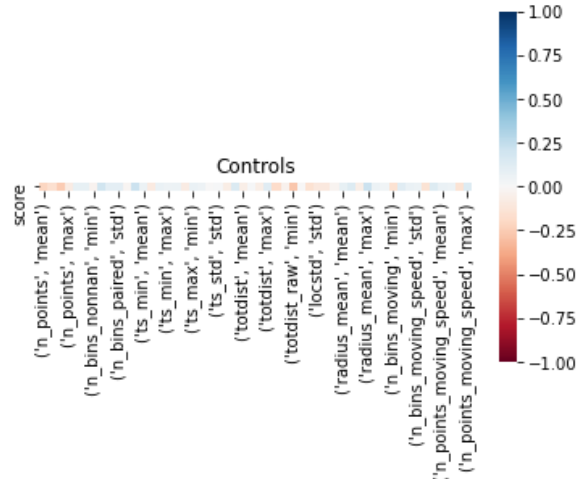

Figure 11: Correlation between PHQ-9 score and location data for controls.

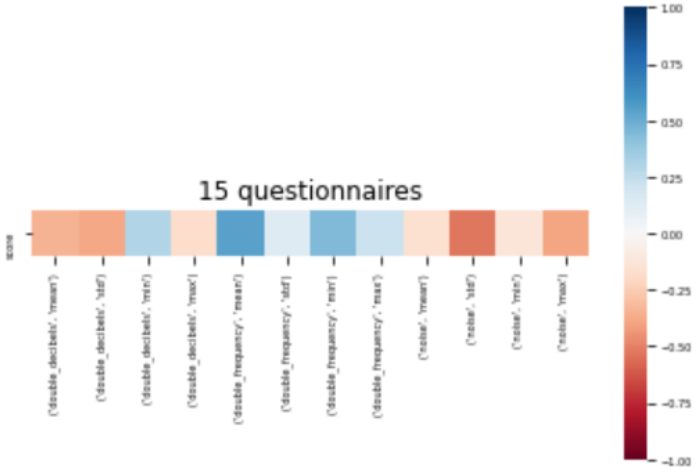

Figure 12: Correlation between PHQ-9 score and noise data for control with 15 answered questionnaires.

k-means clustering to the actual labels in figure 21 it is however seen that there is false classifications.

The kmeans results for location has sorted the subjects into two clear groups, see 22. However looking at the original labels in figure 23 the control is more spread and the patients are found clustered together in a small area.

When comparing the kmeans groups and the original labeling of the communication data, see figures 24 and 25, they seem quite similar. It is interesting to see that the cluster centers are very close to each other. This means that the clusters are close to each other and indicate that the groups are similar.

The social application data is grouped quite well. It does however not cluster the control group outliers into the same group. As a test the amount of groups was

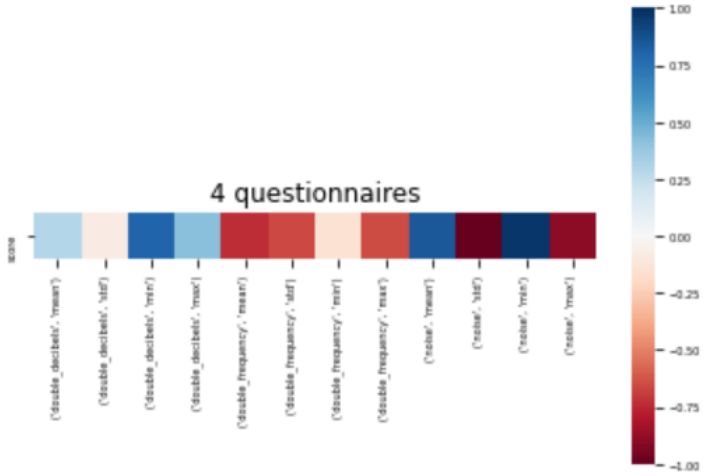

Figure 13: Correlation between PHQ-9 score and noise data for patient with 4 answered questionnaires.

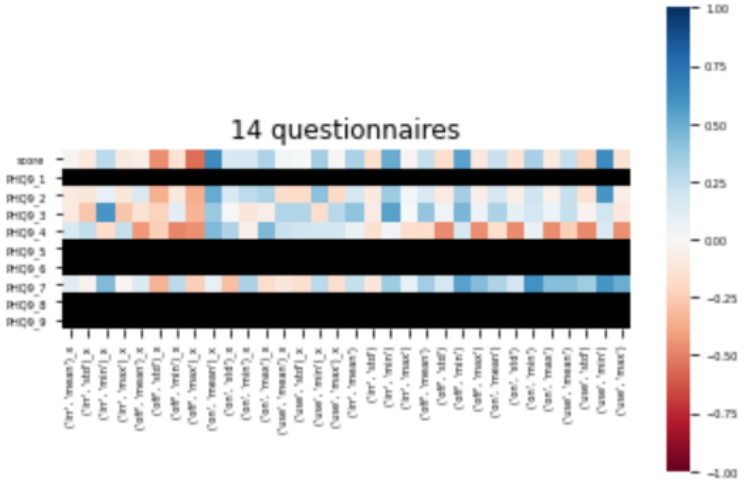

Figure 14: Correlation between PHQ-9 score and screen data for control with 14 answered questionnaires.

increased to three, see figure 28. The third group consists of the control outliers. Increasing the amount of groups seems to find the controls and place them in a separate group. However increasing the amount of groups does not add value, as the comparison is only possible due to that the labels are known. One could argue that the patient group has similar behavioural patterns whereas the control group can be divided into smaller groups with different behavioural patterns.

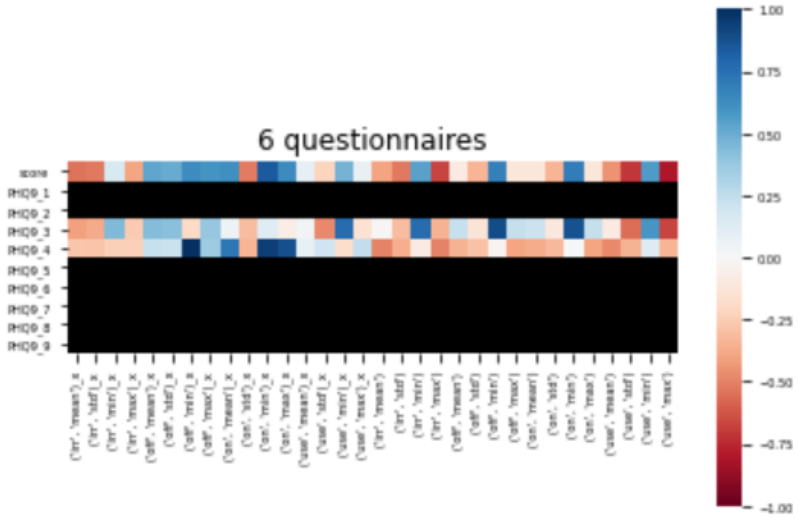

Figure 15: Correlation between PHQ-9 score and screen data for control with 6 answered questionnaires.

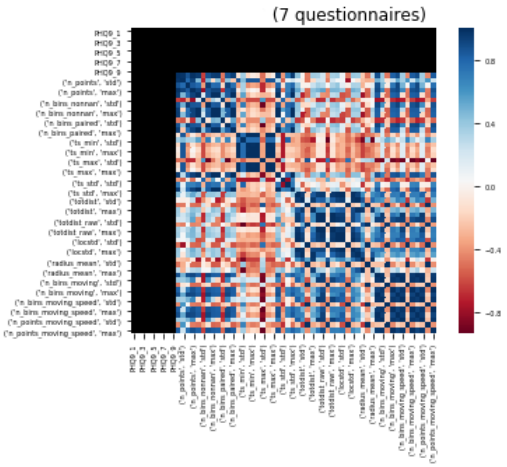

Figure 16: Correlation between PHQ-9 score and location data for control with 7 answered questionnaires.

4.3 Linear Discriminant Analysis and Decision Tree Classification results

Results and conclusions for Linear Discriminant Analysis (LDA) and decision tree classification. The LDA and decision tree classification models were created as described in Sections 3.4.2 and 3.4.3. The validation of the models is described in Section 3.4.4.

Different models were trained for each sensor and one model was trained for all data combined. All models were trained by changing the interval of days, and by

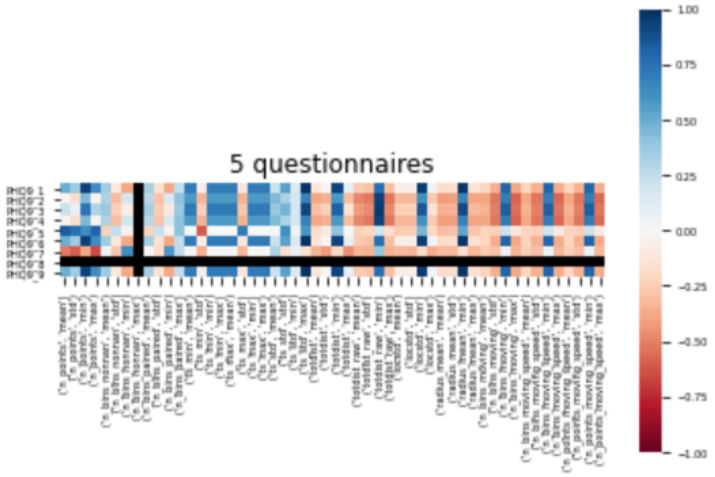

Figure 17: Correlation between PHQ-9 score and location data for patient with 5 answered questionnaires.

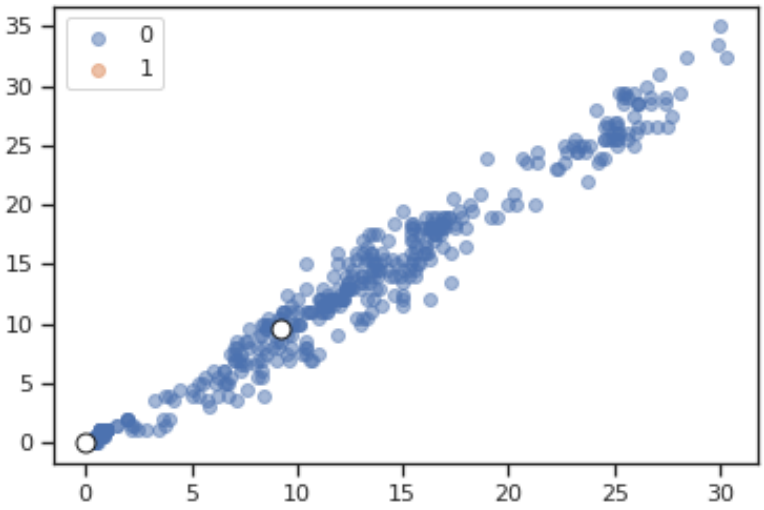

Figure 18: k-means clustering of noise data. Three subjects have been clustered close to the lower cluster center.

rotating the subjects used in the training and testing. The data was split into pieces of 7,14,21,28,30 and 35 days. When looking into the differences in the models, most models were quite consistent, with a good cross-validation accuracy.

Bellow, first the LDA results before the decision tree and then the classification results will be presented for each model.

As the classification is between two classes the scaling matrix produced by the LDA will be one-dimensional. LDA tries to separate the two groups from each other

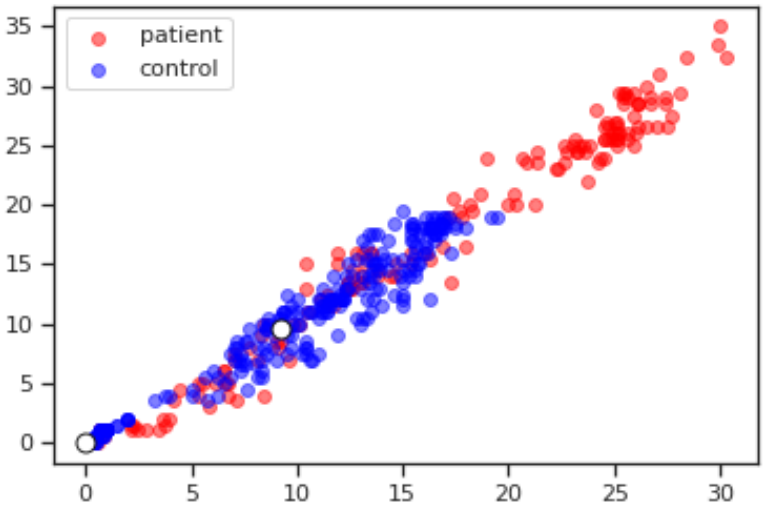

Figure 19: Visualisation of the real labels on k-means clustering of noise data.

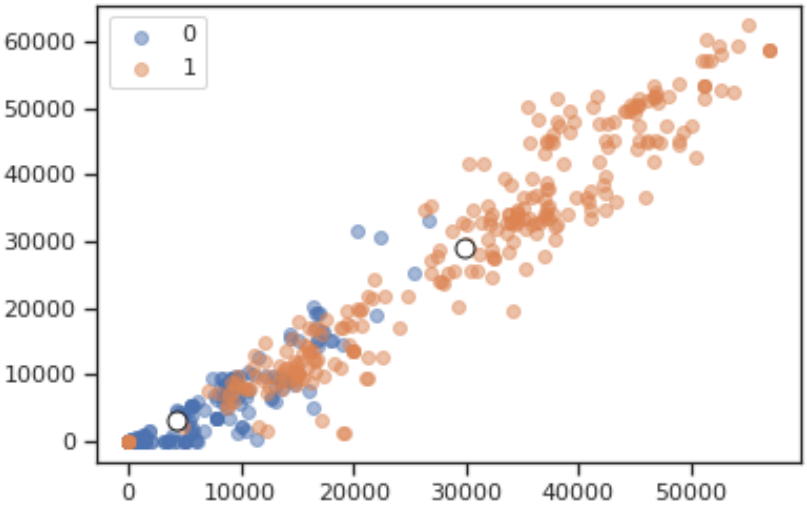

Figure 20: k-means clustering of screen data.

as well as possible. To see this separation, the fitted and transformed data is plotted to see how well the groups can be distinguished from each other. As a reference the battery data has undergone the same process. The expectation is that the battery data should not be different between the groups. In figure 29 no separation between the groups is seen. However there is one outlier which is interesting.

As mentioned in Section 3.4.2, PCA is a similar technique to LDA, but is not

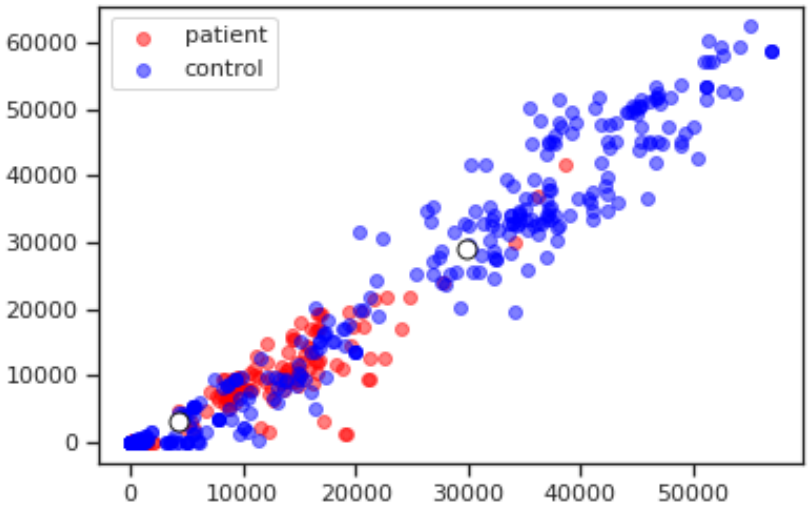

Figure 21: Visualisation of the real labels on k-means clustering of screen data.

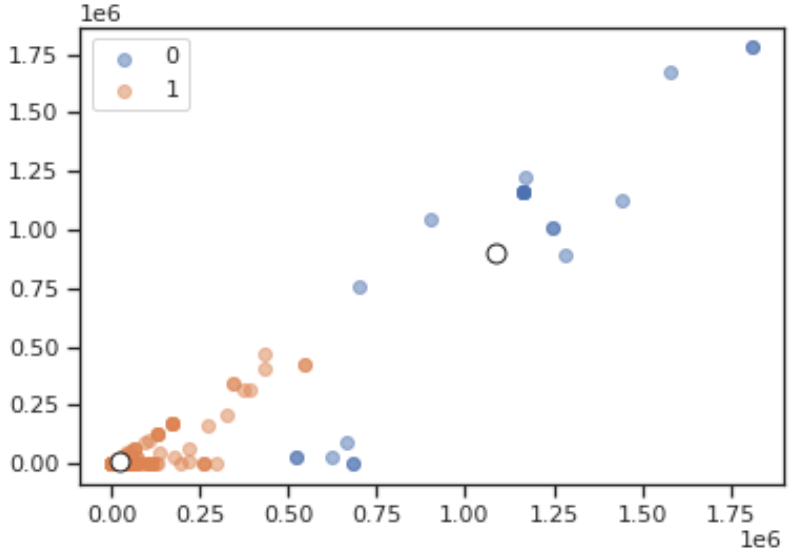

Figure 22: k-means clustering of location data.

expected to work well in classification tasks. PCA is performed so that it can be compared with LDA. For this battery is again a reference. It is seen in figure 30 that PCA is not able to separate the data points into groups. The same outlier is seen again.

The noise plugin measures ambient noise captured from the phones surroundings. This gives an idea of how much noise the subject is exposed to during the day.

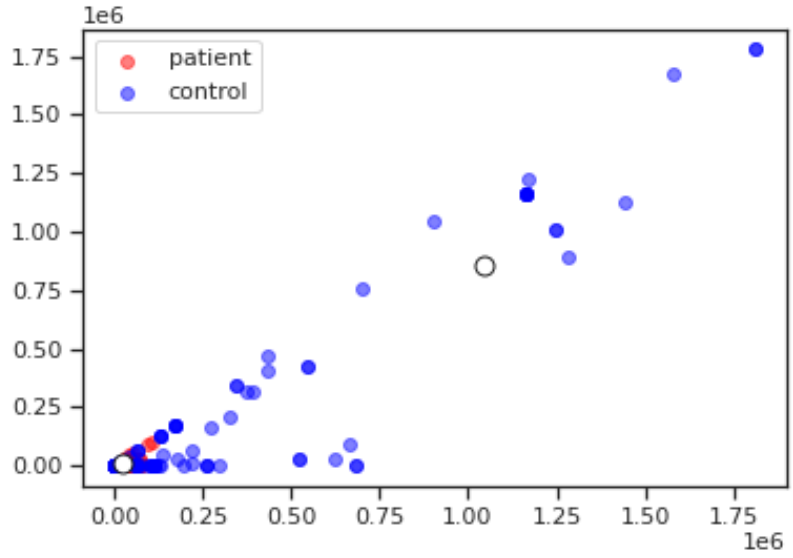

Figure 23: Visualisation of the real labels on k-means clustering of location data.

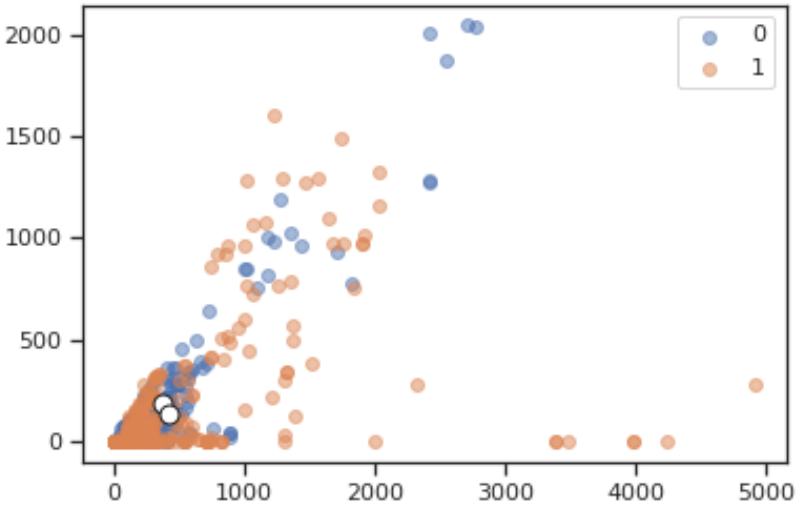

Figure 24: k-means clustering of communication data.

