Exploring Behavioral Patterns of Patients with Mental Disorders Using the MoMo-Mood Dataset

Amirmohammad Ziaei Bideh

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.
Espoo 28.06.2022

Supervisor

Dr Talayeh Aledavood

Advisor

Dr Talayeh Aledavood
Mental disorders are major problems for people’s wellbeing in societies due to the increasing amount of stress and challenges of living in modern cities. Understanding these disorders and diagnosing them in a timely manner is crucial for people to enjoy satisfactory life quality and to function well in society. Previous studies on diagnosing mental disorders and their development over time rely on questionnaires filled by patients and visiting clinicians on a regular basis. For instance, clinicians employ The Standard for Clinicians’ Interview in Psychiatry (SCIP) to interview adult patients and diagnose a psychiatric disorder based on their answers. In recent years, technological advancements and the fact that people are using technologies like mobile phones in their daily lives provide us new opportunities for having a more realistic image of mental disorders.

However, since smartphones and digital tools have emerged only recently, their application in the mental health context calls for extensive research. The overall objective of this study is to find interpretable behavioral markers of psychiatric disorders and depressed moods in patients, using digital wearables. More specifically, this work attempts to find differences in disorder and mood levels between healthy controls and patients using features extracted from the data, their correlations, social signature, and daily rhythm analysis. To this end, this study employs the MoMo-Mood dataset, a dataset containing the digital data and mood scores (PHQ9) of 164 individuals categorized into healthy control; major depressive disorder; borderline personality disorder; and bipolar disorder. The results suggest that depressed moods are associated with a smaller but closer social network as well as higher time spent at home and reduced physical activity and variance in the movement.

Keywords Mental health, Digital phenotyping, Depression, Statistical methods, Behavioral marker
Preface

I would like to thank Talayeh Aledavood for her invaluable support throughout all the phases of writing this thesis, Erkki Isometsä for his constructive feedback, Arsi Ikäheimonen; Ana Triana Hoyos; Sara Heydari; and Nguyen Luong for their helpful discussions, Richard Darst for his technical support and helpful advice, Mikko Kivelä and Markus Heinonen for supervising the interesting research projects I have done during my studies, and Pekka Orponen for arranging the Doctoral Track program and providing the opportunity to be involved in multiple research projects.

We acknowledge the computational resources provided by the Aalto Science-IT project.

Otaniemi, 27.06.2022

Amirmohammad Ziaei Bideh
Contents

Abstract 3
Preface 4
Contents 5
Symbols and abbreviations 7

1 Introduction 8
  1.1 Focus and goals ............................................. 8
  1.2 Contributions ............................................ 9
  1.3 Organization ............................................ 9

2 Background 11
  2.1 Mental disorders ........................................... 11
  2.2 Digital phenotyping ......................................... 11
  2.3 Digital phenotyping studies ................................. 12

3 Dataset 14
  3.1 Data collection ............................................ 14
  3.2 Dataset properties ......................................... 15
    3.2.1 Message .............................................. 15
    3.2.2 Call ............................................... 15
    3.2.3 Location .......................................... 16
  3.3 Location preprocessing .................................... 17
    3.3.1 Removing (0, 0) points .............................. 19
    3.3.2 Keeping only gps data points ....................... 19
    3.3.3 Down sampling (binning) ............................ 22

4 Research material and methods 23
  4.1 Features ................................................ 23
    4.1.1 Location ........................................... 24
    4.1.2 Message ........................................... 25
    4.1.3 Call .............................................. 25
  4.2 Feature Analysis ........................................... 26
    4.2.1 Welch’s t-test ..................................... 26
    4.2.2 Comparing groups ................................... 27
  4.3 Correlation Analysis ...................................... 27
  4.4 Signature Analysis ....................................... 29
    4.4.1 Social Signature ................................... 29
    4.4.2 Location Signature ................................ 30
  4.5 Rhythm Analysis .......................................... 30
5 Results

5.1 Feature Analysis
   5.1.1 Messages
   5.1.2 Call
   5.1.3 Location

5.2 Correlation Analysis

5.3 Signature Analysis
   5.3.1 Social Signature in calls
   5.3.2 Social Signature in messages
   5.3.3 Location Signature

5.4 Rhythm Analysis
   5.4.1 Group comparison
   5.4.2 Relationship between change in PHQ9 and change in rhythm

6 Conclusions and discussion

6.1 Feature analysis
6.2 Correlation analysis
6.3 Signature analysis
6.4 Rhythm analysis
6.5 Limitations
6.6 Future work

References

A Appendix
Symbols and abbreviations

Abbreviations

WHO  World Health Organization
ICD  International Classification of Diseases
DSM  Diagnostic and Statistical Manual of Mental Disorders
mdd  Major Depressive Disorder
bpd  Borderline Personality Disorder
bd  Bipolar Disorder
PHQ9  Patient Health Questionnaire-9
GPS  Global Positioning System
CDF  Cumulative distribution function
JSD  Jensen-Shannon Divergence
SSD  Social Singnature Distance
LSD  Location Singnature Distance
SP  Significant Place
MMM  MoMo-Mood
1 Introduction

Mental health is a crucial part of people’s well-being, allowing them to have a fulfilling life and to succeed in their careers, relationships, and society. In the EU, over one in six people were affected by a form of mental disorder in 2016 [1]. In Finland, close to 20% of the population is influenced by mental disorders, which is the highest rate in the EU [1].

Mental health disorders decrease the quality of life of patients and their families [2]. It is estimated that up to 90% of people who died from committing suicide suffered from some form of mental disorder [3, 4]. Compared to the general population, those who suffer from mental disorders have a decreased life expectancy of 10 - 15 years [5, 6].

Despite the serious life challenges patients face, there are effective treatments for mental disorders [7]. Mood and anxiety disorders, which are some of the most prevalent mental disorders, could be treated with psychotherapy, pharmacotherapy, or a combination of both [8].

The traditional forms of mental health treatments involve visiting mental health practitioners and filling out surveys and self-reports [9]. However, several challenges are in effect in the traditional treatment forms [10]. First, it cannot be considered a persistent approach for monitoring mental disorders because patients might forget their previous states or lose motivation to answer the questions [9]. Second, it is subjective and biased toward the state of the individual at the time of visiting the clinician or answering questions, which might not reflect the state of the patient during the previous days and weeks [9]. Third, it is an active method meaning that the person needs to devote time and money to commute and visit the practitioner and also actively answer the surveys [9]. Therefore, more efficient methods for monitoring and diagnosing mental disorders are needed.

The widespread usage of digital devices has created new opportunities for mental health. According to a study conducted in 2018, more than 90% of Finns aged 16-54 have a personal smartphone in use [11]. The built-in sensors in smartphones and smartwatches, including GPS; Bluetooth; Wifi; accelerometer; and app usage logs to name a few, provide new possibilities to understand and monitor mental disorders. Location (GPS) data captures mobility patterns and physical activities, app usage data reflects user interests, and call and message logs show communication patterns. It has been shown that this information is linked to mental disorders symptoms [12].

The way digital devices collect data resolves the three issues regarding traditional treatment and monitoring methods. Since the data collection process does not require active engagement from the patient’s side, and it is always happening when the patient carries the device, this method is persistent and active. Besides, all the data come from sensors, allowing to capture data in an objective manner.

1.1 Focus and goals

This thesis attempts to explore the possibility of leveraging behavioral data captured from digital devices to monitor mental disorders and to understand patient’s
behavioral patterns. This work focuses on two broad goals:

1. Finding behavioral markers which differentiate healthy people from those suffering from mental disorders and patients’ disorder types. In other words, in what ways do people with different disorder types behave differently? Three types of mental disorders are investigated in this study: Major Depressive Disorder, Borderline Personality Disorder, and Bipolar Disorder, all of which are prevalent mental issues (more details are discussed in the next section).