Performing LDA on the noise data gives the results presented in figure 31. It is seen that the separation between the groups is quite good. However there seems to be some overlap. Looking at the PCA of the noise data no distinct groups can be made out of the data points in the plot.

When looking at the LDA component weights for noise it seems that decibel mean, decibel median, decibel std, noise mean, noise median and noise std are the

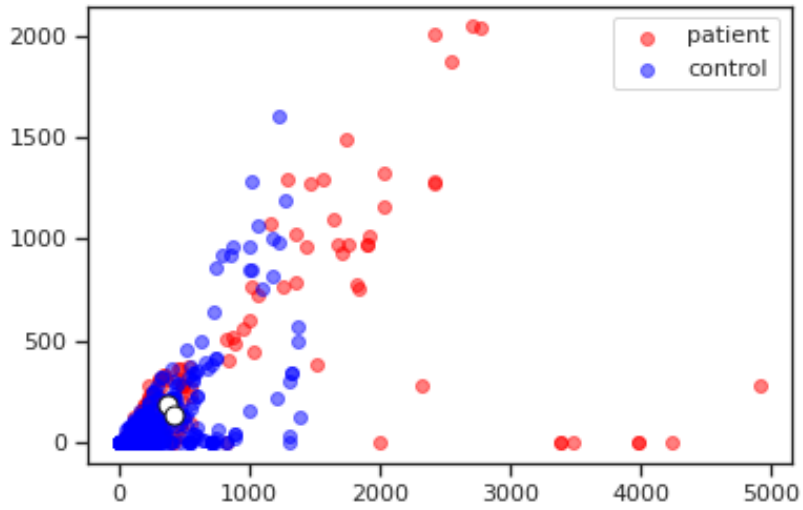

Figure 25: Visualisation of the real labels on k-means clustering of communication data.

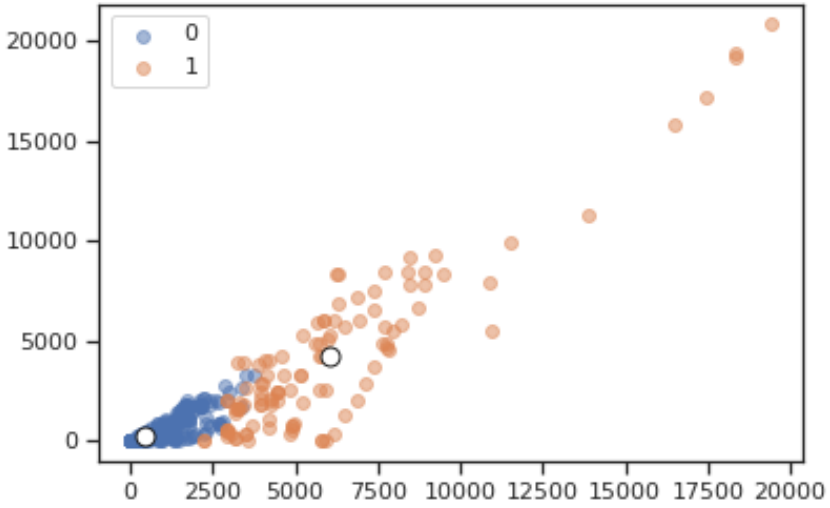

Figure 26: k-means clustering of social data.

most important features. Minimum and maximum does not provide added value, so they can be dropped.

For noise a classification was performed using the LDA model. The results are seen in figure 33. It is seen that cross-validation accuracy is lower for fewer day splits at a mean under 0.871 and at its highest with more days in each split with an mean

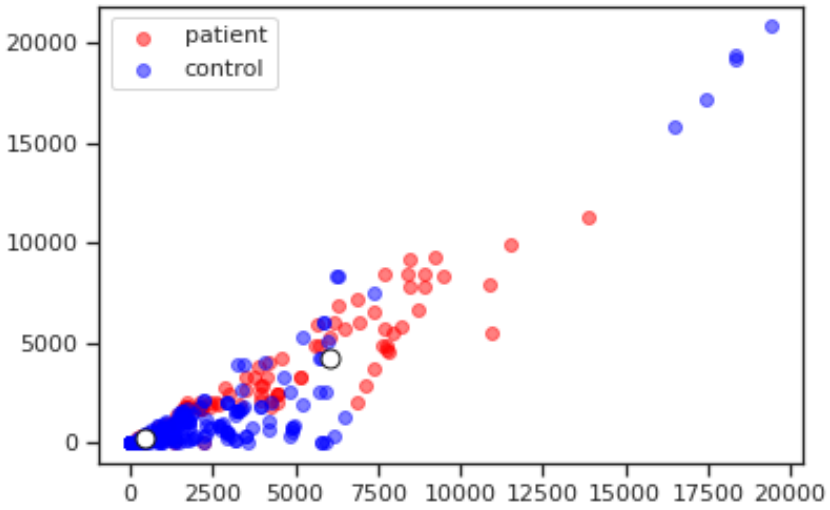

Figure 27: Visualisation of the real labels on k-means clustering of social data.

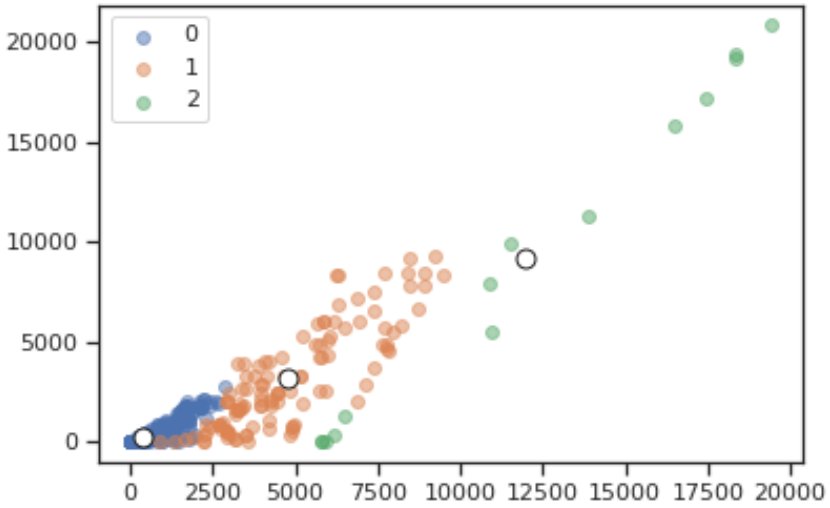

Figure 28: k-means clustering of social data into three groups.

accuracy over 0.876.

The LDA results are used to train the decision tree classifier. As an example, the representation of the tree created for noise data can be seen in figure 34. It is seen in figure 33, that using the decision tree classifier gave higher accuracy results compared to the only LDA model in figure 33. The cross-validation accuracy is ranging from approximately 0.965 to 0.969, with lower accuracy for fewer day splits and higher

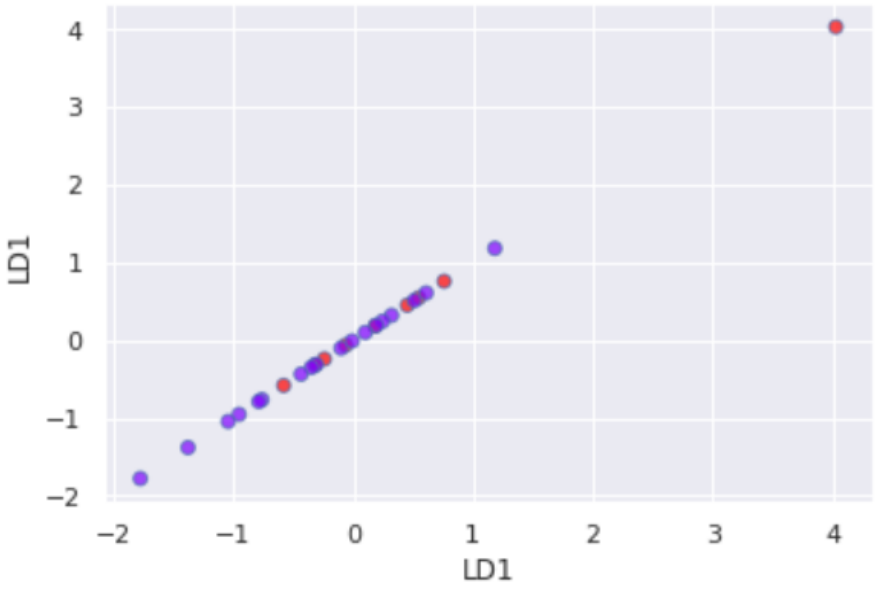

Figure 29: Visualisation of how well LDA was able to separate battery data. The LDA calculated for battery data should be seen as a reference, as it should not be well separated.

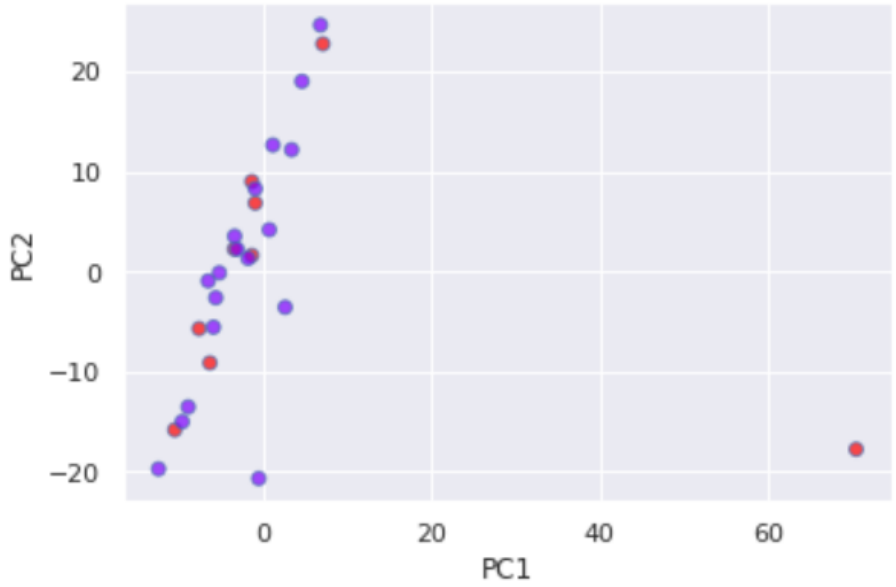

Figure 30: PCA calculated on battery data. The PCA calculated for battery data should be seen as a reference, as it should not be well separated.

accuracy for high amount of day splits.

Performing LDA reduction on screen data gave the separation shown in figure 36.

Figure 31: Visualisation of how well LDA was able to separate noise data.

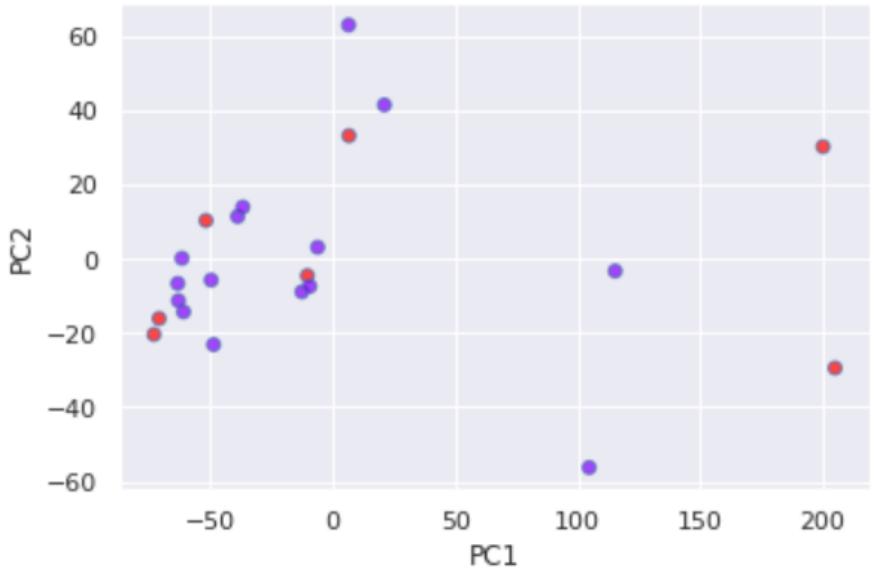

Figure 32: PCA calculated on noise data.

There is only one control who overlaps to the patient side, otherwise the groups are well separated. Comparing it to the PCA in figure 37, it is seen that no distinguishable groups are found in PCA. The decision tree classifier gives the results shown in figure 38. The accuracy is high, with a range from a bit over 0.962 and under 0.950. It is interesting to see that the accuracy drops for splits that contain more days.

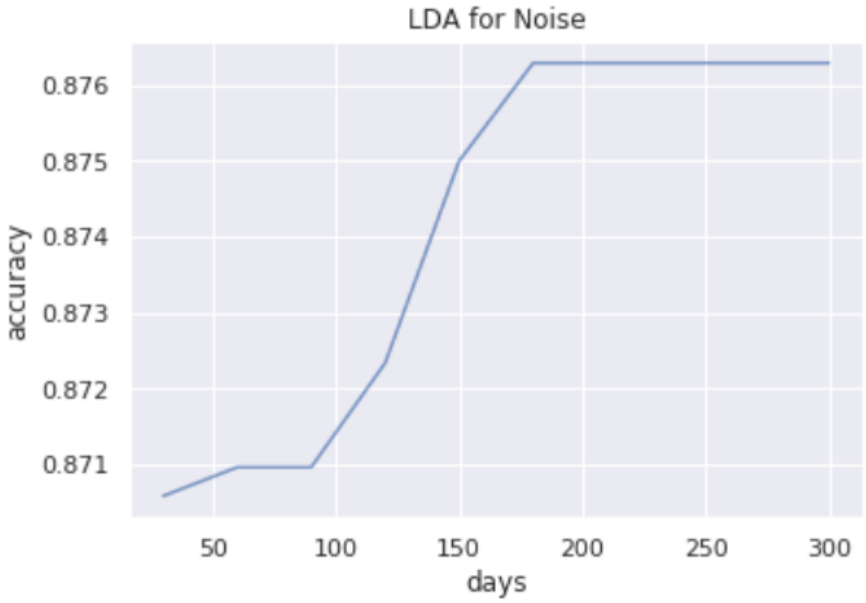

Figure 33: LDA classifier accuracy for noise data

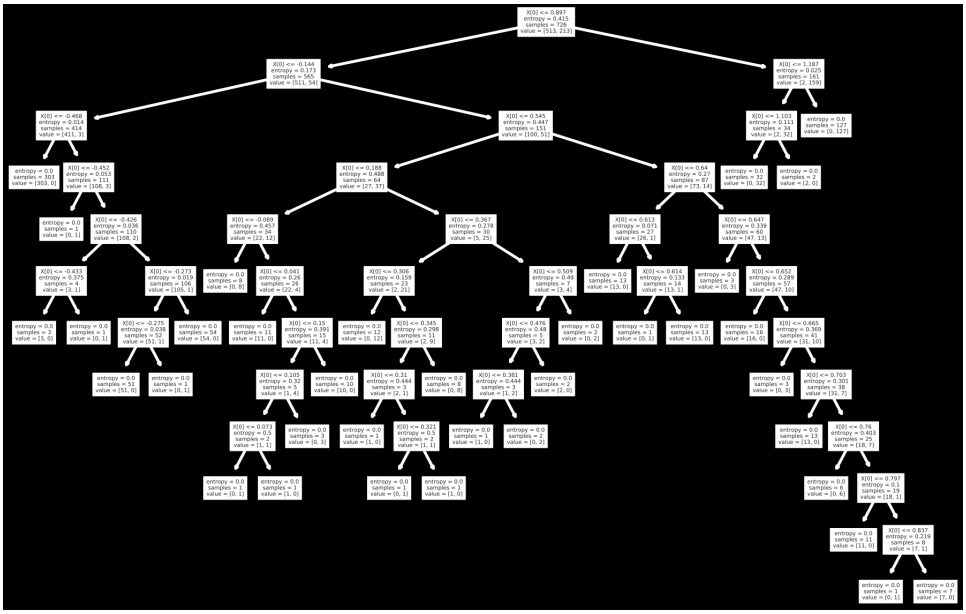

Figure 34: Decision tree for noise

The LDA performed on location data separates the two groups well, see figure 39, but has one patient that is on the control side right next to the control that is furthest away from the rest of the controls. The PCA method creates three different groupings, where the first group consists of one patient, the second group is mixed with both controls and patients and the last group contains only controls. The classification results are shown in figure 41. The location results also hold a high

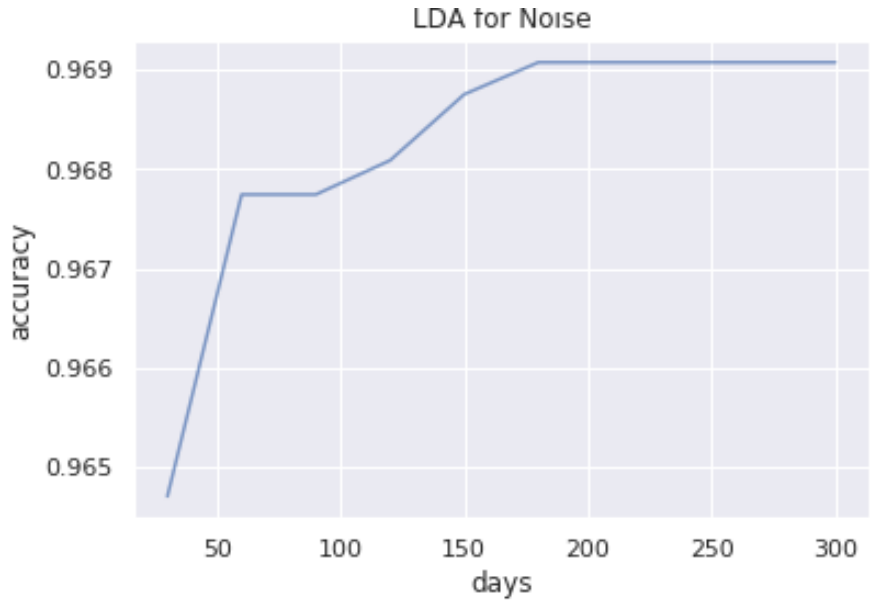

Figure 35: LDA and Decision Tree classifier mean accuracy for cross-validation for the different ranges for Noise

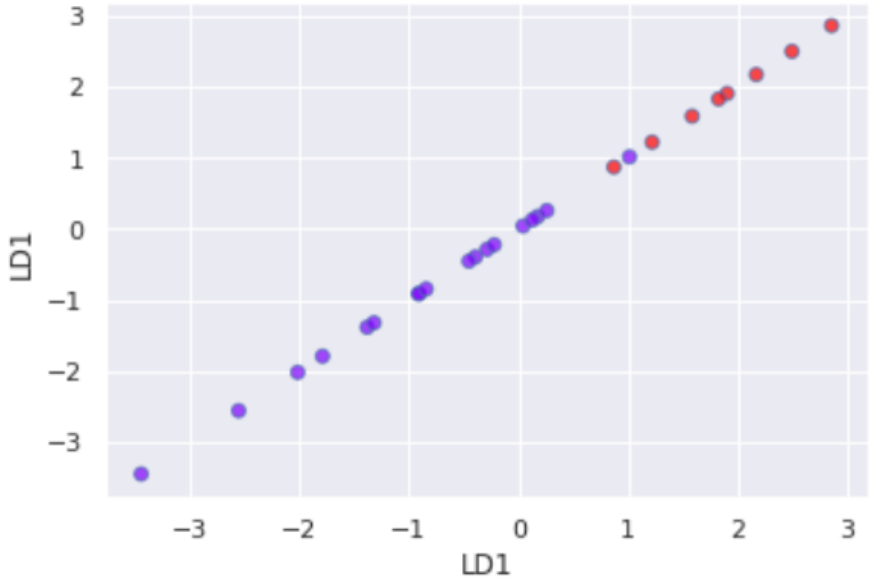

Figure 36: Visualisation of how well LDA was able to separate screen data.

accuracy varying between under 0.9728 and under 0.9742. The location model has, similarly as noise, a lower accuracy at fewer day splits and acquires a higher accuracy when more days are added to the splits.

The communication LDA separation is the most clear of all the data separations,

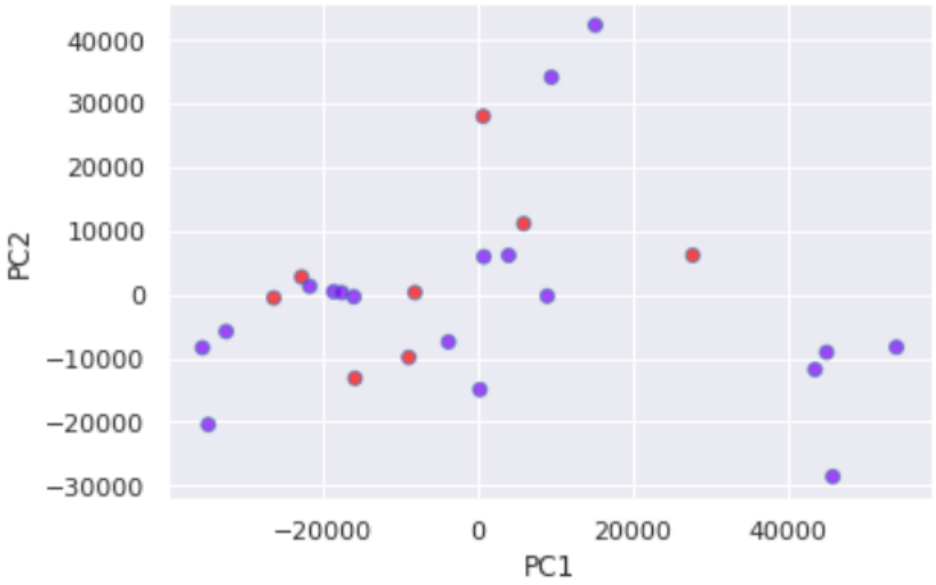

Figure 37: PCA calculated on screen data.

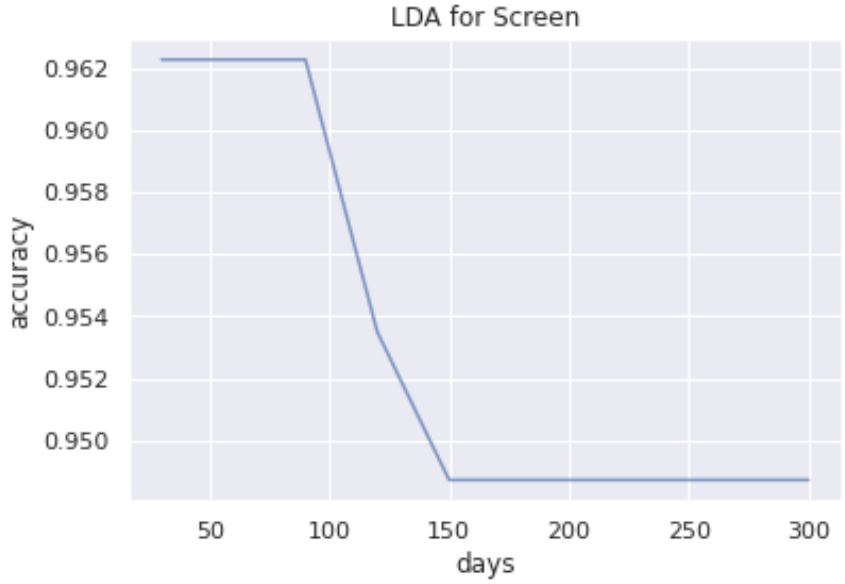

Figure 38: LDA and Decision Tree classifier mean accuracy for cross-validation for the different ranges for screen data.

see figure 42. The accuracy of the classification results was 1 for each split, and is therefore not plotted as a figure.

The application dataset is also well separated by LDA, see figure 43. The gap between the groups was however not the largest. The PCA was not able to distinguish

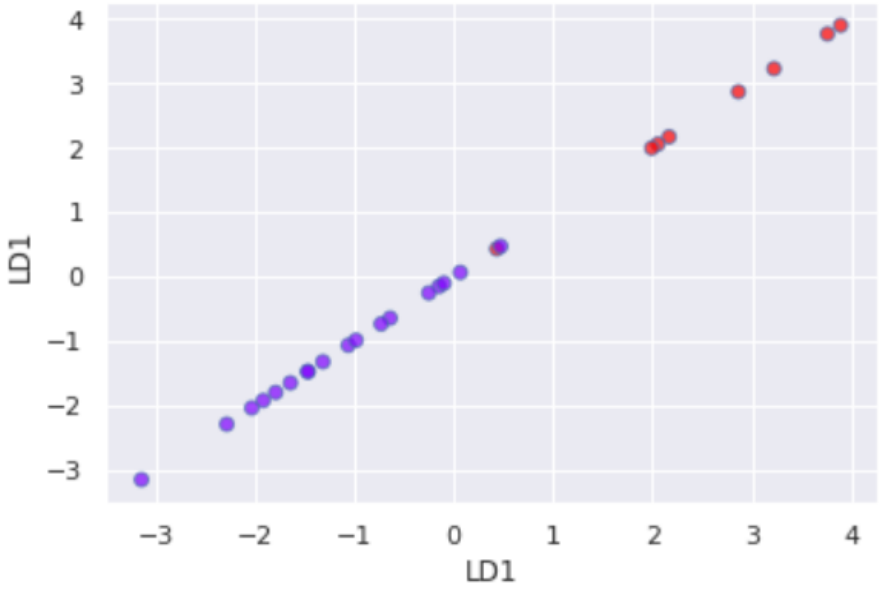

Figure 39: Visualisation of how well LDA was able to separate location data.

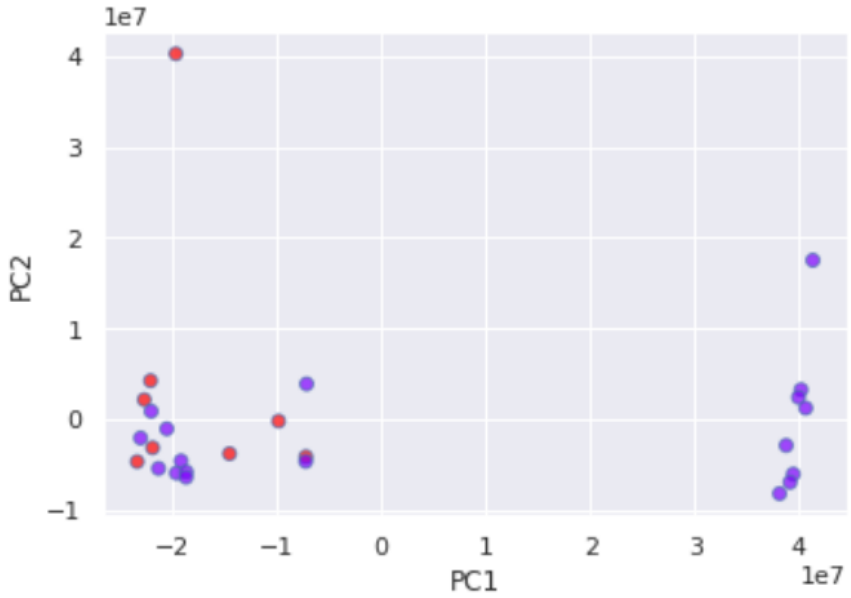

Figure 40: PCA calculated on location data.

any groups. The accuracy of the classification results was 1 for each split, and is therefore not plotted as a figure.

The last model consisted of all features in one model. First a model with monthly splits was created. The LDA separation is seen in figure 45. The accuracy of the model was 1 for all splits. The tree was plotted to see the structure. see figure 46.

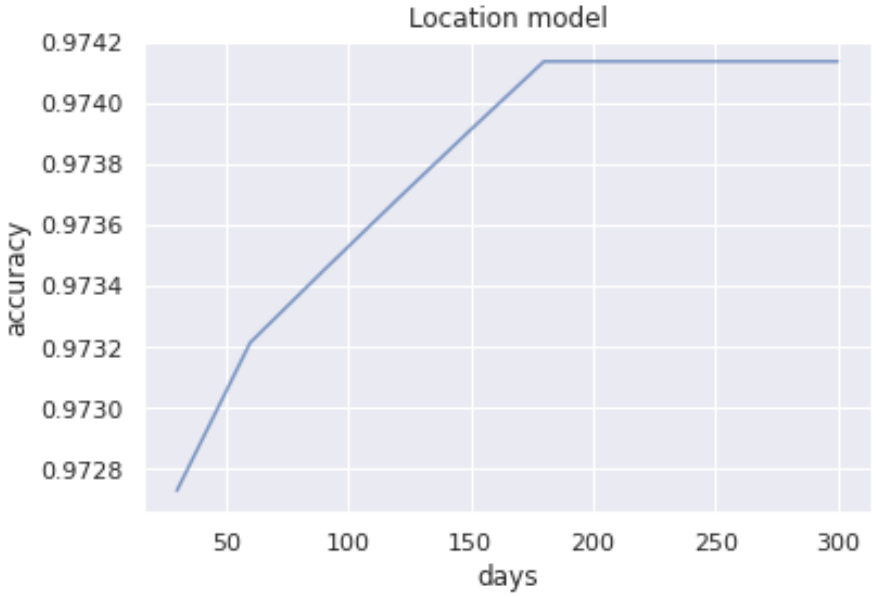

Figure 41: LDA and Decision Tree classifier mean accuracy for cross-validation for the different ranges for location data.

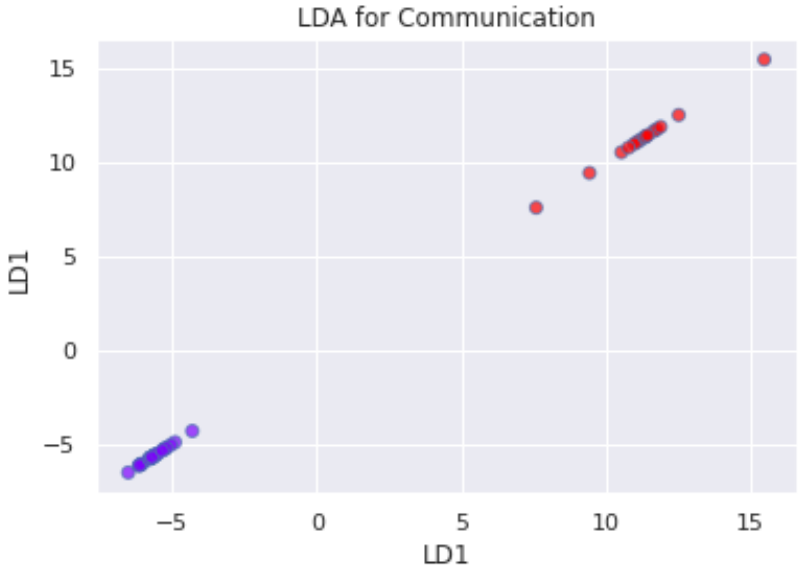

Figure 42: Visualisation of how well LDA was able to separate communication data.

Looking at the tree it seems like the model using months is underfitted, this is due to that a smaller depth of the tree increases the chances of bias. One way to prevent underfitting is to increase the number of samples, so next the splits were made smaller.

An method used to prevent underfitting was splitting the subject data into pieces

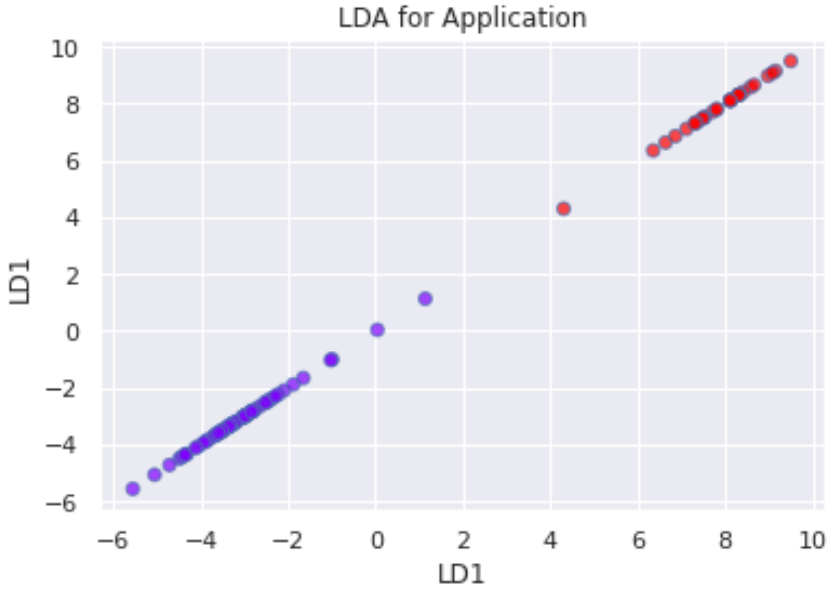

Figure 43: Visualisation of how well LDA was able to separate application data.

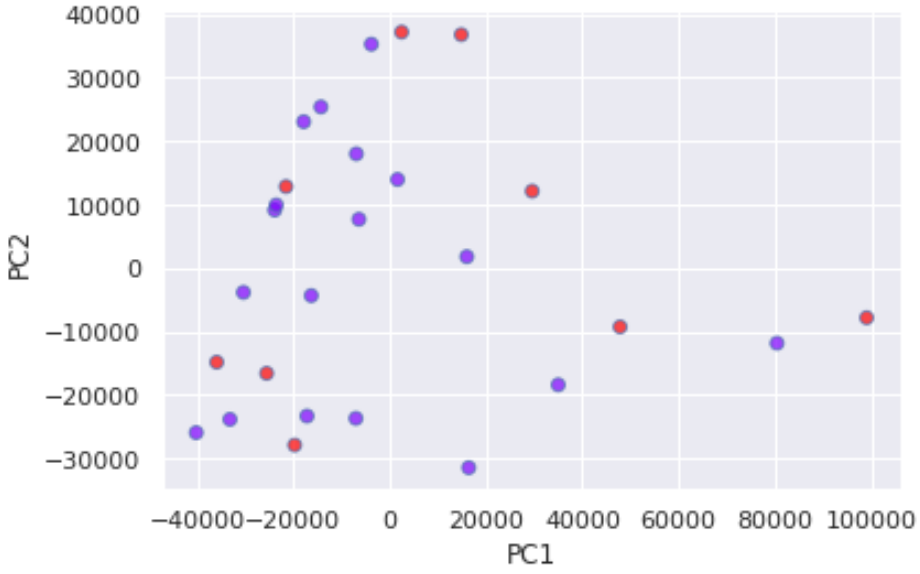

Figure 44: PCA calculated on application data.

based on the number of days. Deciding on the allowed number of, for example, 14-day intervals for each subject in the training set should balance the data. The same splitting was done for the test data, but each model was tested with all splits for each test subject. The smaller splits were 7, 14, 21, 28, 30 and 35. The resulting LDA is seen in figure 47 and contains all different split sizes. It is seen that there is plenty of overlap in the middle. The classification accuracy is seen in figure 49. The

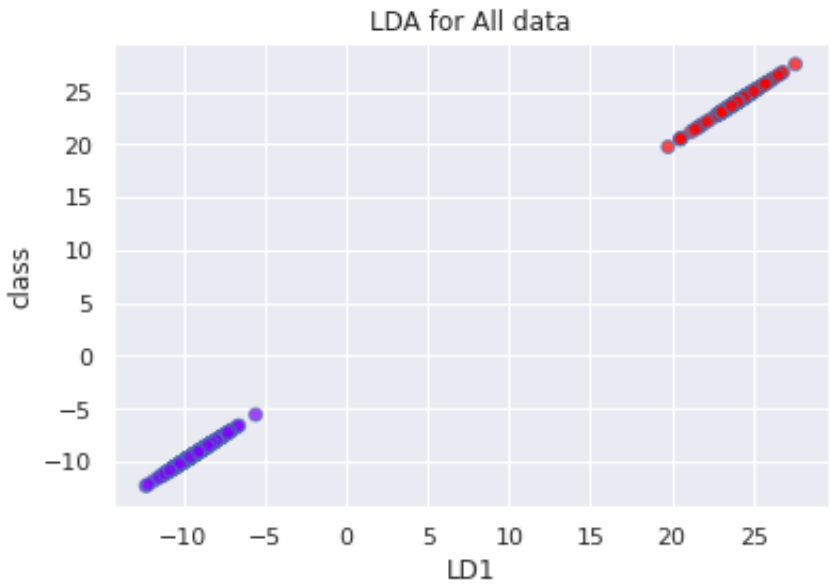

Figure 45: Visualisation of how well LDA was able to separate all data, when using more than 7 day intervals.

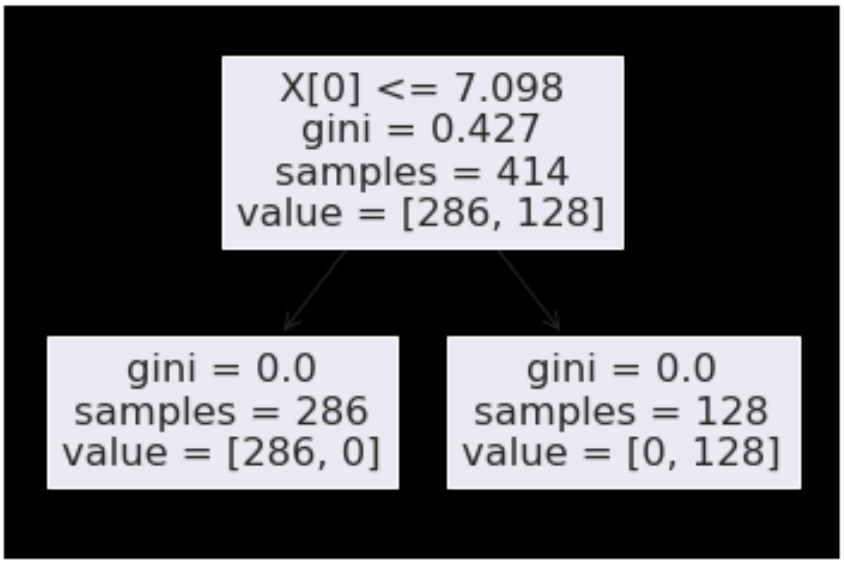

Figure 46: Decision tree for all data and using intervals of more than 7 days.

mean accuracy clearly drops for bigger split sizes than for the smaller. The tree in figure 48 contains more branches, which should indicate less underfitting.