2. Understanding how depressed moods are reflected in individuals’ behaviors and discovering related behavioral markers. People, either healthy or with mental disorders, experience fluctuations in their moods throughout their lifetime. The changes in mood could be subtle or drastic. The behavioral markers can help to identify alterations in mood level and address them in a timely manner. Also, it might give insights into the causes of depressed moods.

1.2 Contributions
To achieve the aforementioned goals, a number of contributions are made by this work:

- A representative dataset of digital data from healthy subjects and people with mental disorders along with their depressed mood scores is preprocessed and analyzed.

- A number of behavioral features from location, call, and message raw data are computed. This enables us to study the individuals’ behaviors.

- The code for preprocessing and extracting location features is integrated into an open-source software package called niimpy [13].

- Each of the computed features is examined to find those which differentiate control people from patients.

- Relationship between behavioral features and mood scores is investigated in order to find behaviors that reflect the mood.

- Difference in Social Signature and Location Signature of subjects belonging to different patient groups is studied.

- Daily and weekly rhythm of subjects along with the link between change in rhythm and change in mood is analyzed.

1.3 Organization
The rest of this thesis is organized as follows: Chapter 2 provides relevant background information on mental disorders, digital phenotyping, and statistical methods for studying mental health data. Chapter 3 gives an overview of dataset properties and
preprocessing steps done on the location data. Chapter 4 discusses in detail the way behavioral features are extracted from raw data and how statistical methods are employed in this work. Chapter 5 presents the findings and their implications. Finally, Chapter 6 discusses the conclusions made from the results along with a discussion of limitations of this work and possible future works.
2 Background

2.1 Mental disorders

Mental Disorders or Psychiatric Disorders are characterized by a cluster of behavioral or emotional symptoms [12]. Two widely established references for diagnostics of mental disorders are available: 1) the Diagnostic and Statistical Manual of Mental Disorders (DSM) [12] which is published by American Psychiatric Association and 2) the International Classification of Diseases (ICD) [14] which is published by the World Health Organization (WHO).

According to DSM-5, a mental disorder is defined as “a syndrome characterized by clinically significant disturbance in an individual’s cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning” [12]. Mood disorders [15] and personality disorders [16] are among the most prevalent groups of mental disorders. In the following, some of the most common examples of these disorders are discussed.

Major Depressive Disorder (mdd) is among the most prevalent mood disorders in youths [12]. People who suffer from mdd share similar symptoms over various periods of time (at least 2 weeks) including depressed mood, diminished interest in activities, sleep disturbances, fatigue or loss of energy, low mood, feeling of helplessness or hopelessness, thoughts and feelings of worthlessness, and diminished concentration. These symptoms manifest themselves in verbal and nonverbal behaviors [12].

Bipolar Disorder (bd), previously called maniac depressive illness, is a mood disorder characterized by recurrent periods of depression and abnormally-elevated states of happiness and high energy [17, 12]. It has adverse influences on the lives of patients, leading to a high suicide rate of 6% over a period of 20 years among the patients. In a meta-analysis study, it has been found that Suicide Attempt (SA) rate in bd patients is 30-40% [18].

Borderline Personality Disorder (bpd) is a complex personality disorder characterized by instability in several life domains, including interpersonal relationships, emotions, and self-image [19, 12]. Among those affected, up to 10% have committed suicide, a rate 50 times higher than the general population [20].

2.2 Digital phenotyping

Traditional methods of diagnosing mental disorders are mainly based on periodic self-assessments using questionnaires and clinical interviews. These methods suffer from a handful of disadvantages. For instance, retrospective recall of psychiatric symptoms over weeks or even months tends to be subjective [21] and is prone to recall inaccuracies and reporting biases [22, 23]. Therefore, a more objective and accurate tool for assessing mental health signals is needed.