Figure 49 is a mean of results for different training and test sizes, and different time intervals. When looking at the them separately the worst accuracy was found

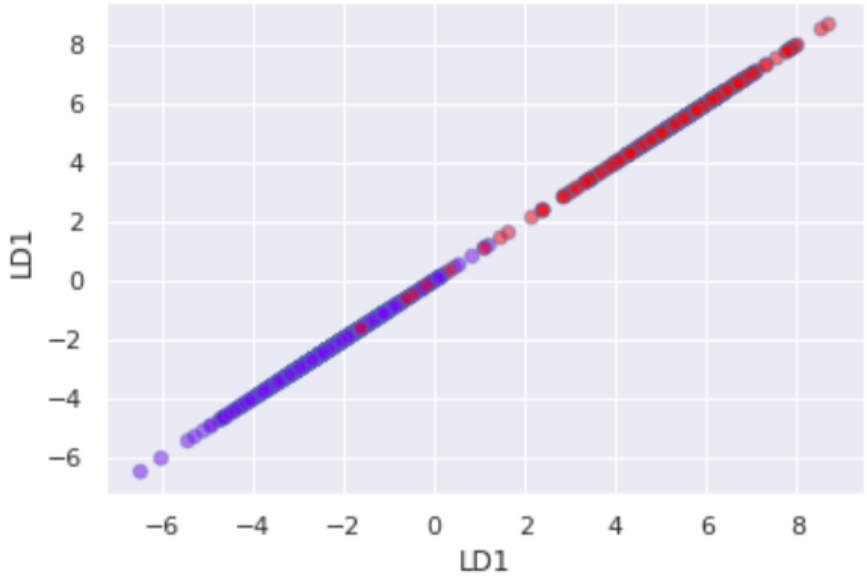

Figure 47: Visualisation of how well LDA was able to separate all data when using 7 day intervals.

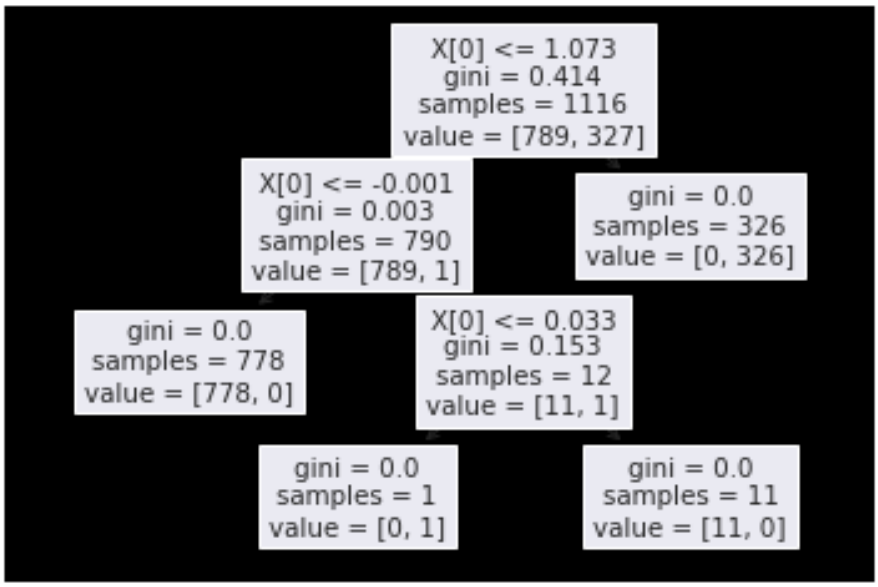

Figure 48: Decision tree for all data when using 7 day intervals.

in figure 50. It drops to an accuracy close to 0.2 at 27 days.

When comparing the results from the separate features and all features, it is seen that the separate features are giving more accurate predictions. This is probably due to that when a subject is missing data from one of the sensors, this has been

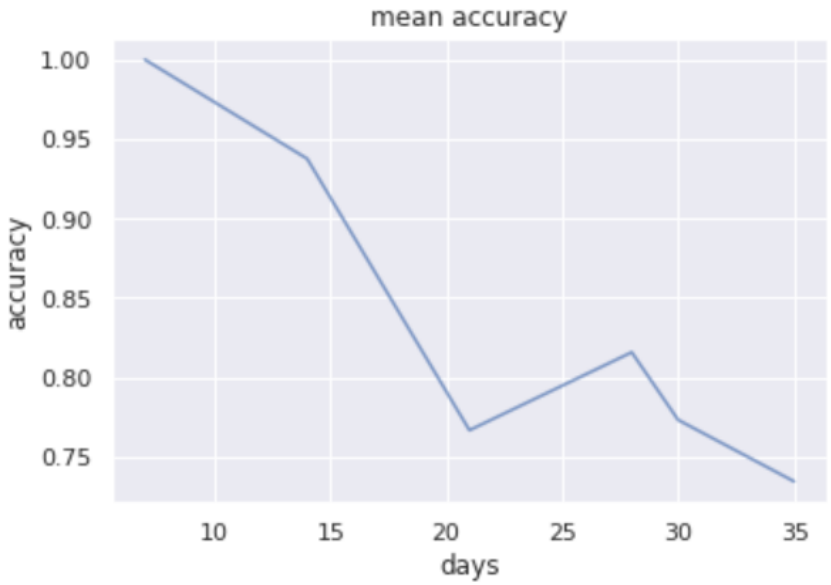

Figure 49: The mean accuracy for cross-validation of classification for different training and test set sizes, and different time intervals, when using all data.

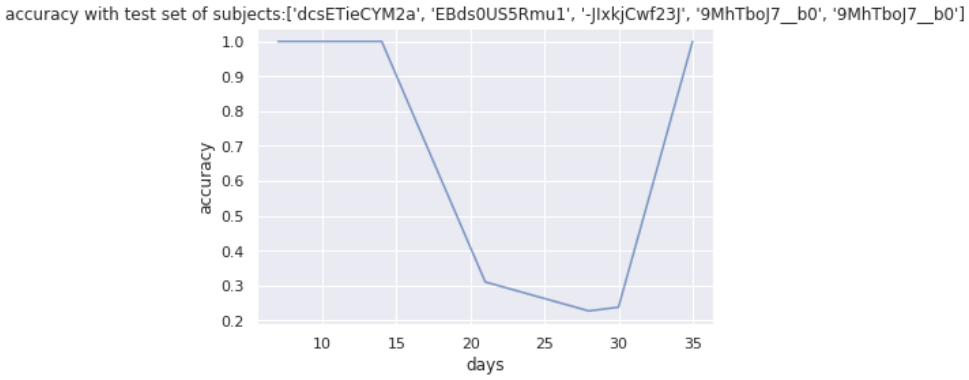

Figure 50: The model with worst accuracy.

imputed instead of not being discarded. Also models trained with machine learning tend to become worse with too many features compared to the amount of samples. With this in mind a better solution would be to classify subjects with separately trained models for each sensor and then summing the results of the different models to get the final classification result.

As the sample size is small it is possible the models are skewed or overfitted. The ratio between the subjects is also twenty-three controls versus fourteen patients, which also affects the results. An interesting point is that the control group consists of data collected by students. It would be interesting to see how the models would be affected by a more diverse control group.

Manual pruning could be useful in the future when there are more samples and more variance.

4.4 kh-segmentation results

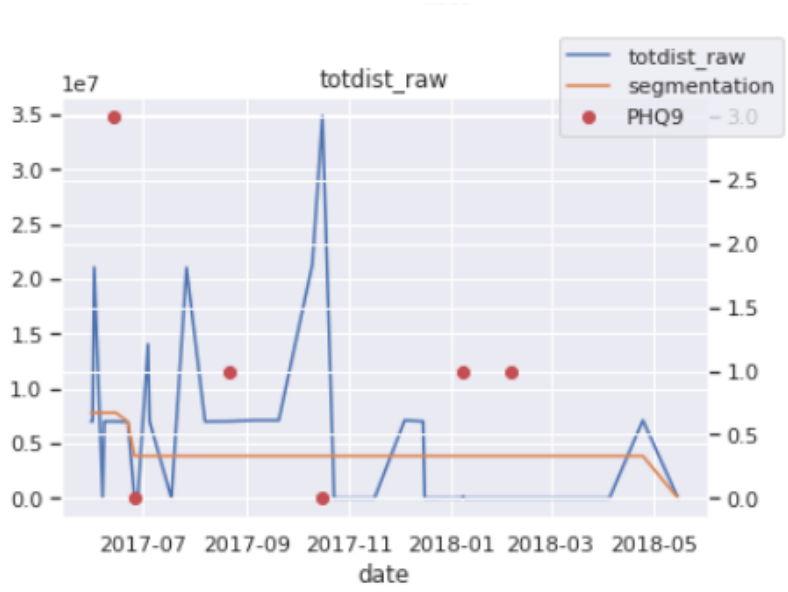

Figure 51: Example of kh-segmentation with k=4 and h=4 on low quality location data.

The hypothesis is that if there is a change in the PHQ-9 score, then there is a behavioural change in the smartphone usage.

First, the kh-segmentation code was tested with location data, this is due to that other studies have found preliminary support for the reliability and validity of location data as an objective measure of behaviour changes [23][51]. The code is first tested looking at separate sensor features and setting k and h to be the same value. Meaning the number of segments and sources are the same. The test values were 5 and 15 for both k and h, these were chosen arbitrarily but small enough to analyse intuitively. As stated in the Method section 3.4.5, the data was aggregated into days and looking at the 14 days before the taken PHQ-9 questionnaire. By plotting the aggregated data together with the segmentation, it is seen that the results produced by the kh-segmentation method are intuitively appealing. Results are shown in figures 51 and 52. The algorithm seems to work well with the location data, even though the location data seems to be very irregular and have some missing data. This could indicate that data is lost somewhere during the data processing pipeline, as there are timestamps for these data points. Nevertheless, the algorithm still seems

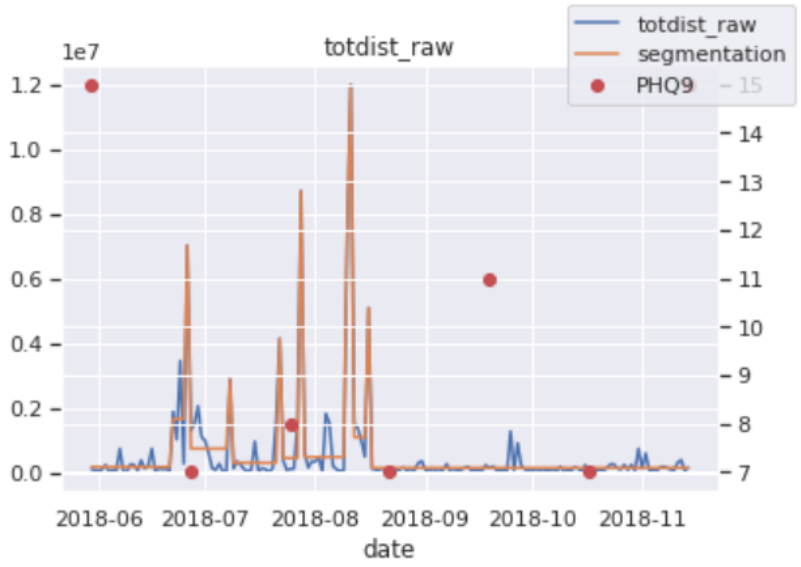

Figure 52: Example of kh-segmentation with k=14 and h=14 on high quality location data.

to find the most optimal segments for the given k and h. Clear segments can be seen and they seem to match changes in the data pattern.

Next, the segmentation was done including more than one feature, meaning comparing more than one data-sequence for getting the most optimal segmentation. The chosen numbers for k and h were set based on the amount of available PHQ-9 scores. k is set to two times the amount of PHQ9 and h is set to amount of unique PHQ9 scores. The results are shown in figures 53, 54 and 55.

The location data again shows problems in figure 54 due to missing data. However the segmentation can again be seen as correct. Figures 53 and 55 do not have missing data to the same extent and produces better segments. Looking at figures 53 and 55, it is also seen that peaks tend to get their own segment. This due to a considerable change in the data sequence.

It can now be seen that, even though the segmentation of the location data is correct, the segmentation does not add value as the location data used in this thesis is missing too much information. Thus the segmentation method is not to be used on sensor data with too few data points or with too much missing data. Too few data points in this case means that a describing aggregation of a day cannot be made from the data. If it is possible to improve the quality of the location data, segmentation could be used.

It seems as if the areas of missing data are marked with a segment of its own. This is seen especially for the subjects that have more data points to look at than other subjects, but still have missing data points. This can be seen in figure 54, where the first segment contains peaks in the data, but are not enough to create a new segment.

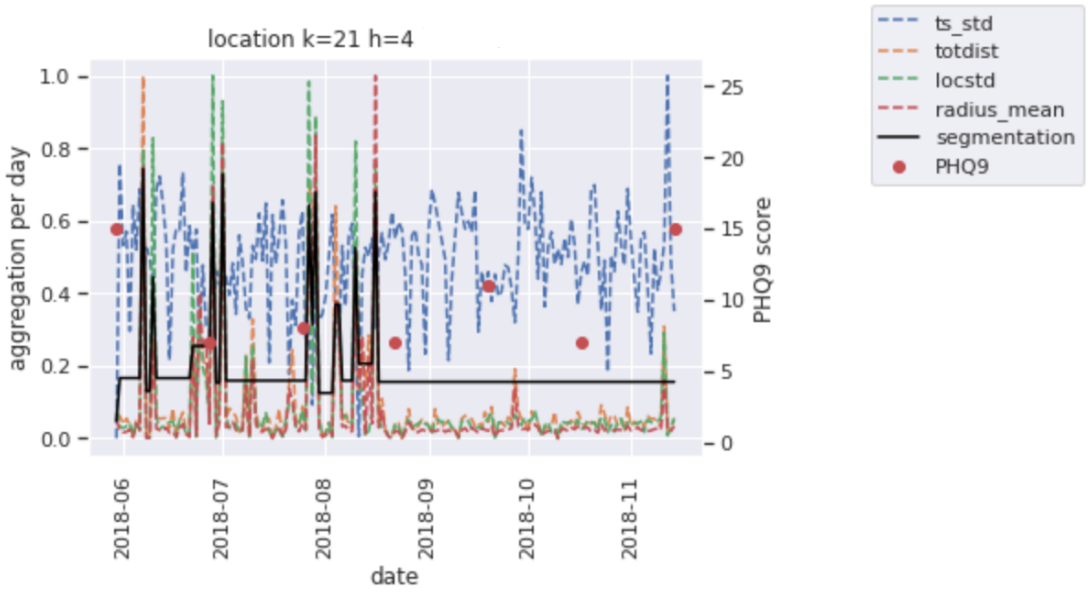

Figure 53: Example of kh-segmentation with k=21 and h=4 on location data. k is set based on the amount of PHQ9 times three and h is set based on the amount of unique PHQ-9 scores.

Next, the screen data was tested as it has more data points and less missing data than location data. k and h were again set based on the amount of available PHQ-9 scores. k is set to two times the amount of PHQ9 and h is set to amount of unique PHQ9 scores. The algorithm seems to work well with the screen data, as clear segments can be seen and they seem to match changes in the data pattern. This can be seen in figure 56.

More data points seems to result in a better segmentation, which agrees with that the ClusteringSegments method obtains a better accuracy when provided a longer sequence [24]. The correct segments and sources are easier to find with more data points. Due to this only screen data will be used for segmentation. As screen data is of better quality with regards to data points and missing data, and the aim is to find the most suitable k and h for each subjects data, using different criteria. The different criteria to be tested were presented in the method section 3.4.5.

The first criteria was based on having the PHQ-9 score as ground truth and measuring how well it correlates with the segments. As each subject has a different amount of PHQ-9 scores, the max range for k was set to 30 times the number of PHQ-9 scores. This was an arbitrary number that was manually tweaked to work based on that the optimal k was most probably found before the max k. Most probably means that the optimal k had not changed in several cycles before reaching the max k. Another option could be choosing the max range for k by the changes in the PHQ-9. However, in a few cases the max allowed k was the most optimal, according to the criteria. It should be taken to consideration that the longer the range is for searching the most optimal k, the longer it takes to run the code. One

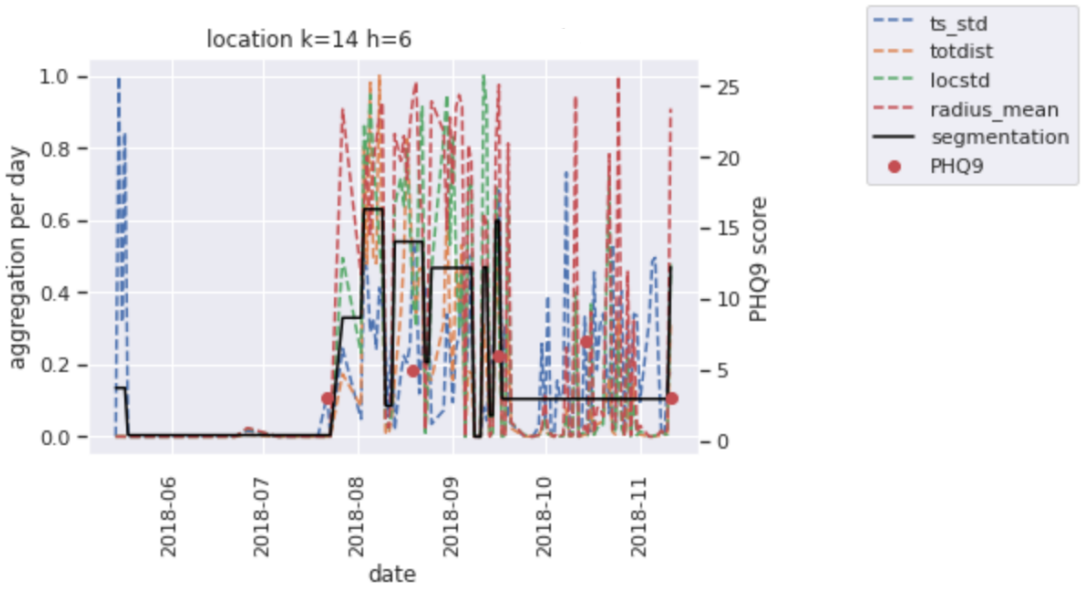

Figure 54: Example of kh-segmentation with k=14 and h=6 on location data. k is set based on the amount of PHQ9 times two and h is set based on the amount of unique PHQ-9 scores.

possible case is that some data sequence achieves a better correlation with PHQ-9 the higher the k is, meaning that k equal to the number of data points gives the most optimal result. So a restriction for max k is in order.

When applying the PHQ-9 error measure, it was first optimised based on k. This resulted in a more granular result, with more segments, see figure 58. The correlation between the segments and the PHQ-9 score is shown in figure 59. The max correlation score is 15 and by only optimizing k the score was 6.6.

Secondly, it was optimised on either k or h, choosing the k or h that gave the highest correlation value. This resulted in less segments k and also a variation of h for most subjects. See figure 62.

For k or h, the h is often lower than when maxing both k and h. This seems to be the case for controls. But for patients the k and especially h changes for maxing both k and h. This could be explained by the changing PHQ-9 score. This is another reason for why it is hard to say if the PHQ-9 criteria is good or not.

Lastly, it was optimised using the k and h that together gave the highest correlation. See figures 63, 64, 65 and 66. This resulted in a high k and a high h.

One major problem with having the PHQ-9 score as ground truth and as an error measure for the kh-segmentation is that some subjects do not have a change in the PHQ-9. This is a problem due to that when there is no change in the PHQ-9 score no correlation between the PHQ-9 score and the segmentation can be calculated.

The second criteria was based on the BIC error measure, which was implemented in [24]. A lower BIC value means a better segmentation. See figures 57 and 60 for results.

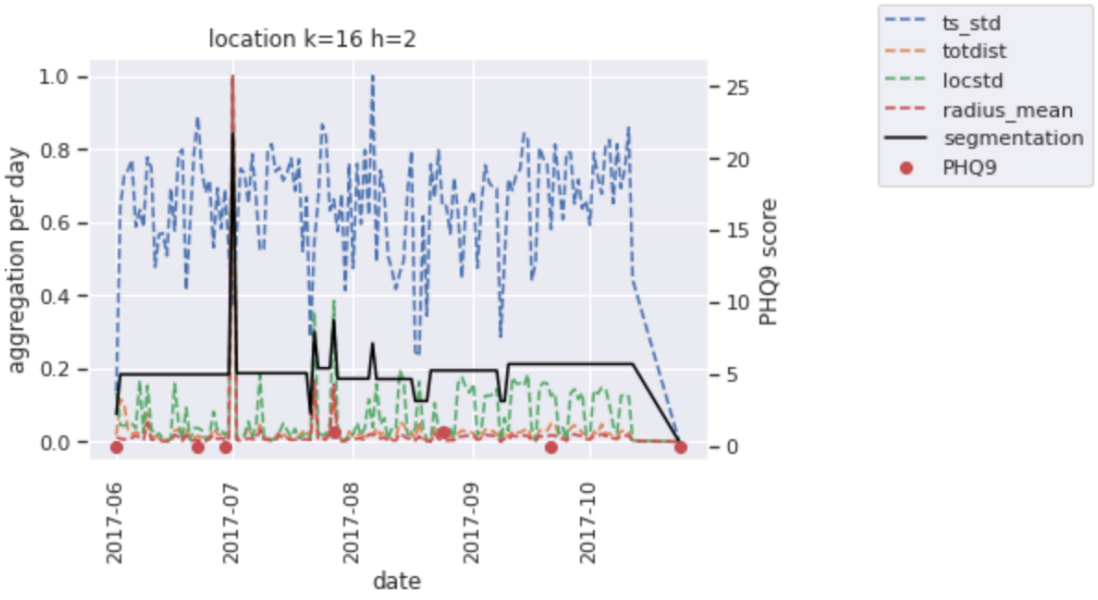

Figure 55: Example of kh-segmentation with k=16 and h=2 on location data. k is set based on the amount of PHQ9 times two and h is set based on the amount of unique PHQ-9 scores.

Figure 56: kh-segmentation of screen data for patient, with k=70 and h=4, where k is chosen 10 times number of PHQ-9 and h is the number of unique PHQ-9 scores.

An interesting property of the segmentation, when using the BIC error measure, was that the BIC value was lowest when k and h was the same. For future research the h could be chosen in a different way, for example, allowing an maximum BIC error when reducing h.

When comparing the error measure of PHQ-9 and BIC, see figures 58 and 60, it

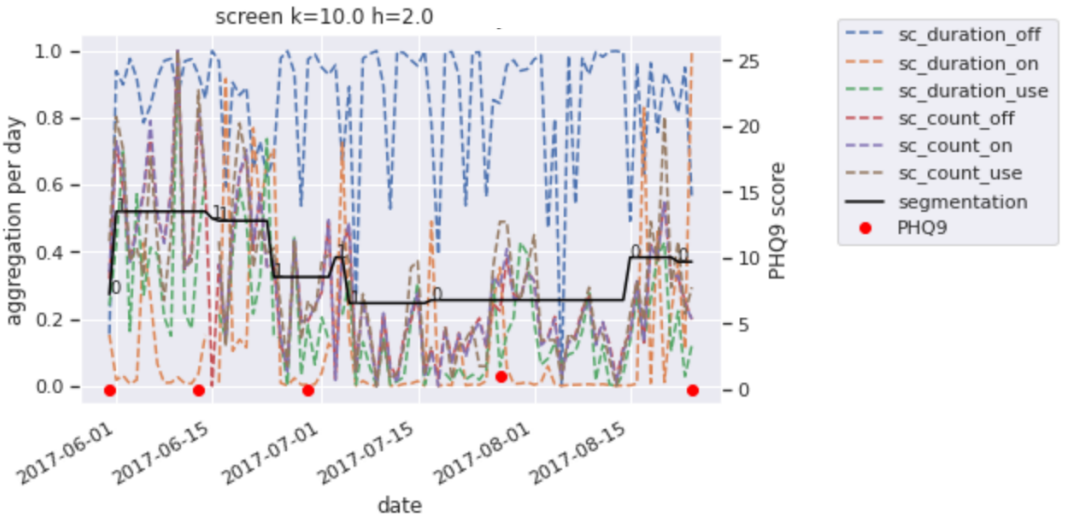

Figure 57: kh-segmentation of screen data for control, with lowest BIC of 156.312 when k=10 and h=2.

is seen that the resulting segmentation is very different. It seems like the PHQ-9 error measure gives a more granular result, with more segments and the BIC measure results in a less granular result, with fewer segments.

To validate the different criteria, the correlation was calculated between the segmentation and PHQ-9 score. This was compared to the average segmentation were k was set highest possible. First a comparison for a control is made and the average segmentation is seen in figure 73. The segmentation for maximising k and h is seen in figure 67 and the correlation is seen in figure 68. It is seen that the correlation for the criteria segmentation is lower than for average.

The segmentation for maximising k or h is seen in figure 69 and the correlation is seen in figure 70. It is seen that the k correlation for the criteria segmentation is higher than for average.

The segmentation for BIC is seen in figure 71 and the correlation is seen in figure 72. It is seen that the correlation for the criteria segmentation is higher than for average.

Second a comparison for a patient is made and the average segmentation is seen in figure 80. The segmentation for maximising k and h is seen in figure 74 and the correlation is seen in figure 75. It is seen that the correlation for the criteria segmentation is lower than for average.

The segmentation for maximising k or h is seen in figure 76 and the correlation is seen in figure 77. It is seen that the h correlation for the criteria segmentation is higher than for average.

The segmentation for BIC is seen in figure 78 and the correlation is seen in figure 79. It is seen that the correlation for the criteria segmentation is lower than for average.

When comparing the criteria by looking at the correlations between the segmen-

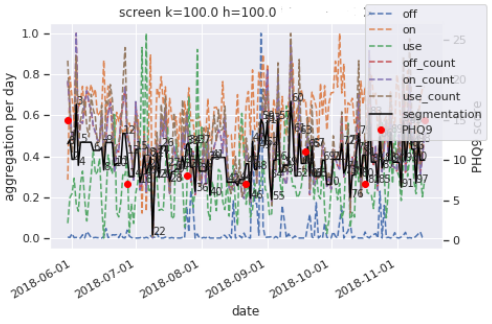

Figure 58: kh-segmentation of patient screen data using the correlation between PHQ-9 and the segmentation as the error measure. Here the optimal $k = 100$ and $h = 100$, where the correlation error measure was 6.63.

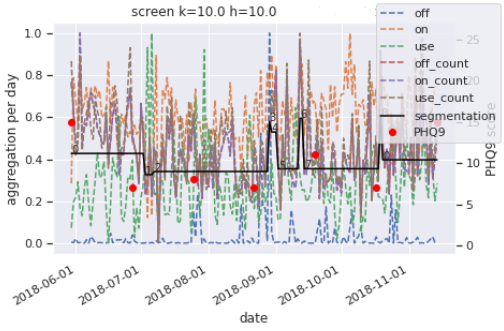

Figure 60: kh-segmentation of patient screen data using BIC as the error measure. Here the optimal $k = 10$ and $h = 10$, where the best BIC was 392.166.

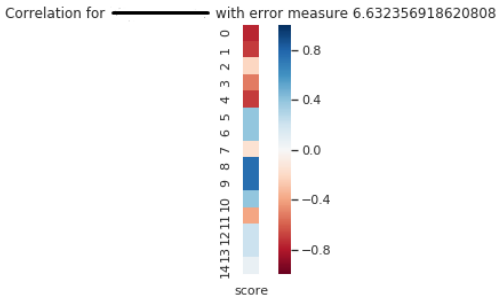

Figure 59: The correlation between the segmented fortnights and their corresponding PHQ-9 scores, for kh-segmentation of patient screen data. Random identifier redacted for information security.

Figure 61: Different segmentation results for same patients screen data.

Figure 62: kh-segmentation of screen data for control, when maximising correlation with PHQ-9 for either k or h.

tations and PHQ-9 score, and the k and h that give the most optimal segmentation there seems to be some connections. Firstly, when maximising both k and h it seems like the it tends to set k to maximal allowed value. This means it will be very close to average or average. This criteria will also tends to perform worse than average when comparing the correlations, see figures 68 75.

Secondly, when maximising for k or h both k and h tend to stay low. Looking at the correlation with PHQ-9 score, see figures 70 and 77, the correlation for either k or h is higher than for the average correlation.

Lastly, when minimising BIC k and h are set to a low value. The correlation is

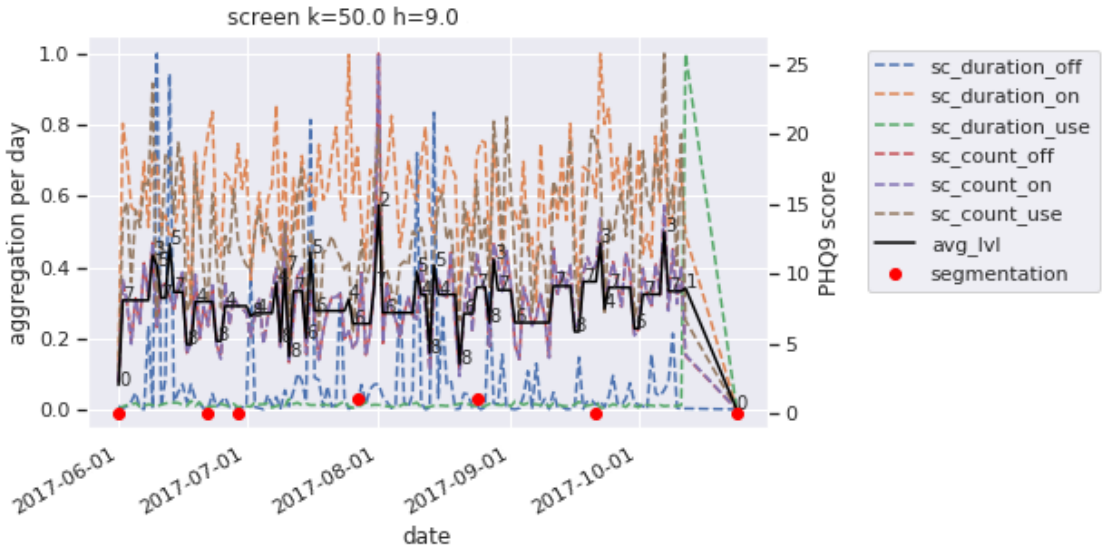

Figure 63: kh-segmentation of screen data for control, when maximising correlation with PHQ-9 for both k and h.

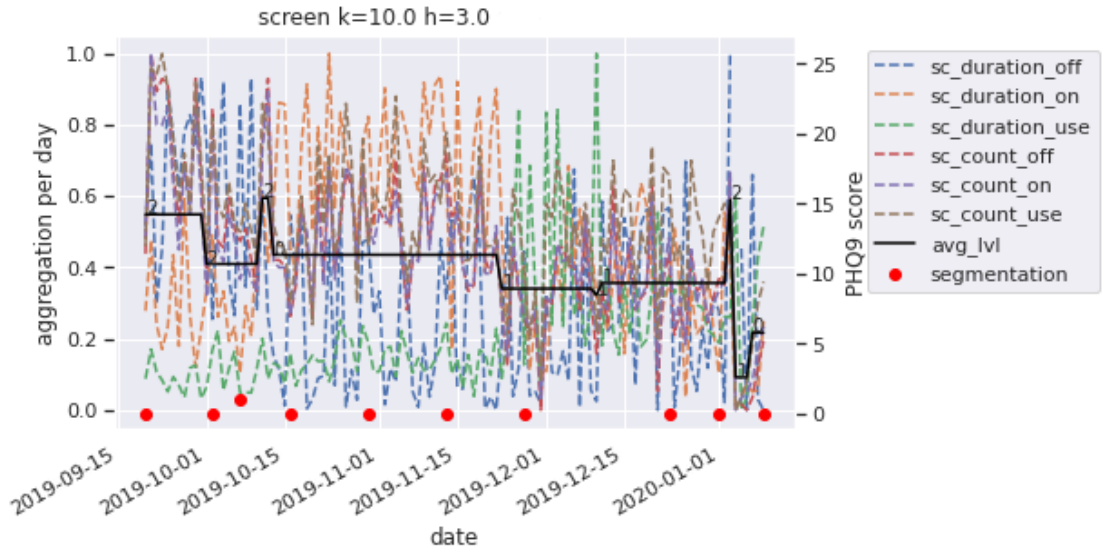

Figure 64: kh-segmentation of screen data for control, when maximising correlation with PHQ-9 for both k and h.

mostly better than average, but there are cases where the correlation is lower than average e.g. figure 79.

When writing the results section of this thesis, a bug was found in the location data converter code. It seemed as if the longitude and latitude was calculated incorrectly dropping a whole data channel, causing the distance measuring to be incorrect. Hopefully a re-work and re-run of the code would be enough to improve

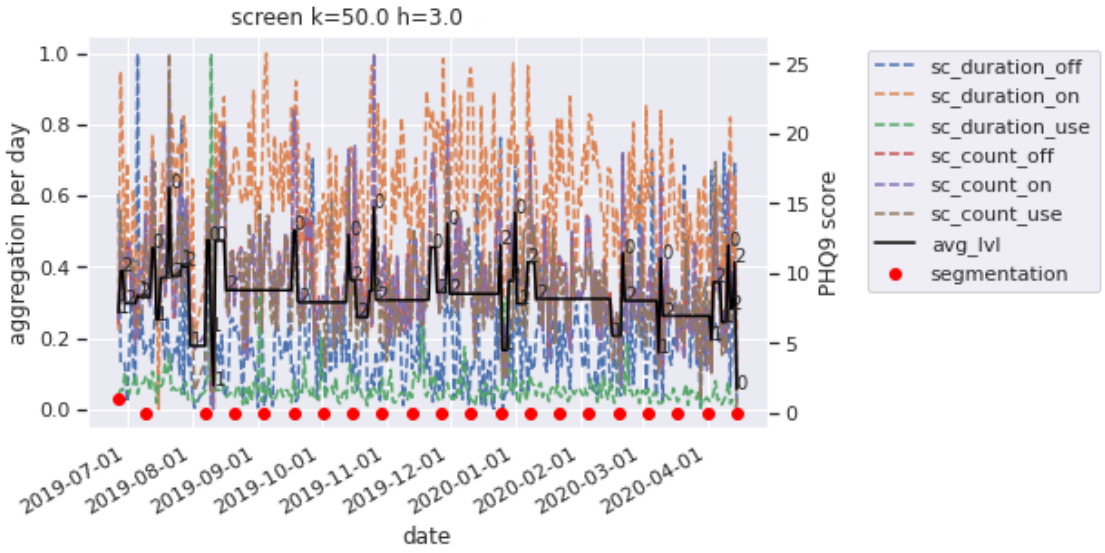

Figure 65: kh-segmentation of screen data for control, when maximising correlation with PHQ-9 for both k and h.

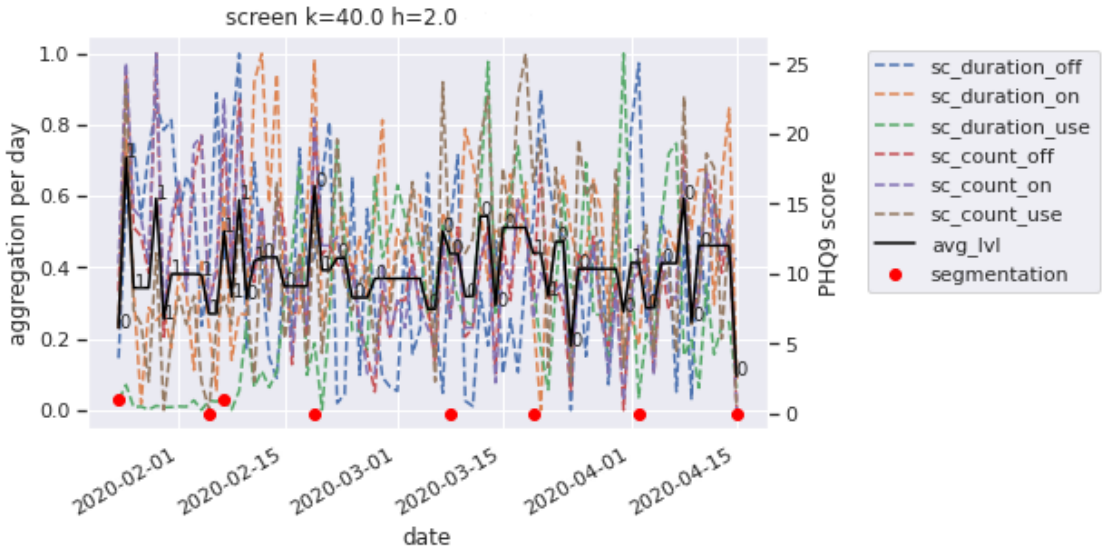

Figure 66: kh-segmentation of screen data for control, when maximising correlation with PHQ-9 for both k and h.

results. Unfortunately the data converter is outside the scope of this thesis.

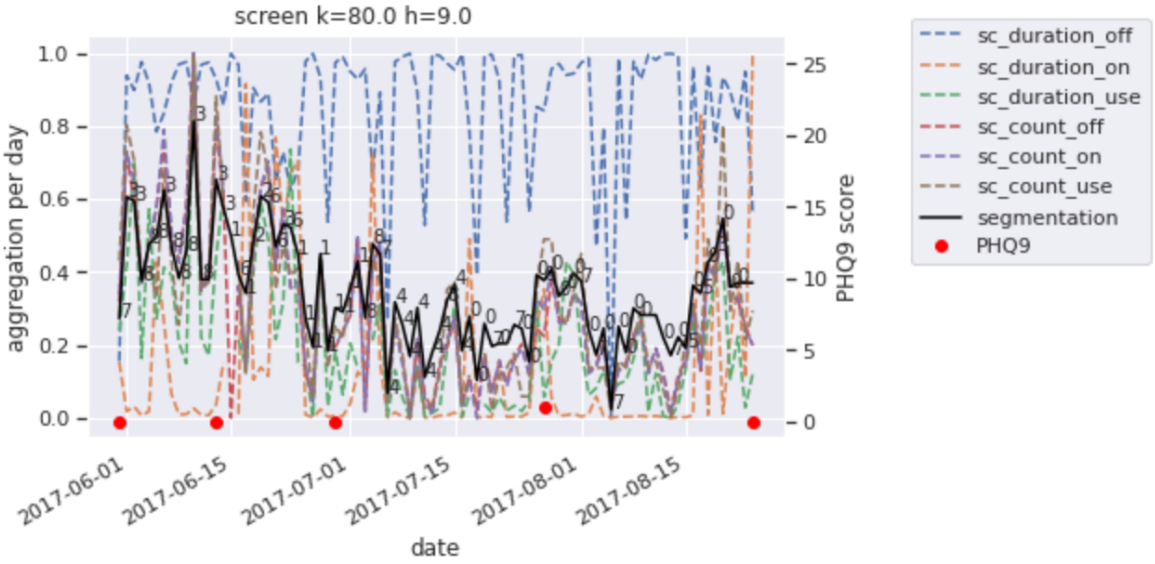

Figure 67: kh-segmentation of screen data for control, when maximising correlation with PHQ-9 for both k and h.

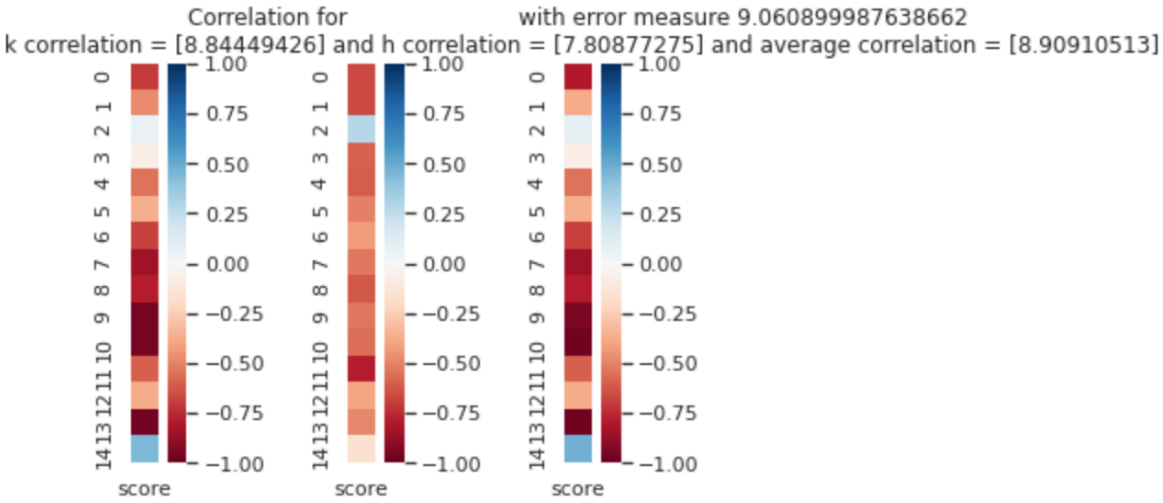

Figure 68: Correlations for control, when maximising correlation with PHQ-9 for both k and h, and the average correlation.

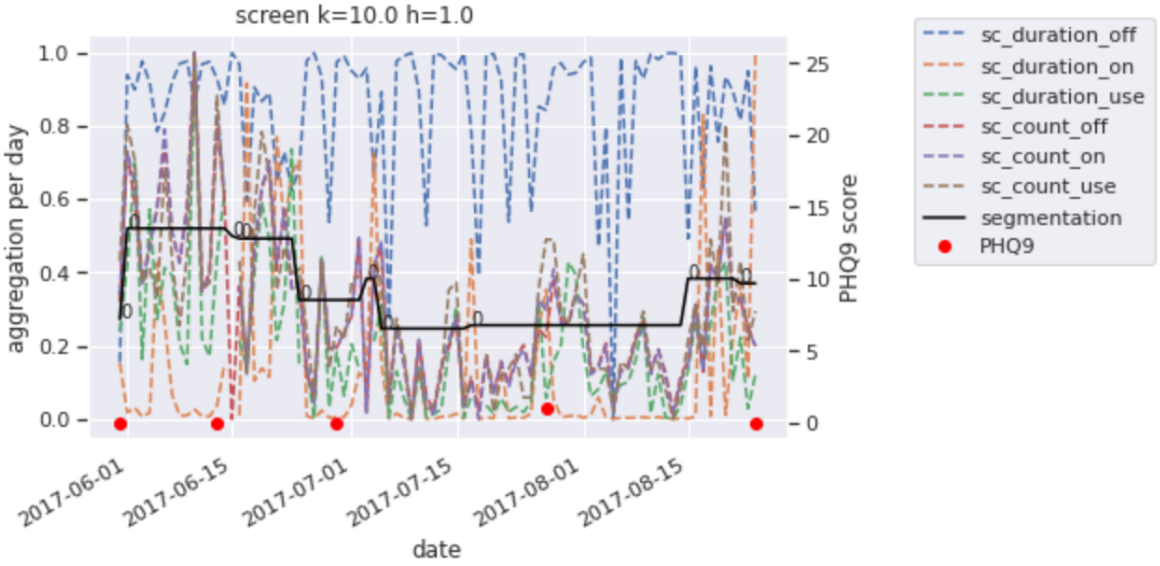

Figure 69: kh-segmentation of screen data for control, when maximising correlation with PHQ-9 for k or h.

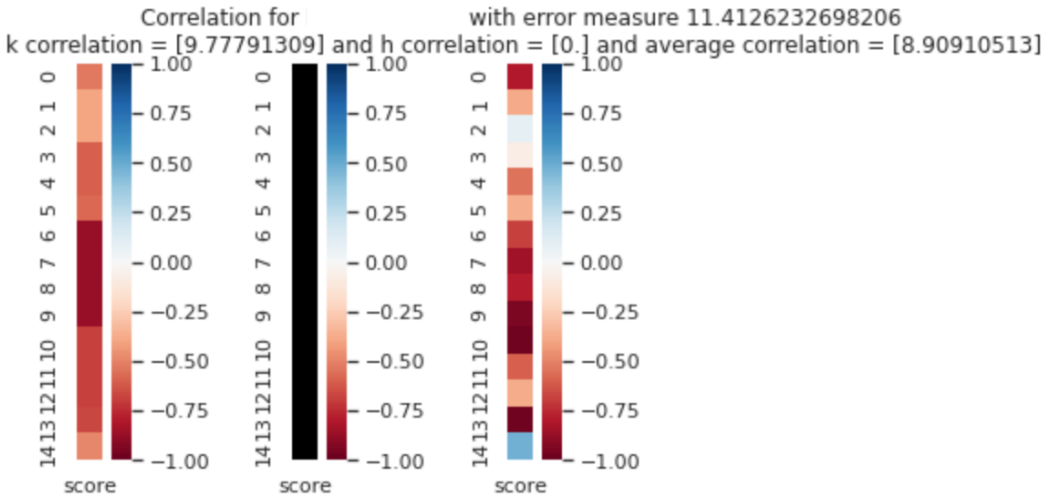

Figure 70: Correlations for control, when maximising correlation with PHQ-9 for k or h, and the average correlation.

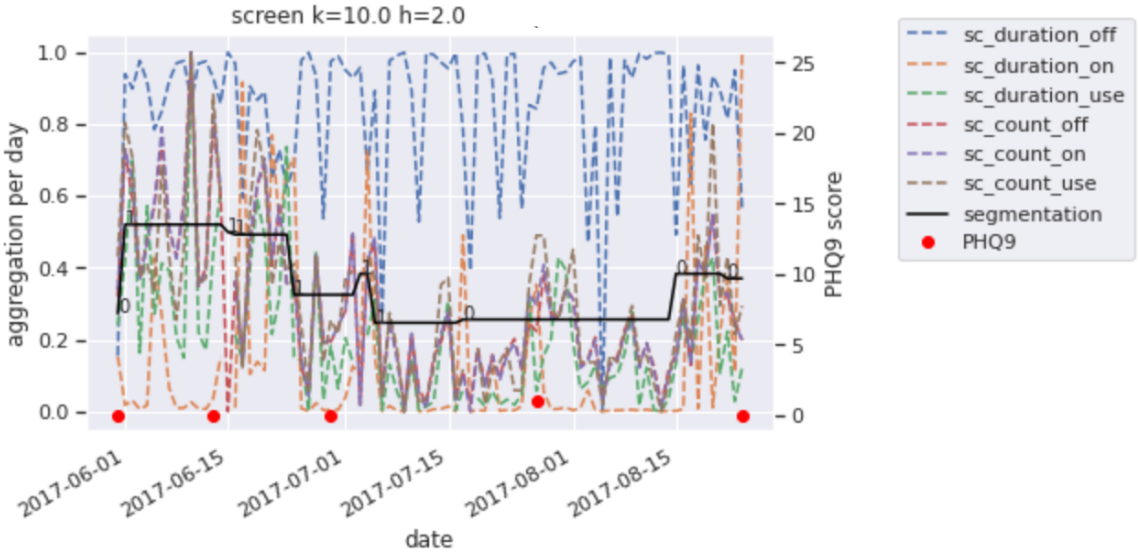

Figure 71: kh-segmentation of screen data for control, when minimizing BIC.

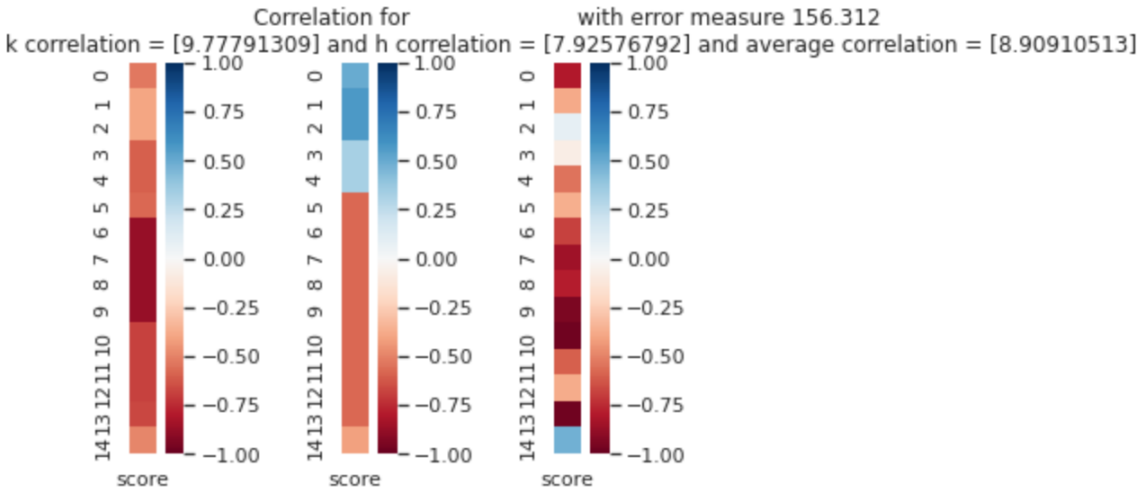

Figure 72: Correlations for control, when minimizing BIC and the average correlation.

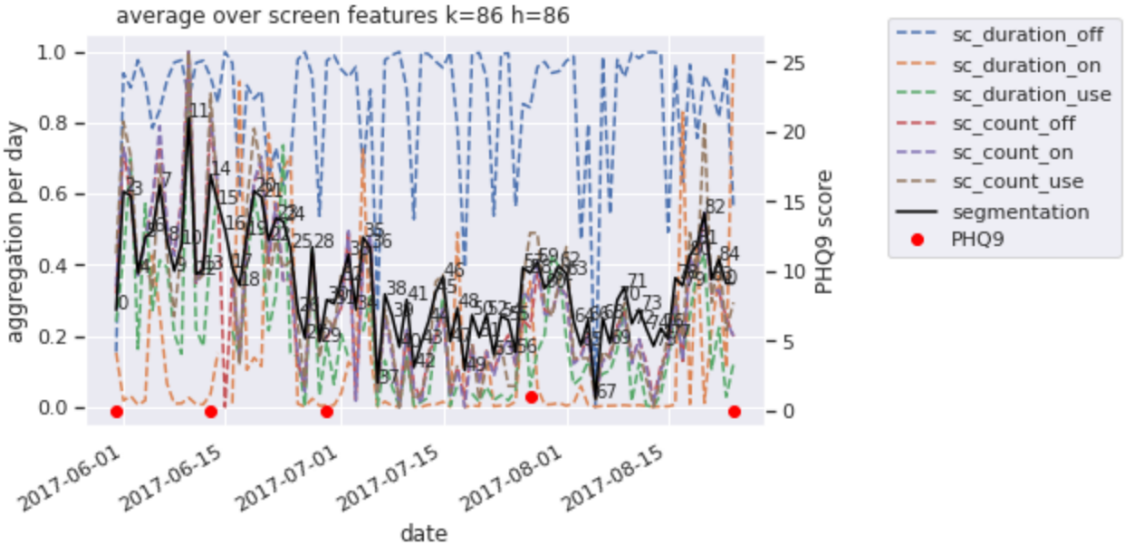

Figure 73: kh-segmentation of screen data for control for average.

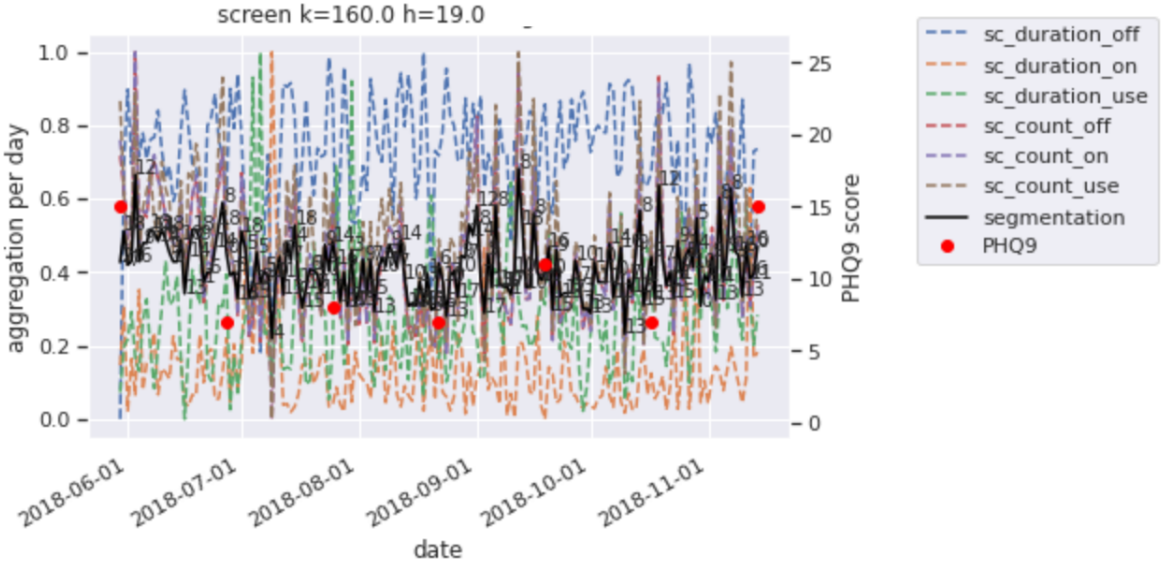

Figure 74: kh-segmentation of screen data for patient, when maximising correlation with PHQ-9 for both k and h.

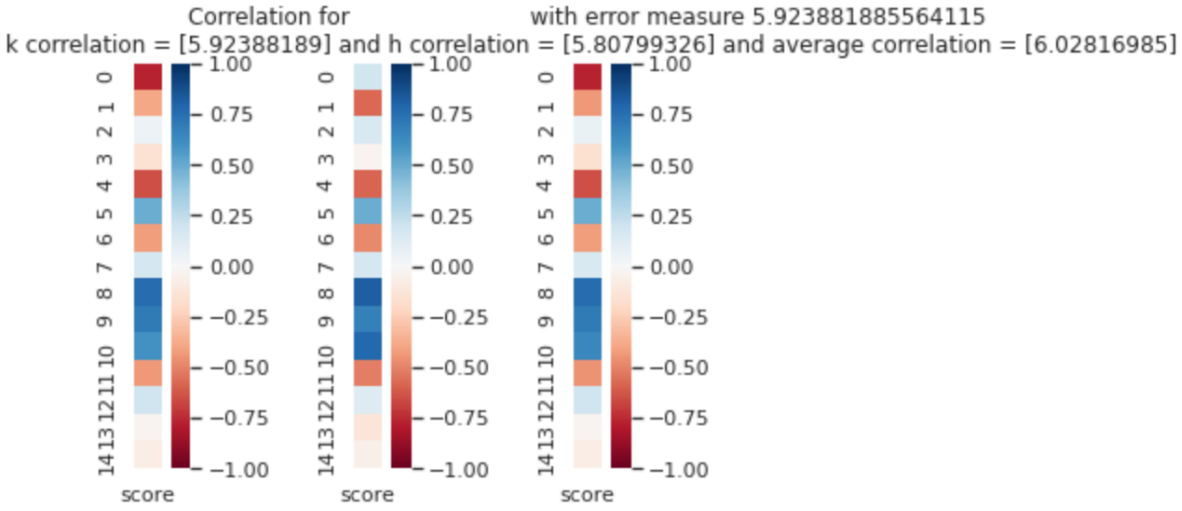

Figure 75: Correlations for patient, when maximising correlation with PHQ-9 for both k and h, and the average correlation.

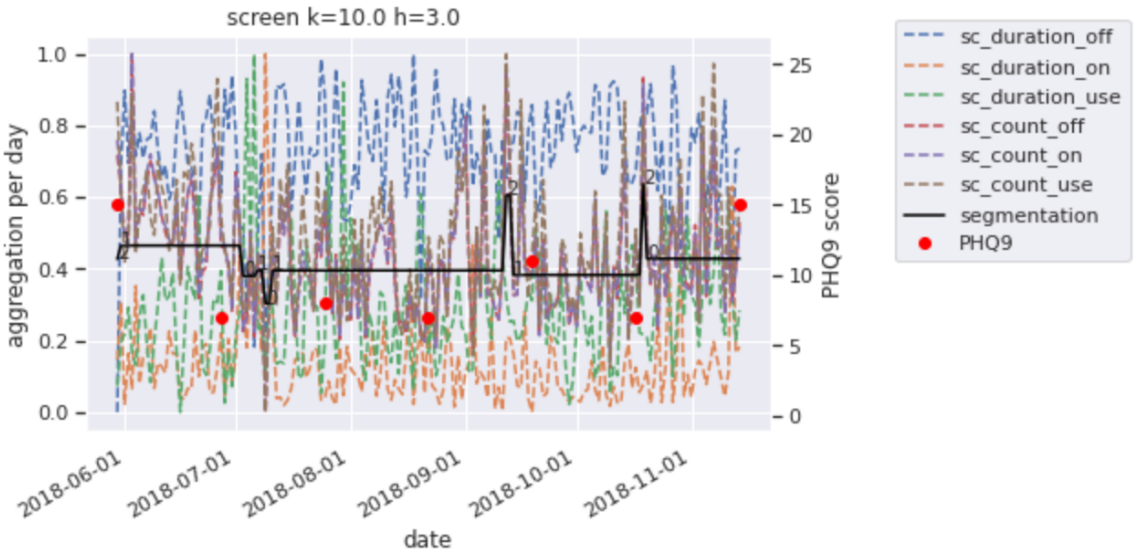

Figure 76: kh-segmentation of screen data for patient, when maximising correlation with PHQ-9 for k or h.

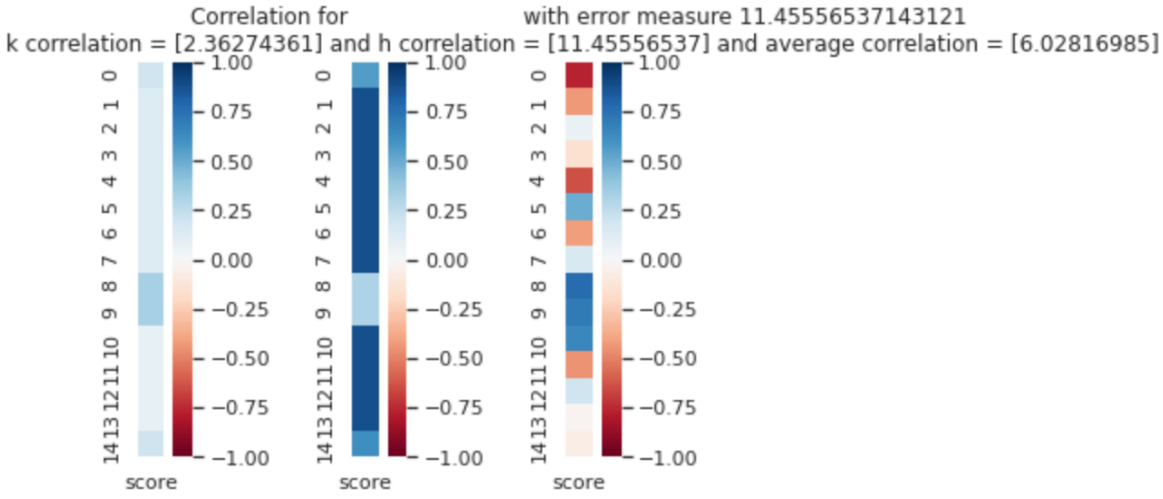

Figure 77: Correlations for patient, when maximising correlation with PHQ-9 for k or h, and the average correlation.

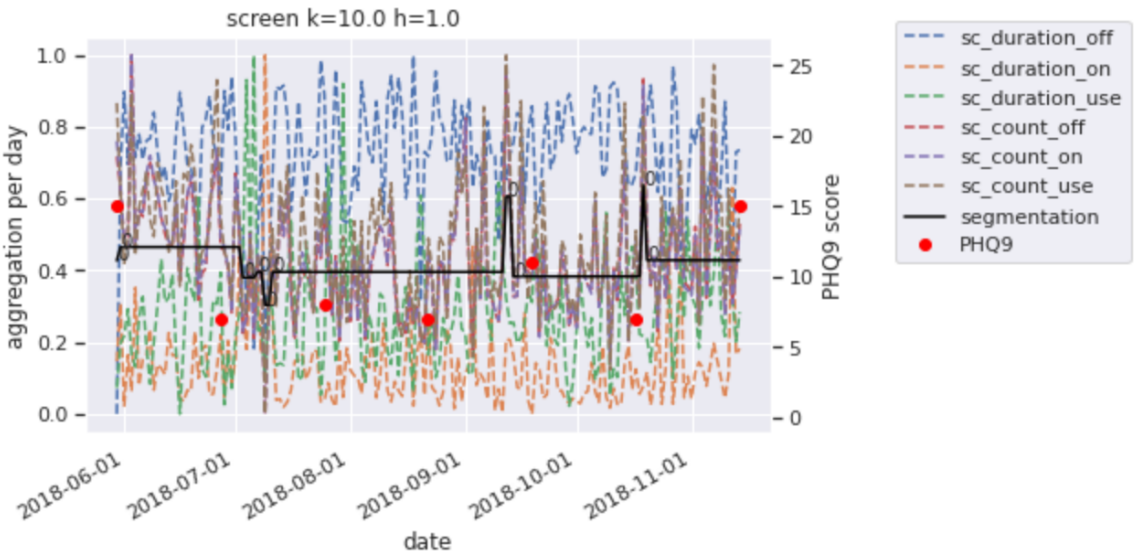

Figure 78: kh-segmentation of screen data for patient, when minimizing BIC.

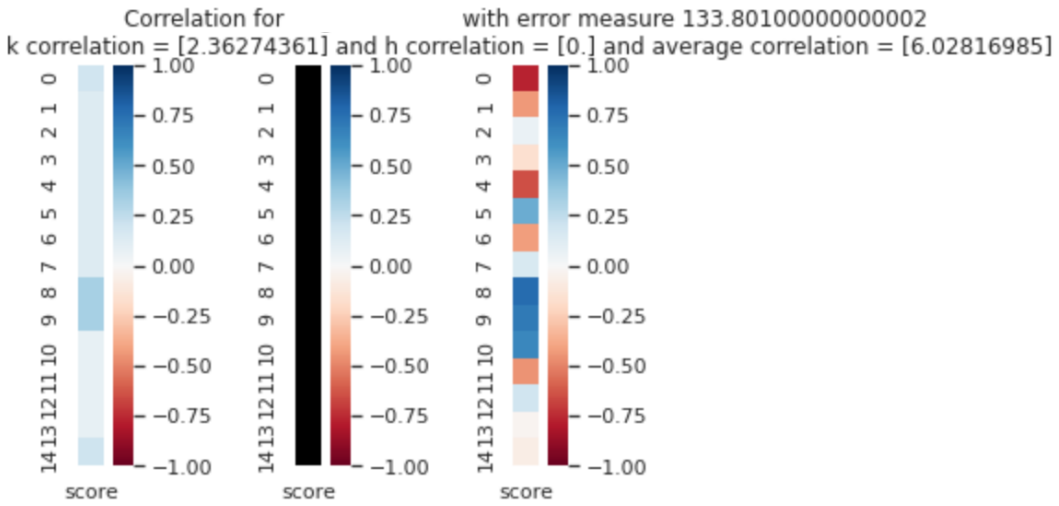

Figure 79: Correlations for patient, when minimizing BIC and the average correlation.

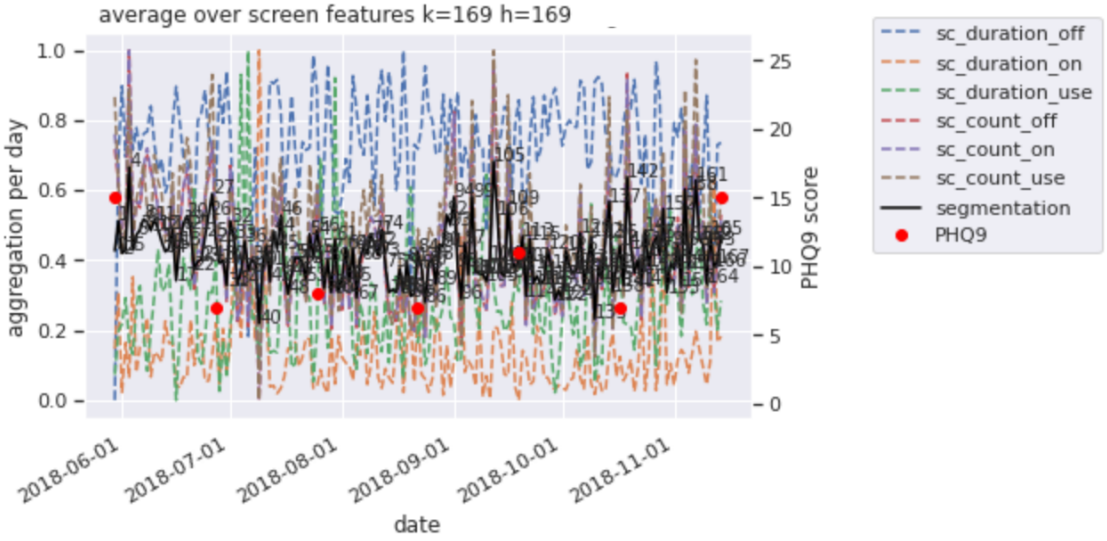

Figure 80: kh-segmentation of screen data for patient for average.

5 Conclusions and discussion

When looking at the results it is good to remember that they are not predicting behavioural changes or the mood. All methods look at the behavioural patterns and compare the changes in the behavioral patterns to the changes in the PHQ-9 score or compare the different behavioural patterns between each other to classify if the subject is a patient or a control.

5.1 Correlation

Patients showed correlations in the ranges of -1.0 to -0.5 and 0.5 to 1.0. Where as, controls had correlations only in the range of -0.25 to 0.25. This could be due to that the controls are more diverse as a group whereas patients behavioural changes are similar. Another reason could be the fact that the control group's PHQ-9 score does not vary much. Some notable results were indication of that patients surrounded by noise were feeling better than those who were not, and patients having their phone screen on for a longer duration had an improved mood. It was expected that the location data would show correlation results, but it did not. The lack of seen correlation can be due to the problems with the location data and its converter. For future work it could be interesting to see if the location data has correlations when the converter is corrected. When the main study is completed or when there is more subjects available it would be interesting to see how the separate PHQ-9 questions correlate with different sensor data. If the control group grows big enough there could be a possibility to split the controls into groups and see if the different groups phone sensor data correlates with the PHQ-9 score. This could for example be done by combining the k-means clustering and the correlation methods.

5.2 k-means clustering

It is interesting to see how the k-means clustering grouped the subjects into different groups based on different sensor data. For noise and location data, k-means clustering sorted most of the subjects into the same group. For screen data, patients were sorted to the same group as controls, but there were clearly two different groups. The communication and social application both grouped well. It can be seen as expected that the results are similar for both communication and social application as they overlap with each other in certain ways. It can be concluded that the k-means clustering method is not a good method for grouping patients and controls into different groups. However, another application of the method would be to see how k-means clustering sorts the controls into different groups. The patients are labeled, but the controls are assumed to have the same type of behaviour even-though they most probably can be divided into different groups too. This could also be applied to the patients' data to acquire subgroups.

5.3 Linear Discriminant Analysis and Decision Tree

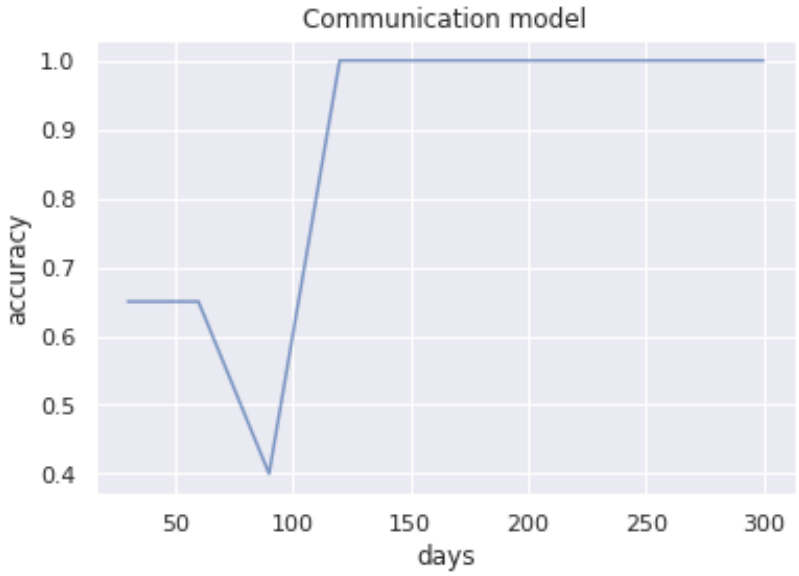

Figure 81: LDA and Decision Tree classifier mean accuracy for cross-validation for the different ranges for communication data after the rework.

Figure 82: LDA and Decision Tree classifier mean accuracy for cross-validation for all data with monthly splits after the rework.

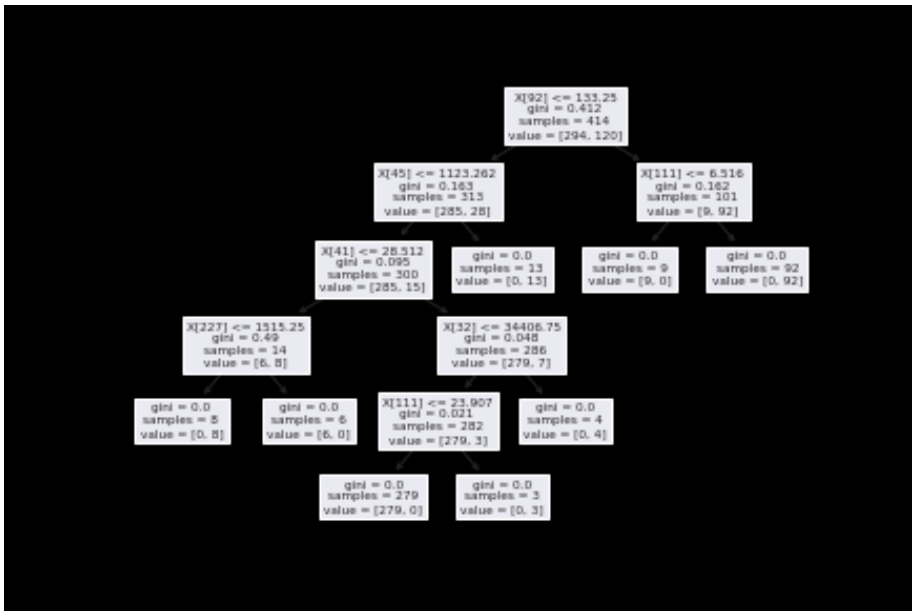

Figure 83: Decision Tree for all data with monthly splits

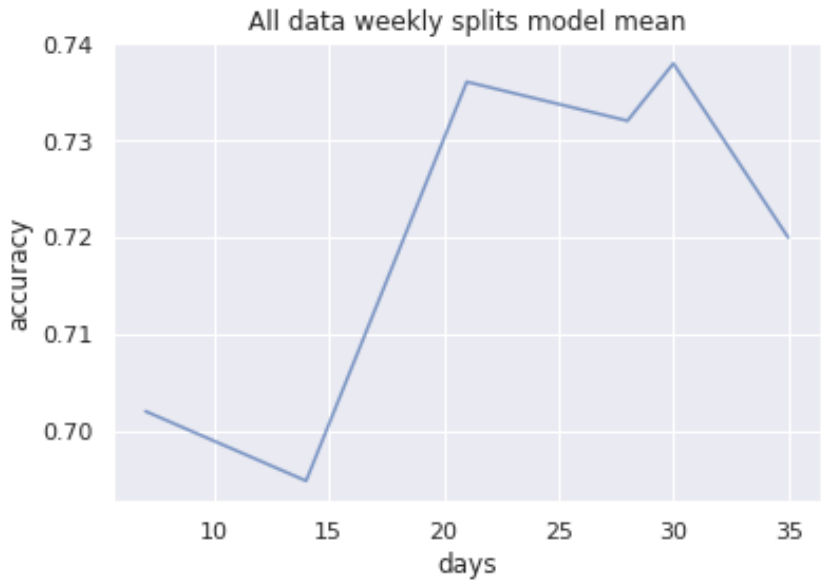

Figure 84: LDA and Decision Tree classifier mean accuracy for cross-validation for all data after the rework.

After analysing and writing the results section for the classification, it was seen that the models were performing unrealistically well. Reviewing the code again a crucial problem was found. The LDA was done on the whole dataset and not a split. This has probably created a bias in the final model and is something that must be considered in future studies. The correct way to build the model was however tested

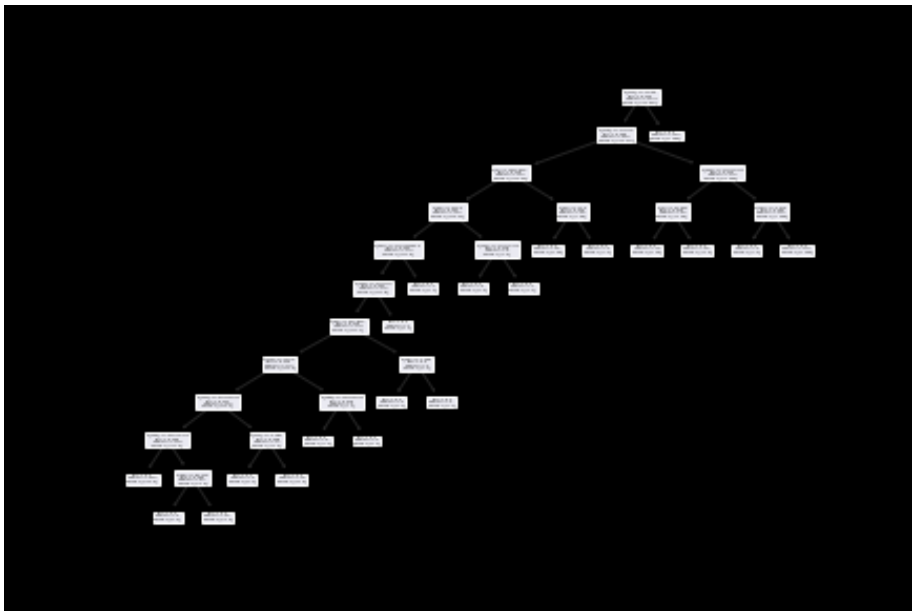

Figure 85: Decision Tree for all data with weekly splits

on the communication data, the all data monthly and all data weekly models. The accuracy for the model based on communication data dropped from an accuracy of 1 to which is seen in figure 81. This is not a bad result, on the contrary it is a more desirable result, as a perfect accuracy in the small dataset would indicate overfitting. When comparing the new trees in figures 83 and 85 to the old trees in figures 46 and 48 it is seen that there should be less underfitting, as described in Section .
For future research it will be interesting to see how the accuracy is affected by a bigger dataset. The main study also contains more cohorts than Major Depressive disorder, which suits the LDA method. In the future it is probably desirable to have models trained based on the different sensor data instead of all data in one model. This is due to the missing data in the different sensors that occur for the subjects. Regarding the tree classifier the automatic branching will probably be sufficient, but when the dataset is bigger, manual pruning could potentially improve the results.

5.4 kh-segmentation

Several conclusions can be drawn from the results. Firstly, the method seems suitable for analysing the data. It finds appealing results for the given k and h.

Secondly, kh-segmentation needs a sufficient amount of data points, with regards to both the amount of collected data and quality of data. The quality of data is discussed in Section 3.1 and the preprocessing is discussed in Section 3.2. For the kh-segmentation especially the time granularity in comparison to the amount of data points is important. For example, if the granularity is at a daily level and the collected data aggregates badly (describes several days incorrectly) due to missing

data, the method will not produce usable results. Another example is having the granularity at hourly level. In this example missing several hours of data will produce gaps in the results or unusable results. If the time span is big enough in comparison to the missing or incorrectly described days, then the kh-segmenation will just mark the bad days as a segment and if the amount of sources is chosen well the incorrect data can be marked as a different source. Meaning incorrect data has an own source number h, see figures 53 and 54. This is something that has to be taken into account in the preprocessing of the data.

When data is missing or is of poor quality the kh-segmentation will either ignore the missing data and not change the segment or the source or mark the missing data as a different state and source.

The different criteria are applicable for different cases. The PHQ-9 correlation criteria works best when there is changes in the PHQ-9 score, but it can not be used for controls or patients that have the same score in each questionnaire taken. For these cases it could be interesting to look at the changes in the separate PHQ-9 question scores.

The BIC error measure criteria had the interesting property that is tended to be optimal, at its lowest, when k and h were the same. This opens up the opportunity to let the BIC error measure get higher so that the k and the h does not have to be the same to get the most optimal result.

Comparing these criteria the PHQ-9 tend to give more granular results, with more segments, whilst the BIC error measure criteria results in less segments.

When looking at the validation of the criteria, it is seen that when maximising correlation for both k and h the correlation is worse than average or as good as average. This is due to that the criteria will be maximised when it is set to the average.

The criteria where k or h is maximised tends to set low values for both k and h. Either k or h has always a better correlation than average.

The criteria where BIC is minimised, sets low values on both k and h. The correlation is most often better than average. This criteria and maximising k or h give very similar results, where the value of h usually varies by one.

When looking at all criteria it could be beneficial to let the starting value be smaller than 10. For instance, the maximising k or h and minimising BIC criteria could get better generalization from smaller values. The value of the smallest possible state could be decided on the length of the time span, i.e. how much data has been collected for the subject. There have probably been less behavioural changes in a shorter range of time, than in a longer.

For future research it could be interesting to look at the sequences on an hourly basis instead of days. This would however drop out the daily features, for example, location, but could give another perspective of the subjects behaviour.

Another interesting question would be if we can hide and find different groups in the same data. Meaning would the different behaviours for certain groups be noticed, would they be of the same source. This could be done by combining different subject data streams to one data stream and localizing and grouping them using kh-segmentation.

When writing the results section of this thesis, a bug was found in the location data converter code. It seemed as if the longitude and latitude was calculated incorrectly dropping a whole data channel, causing the distance measuring to be incorrect. Hopefully a re-work and re-run of the code would be enough to improve results. Unfortunately the data converter is outside the scope of this thesis.

6 Summary

This thesis explored possible analysis methods for passive mobile data collected in Mobile Monitoring of Mood (MoMo-Mood) Pilot study [68]. Exploration of the sensor data was done by searching for correlations between the sensor data and the PHQ-9 score. The results was that the patients showed correlation whereas controls did not.

The k-means clustering algorithm was used for exploring how the subjects would be sorted if they were unlabeled. The patients seem to consist of a similar group, as they were sorted to the same group. The controls, however, were more diverse and could perhaps be divided into smaller subgroups. This is supported by the previous method were the patients showed correlation, but the controls did not.

The classification of patients and controls seems possible with the LDA and Decsision Tree classification method, but due to an error in the original implementation it was hard to draw conclusions from the results of the analysis. However the reworked implementation showed great initial results. This implementation could be interesting to run on the main study when possible.

The kh-segmentation method seems to find the optimal segments in the sensor data in an intuitively good way. When k and h are chosen using a suitable criteria the behavioural changes are visible in the plots. Especially marking in the plots how the state of h changes depending on the segment is useful. This shows which segments show similar behaviour. The segmentation can however be too abstract to be used in the discussions between the clinician and the patient. Further development of the layout is desirable, if it is to be used in clinical context.

This field of study shows great promise and with more data from the upcoming study, better understanding could be gained and a clinically usable tool could be developed. This would help both the clinician and the patient, and ease the burden of mental health problems.

References

[1] Official Statistics of Finland (OSF): Use of information and communications technology by individuals [e-publication]. ISSN=2341-8710. 2018. Helsinki: Statistics Finland [referred: 15.1.2020]. Access method: http://www.stat.fi/til/sutivi/2018/sutivi_2018_2018-12-04_tie_001_en.html.

[2] AWARE, Open-source Context Instrumentation Framework For Everyone. Documentation available from https://awareframework.com/. (Accessed [April 14, 2021]).

[3] koota-serverconverters code. Documentation available from https://github.com/niima-project/koota-server/blob/master/kdata/converter.py. (Accessed [April 14, 2021]).

[4] Niimpy allows you to access Koota sqlite databases easily. Documentation available from https://niimpy.readthedocs.io/en/latest/. (Accessed [April 22, 2021]).

[5] Aledavood, T. *Temporal patterns of human behavior*. Doctoral thesis, School of Science, 2017.

[6] Aledavood, T., Hoyos, A. M. T., Alakörkkö, T., Kaski, K., Saramäki, J., Isometsä, E., and Darst, R. K. Data collection for mental health studies through digital platforms: requirements and design of a prototype. *JMIR research protocols 6*, 6 (2017), e110.

[7] Aledavood, T., Kivimäki, I., Lehmann, S., and Saramäki, J. A non-negative matrix factorization based method for quantifying rhythms of activity and sleep and chronotypes using mobile phone data. *arXiv preprint arXiv:2009.09914* (2020).

[8] Aledavood, T., Lehmann, S., and Saramäki, J. Social network differences of chronotypes identified from mobile phone data. *EPJ Data Science 7*, 1 (2018), 46.

[9] Aledavood, T., López, E., Roberts, S. G., Reed-Tsochas, F., Moro, E., Dunbar, R. I., and Saramäki, J. Daily rhythms in mobile telephone communication. *PloS one 10*, 9 (2015), e0138098.

[10] Aledavood, T., Torous, J., Hoyos, A. M. T., Naslund, J. A., Onnela, J.-P., and Keshavan, M. Smartphone-based tracking of sleep in depression, anxiety, and psychotic disorders. *Current psychiatry reports 21*, 7 (2019), 1–9.

[11] Allen, D. M. The relationship between variable selection and data agumentation and a method for prediction. *technometrics 16*, 1 (1974), 125–127.

[12] Alpaydin, E. *Introduction to machine learning*. MIT press, 2020.

[13] Bandhauer, T. M., Garimella, S., and Fuller, T. F. A critical review of thermal issues in lithium-ion batteries. *Journal of the Electrochemical Society 158*, 3 (2011), R1–R25.

[14] Banos, O., Villalonga, C., Damas, M., Gloesekoetter, P., Pomares, H., and Rojas, I. Physiodroid: Combining wearable health sensors and mobile devices for a ubiquitous, continuous, and personal monitoring. *The Scientific World Journal 2014* (2014).

[15] Barabasi, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature 435*, 7039 (2005), 207–211.

[16] Bellman, R. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM 4*, 6 (1961), 284.

[17] Bidargaddi, N., Musiat, P., Makinen, V.-P., Ermes, M., Schrader, G., and Licinio, J. Digital footprints: facilitating large-scale environmental psychiatric research in naturalistic settings through data from everyday technologies. *Molecular psychiatry 22*, 2 (2017), 164–169.

[18] Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., and Mohr, D. C. Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research 13*, 3 (2011), e55.

[19] Bussemaker, H. J., Li, H., Siggia, E. D., et al. Regulatory element detection using a probabilistic segmentation model. In *Ismb* (2000), pp. 67–74.

[20] Cao, J., Truong, A. L., Banu, S., Shah, A. A., Sabharwal, A., and Moukaddam, N. Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: development and usability study. *JMIR mental health 7*, 1 (2020), e14045.

[21] Cawley, G. C., and Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research 11* (2010), 2079–2107.

[22] Ferreira, D., Kostakos, V., and Dey, A. K. Aware: mobile context instrumentation framework. *Frontiers in ICT 2* (2015), 6.

[23] Fraccaro, P., Lavery-Blackie, S., der Veer Van, S., and Peek, N. Behavioural phenotyping of daily activities relevant to social functioning based on smartphone-collected geolocation data. *Studies in health technology and informatics 264* (2019), 945–949.

[24] Gionis, A., and Mannila, H. Finding recurrent sources in sequences. In *Proceedings of the seventh annual international conference on Research in computational molecular biology* (2003), pp. 123–130.

[25] Gruenerbl, A., Bahle, G., Oehler, S., Banzer, R., Haring, C., and Lukowicz, P. Sensors vs. human: comparing sensor based state monitoring with questionnaire based self-assessment in bipolar disorder patients. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers* (2014), ACM, pp. 133–134.

[26] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[27] Hidalgo-Mazzei, D., Young, A. H., Vieta, E., and Colom, F. Behavioural biomarkers and mobile mental health: a new paradigm. *International journal of bipolar disorders 6*, 1 (2018), 1–4.

[28] Himberg, J., Korpiaho, K., Mannila, H., Tikanmaki, J., and Toivonen, H. T. Time series segmentation for context recognition in mobile devices. In *Proceedings 2001 IEEE International Conference on Data Mining* (2001), IEEE, pp. 203–210.

[29] Hollis, C., Morriss, R., Martin, J., Amani, S., Cotton, R., Denis, M., and Lewis, S. Technological innovations in mental healthcare: harnessing the digital revolution. *The British Journal of Psychiatry 206*, 4 (2015), 263–265.

[30] Insel, T. R. Digital phenotyping: technology for a new science of behavior. *Jama 318*, 13 (2017), 1215–1216.

[31] James, G., Witten, D., Hastie, T., and Tibshirani, R. *An introduction to statistical learning*, vol. 112. Springer, 2013.

[32] John, M. *A dictionary of epidemiology*. Oxford university press, 2001.

[33] Jung, A. Machine learning: Basic principles. *arXiv preprint arXiv:1805.05052* (2018).

[34] Kent, A. D., and Liebrock, L. M. Secure communication via shared knowledge and a salted hash in ad-hoc environments. In *2011 IEEE 35th Annual Computer Software and Applications Conference Workshops* (2011), IEEE, pp. 122–127.

[35] Kroenke, K., Spitzer, R. L., and Williams, J. B. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine 16*, 9 (2001), 606–613.

[36] Lee Rodgers, J., and Nicewander, W. A. Thirteen ways to look at the correlation coefficient. *The American Statistician 42*, 1 (1988), 59–66.

[37] Li, W. Dna segmentation as a model selection process. In *Proceedings of the fifth annual international conference on Computational biology* (2001), pp. 204–210.

[38] Mattingly, S. M., Grover, T., Martinez, G. J., Aledavood, T., Robles-Granda, P., Nies, K., Striegel, A., and Mark, G. The effects of seasons and weather on sleep patterns measured through longitudinal multimodal sensing. *npj Digital Medicine 4*, 1 (2021), 1–15.

[39] McLachlan, G. J. *Discriminant analysis and statistical pattern recognition*, vol. 544. John Wiley & Sons, 2004.

[40] Miritello, G. *Temporal patterns of communication in social networks*. Springer Science & Business Media, 2013.

[41] Mitchell, T. M. Machine learning and data mining. *Communications of the ACM 42*, 11 (1999).

[42] Moore, G. E., et al. Cramming more components onto integrated circuits, 1965.

[43] Obermeyer, Z., and Emanuel, E. J. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine 375*, 13 (2016), 1216.

[44] OECD, and Union, E. *Health at a Glance: Europe 2018*. 2018.

[45] Onnela, J.-P., and Rauch, S. L. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology 41*, 7 (2016), 1691.

[46] Onnela, J.-P., Waber, B. N., Pentland, A., Schnorf, S., and Lazer, D. Using sociometers to quantify social interaction patterns. *Scientific reports 4*, 1 (2014), 1–9.

[47] Organization, W. H., et al. *Promoting mental health: concepts, emerging evidence, practice: a report of the World Health Organization, Department of Mental Health and Substance Abuse in collaboration with the Victorian Health Promotion Foundation and the University of Melbourne*. World Health Organization, 2005.

[48] Pearson, K. Notes on regression and inheritance in the case of two parents proceedings of the royal society of london, 58, 240-242, 1895.

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

[50] Quinlan, J. R. Induction of decision trees. *Machine learning 1*, 1 (1986), 81–106.

[51] Raugh, I. M., James, S. H., Gonzalez, C. M., Chapman, H. C., Cohen, A. S., Kirkpatrick, B., and Strauss, G. P. Geolocation as a digital phenotyping measure of negative symptoms and functional outcome. *Schizophrenia Bulletin* (2020).

[52] Russell, S., and Norvig, P. Artificial intelligence: a modern approach.

[53] Saramäki, J., and Moro, E. From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *The European Physical Journal B 88*, 6 (2015), 1–10.

[54] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks 61* (2015), 85–117.

[55] Schwarz, G., et al. Estimating the dimension of a model. *The annals of statistics 6*, 2 (1978), 461–464.

[56] Seni, G., and Elder, J. F. Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis lectures on data mining and knowledge discovery 2*, 1 (2010), 1–126.

[57] Sharma, A., and Paliwal, K. K. Linear discriminant analysis for the small sample size problem: an overview. *International Journal of Machine Learning and Cybernetics 6*, 3 (Jun 2015), 443–454.

[58] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (2017), IEEE, pp. 3–18.

[59] Spitzer, R. L., Kroenke, K., Williams, J. B., Group, P. H. Q. P. C. S., et al. Validation and utility of a self-report version of prime-md: the phq primary care study. *Jama 282*, 18 (1999), 1737–1744.

[60] Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological) 36*, 2 (1974), 111–133.

[61] Stone, M. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological) 39*, 1 (1977), 44–47.

[62] Takayanagi, Y., Spira, A. P., Roth, K. B., Gallo, J. J., Eaton, W. W., and Mojtabai, R. Accuracy of reports of lifetime mental and physical disorders: results from the baltimore epidemiological catchment area study. *JAMA psychiatry 71*, 3 (2014), 273–280.

[63] Terzi, E., and Tsaparas, P. Efficient algorithms for sequence segmentation. In *Proceedings of the 2006 SIAM International Conference on Data Mining* (2006), SIAM, pp. 316–327.

[64] Torous, J., and Baker, J. T. Why psychiatry needs data science and data science needs psychiatry: connecting with technology. *JAMA psychiatry 73*, 1 (2016), 3–4.

[65] Torous, J., Kiang, M. V., Lorme, J., and Onnela, J.-P. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR mental health 3*, 2 (2016), e16.

[66] Torous, J., and Powell, A. C. Current research and trends in the use of smartphone applications for mood disorders. *Internet Interventions 2*, 2 (2015), 169–173.

[67] Torous, J., Staples, P., Shanahan, M., Lin, C., Peck, P., Keshavan, M., and Onnela, J.-P. Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (phq-9) depressive symptoms in patients with major depressive disorder. *JMIR mental health 2*, 1 (2015), e8.

[68] Triana, A. M., Martikkala, A., Baryshnikov, I., Heikkilä, R., Alakörkkö, T., Darst, R. K., Ekelund, J., Isometsä, E., and Aledavood, T. Mobile monitoring of mood (momo-mood) pilot: A longitudinal, multi-sensor digital phenotyping study of patients with major depressive disorder and healthy controls. *medRxiv* (2020).

[69] Zhang, D., Guo, B., and Yu, Z. The emergence of social and community intelligence. *Computer 44*, 7 (2011), 21–28.

A K-means clustering pair plots

Bellow are pairplots of the k-means clustering results for the different features for each sensor data:

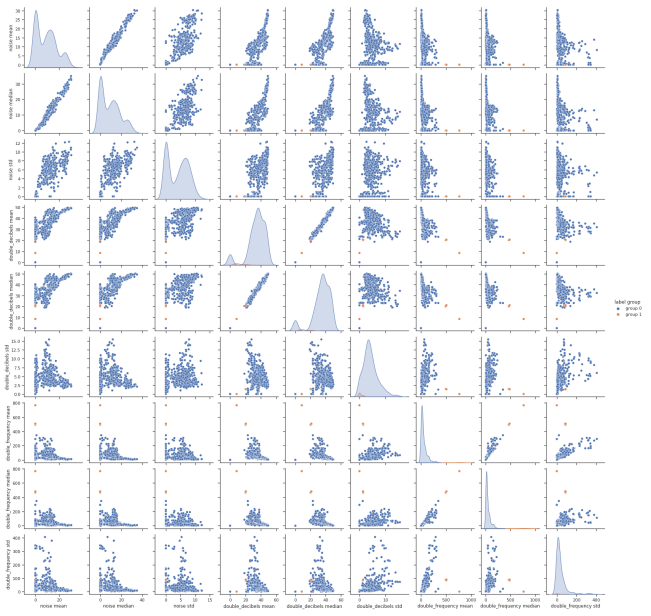

Figure A1: Pair plot of k-means clustering results for noise data.

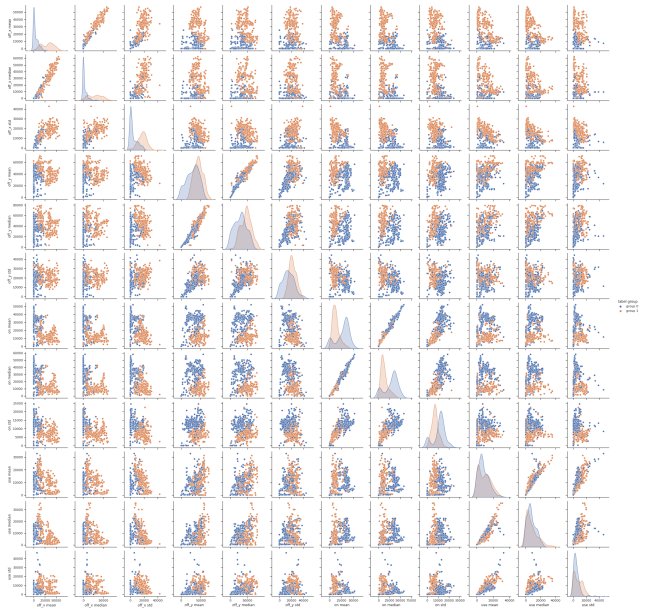

Figure A2: Pair plot of k-means clustering results for screen data.

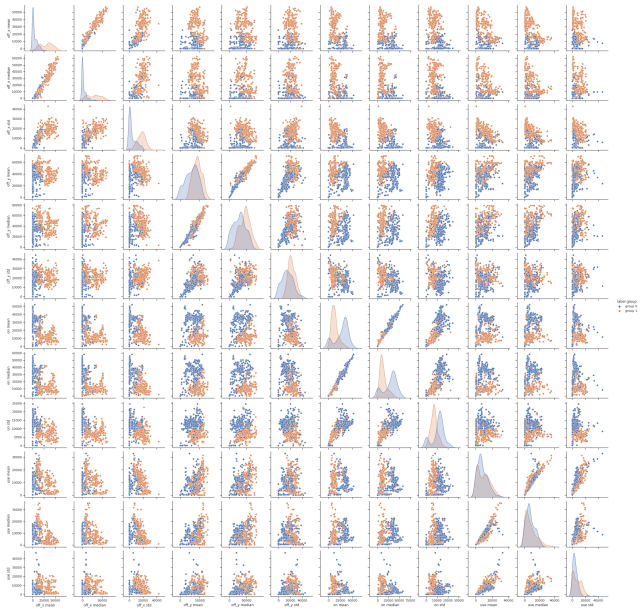

Figure A3: Pair plot of k-means clustering results for location data.

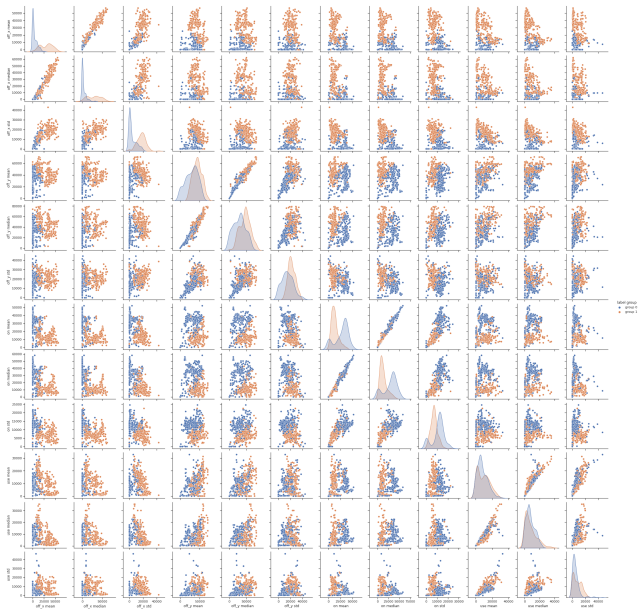

Figure A4: Pair plot of k-means clustering results for communication data.

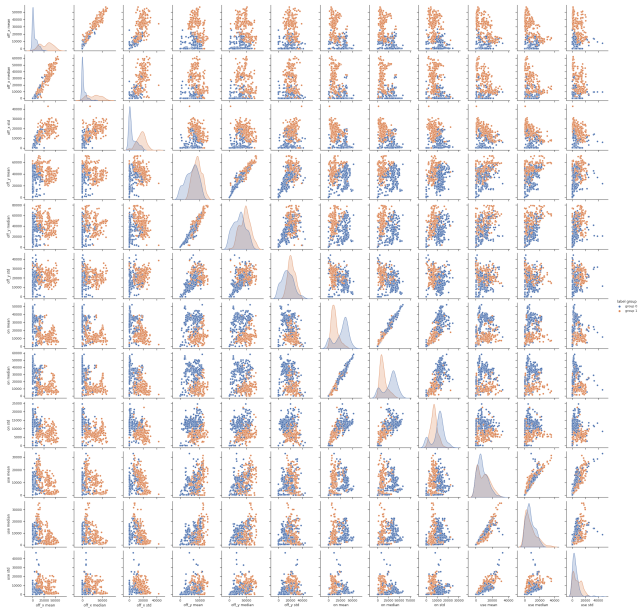

Figure A5: Pair plot of k-means clustering results for social data.