Digital devices such as smartphones and fitness trackers offer us great possibilities for overcoming the mentioned challenges. Their embedded sensors can be repurposed to passively track subjects’ behaviors and to monitor the onset or progression of mental disorders [24]. Digital Phenotyping is defined as the “moment-by-moment
quantification of the individual-level human phenotype in situ using data from personal digital devices”. Digital phenotyping involves collecting various behavioral data from individuals, including spatial trajectories using GPS; social interactions using call, text messages, and Bluetooth; mobility patterns using accelerometer; and sleep patterns using health trackers [25].

The data collected through digital devices can be categorized into two groups: 1) **passive** data which does not require active participation of the individual (e.g. location data or call logs) 2) **active** data which requires active participation (e.g. mood surveys). Raw sensor data is not of use in itself and needs to be processed to gain useful knowledge about the individuals’ state pertaining to behaviors, and clinical disorders [26].

### 2.3 Digital phenotyping studies

Many studies have leveraged the extracted features and behavioral markers using statistical methods to come up with new insights about individuals’ behavioral patterns.

Chow et. al. in [27] have leveraged GPS data from smartphones to examine the association between time stayed at home and self-reported measures of social anxiety. This hypothesis was motivated by the claim that people with social anxiety withdraw to participate in social activities. Time spent at home is used as a proxy for quantifying social isolation. Studied on 72 college students, they found out that high social anxiety is associated with more time spent at home ($p = 0.007$).

Wang et. al. in [28] studied the correlation between self-reported measures of depression mood using PHQ9 score and objective measures from various mobile sensors including GPS and accelerometer to name a few. They concluded that sleep duration as well as number of encounters with other people are negatively correlated with PHQ9 score; with correlation coefficients $r = -0.382$ and $r = -0.362$, respectively.

In yet another study [24] conducted by Ben-Zeev et. al., the association between levels of stress and different behavioral markers is investigated. Using Mixed-effects linear models, they conclude that geospatial activity, sleep duration, and variability in geospatial activity are significantly associated with stronger symptoms of daily stress levels (all p-values < 0.05).

Saeb et. al. in [29] have investigated the relationship between depression levels (measured by PHQ9 survey) and data collected from smartphones. They found that higher number of interactions with phone and greater amount of time working with phone are correlated with higher depression scores. On the other hand, higher location variance and location entropy are negatively correlated with depressed moods. This finding is in line with the statement that depressive symptoms are generally associated with reduced social and physical activity. This study was performed on a small sample size of 28 individuals. The authors replicated their findings in another study ([30]) with a new, bigger sample population.

In [21], Chikersal et. al. made use of Machine Learning methods for detecting depression and the changes in the severity of depression long before its onset. They collected phone-derived data from sensors such as Bluetooth, GPS, Call, and mes-
sages with AWARE framework. Also they collected sleep data and step counts of participants using Fitbit tracker. Using Nested Randomized Logistic Regression, they selected an optimal subset of the numerous features they calculated from the sensor data. Finally, they applied two ML methods, Logistic Regression and Gradient Boosting Classifier, to achieve a best accuracy of 85.7% in detecting post-semester depressive symptoms and 85.4% in predicting change in severity of depression.

Abdullah et. al. tried to leverage the data from smartphone sensors of 7 subjects diagnosed with bipolar disorder to predict the Social Rhythm Metric (SRM) [31]. SRM is an established marker for bipolar disorder, which quantifies the stability of individuals’ social activities [32]. Using Support Vector Regression, they could achieve a Root Mean Square Error of 1.40 in predicting SRM score. Also, they utilized recursive feature elimination [33] to discover the most relevant features for the prediction task. They found that the distance traveled and the number of location clusters were the features with the most importance.

Jakobsen et. al. applied a number of Machine Learning and Deep Learning methods for classifying subjects into healthy or depressed groups based on motor activity data collected from actigraph [34]. Using Random Forrest, Deep Neural Network, and Convolutional Neural Network algorithms, they could obtain a maximum sensitivity of 0.82 and specificity of 0.84.
3 Dataset

3.1 Data collection

The dataset used by this study is from MoMo-Mood (MMM) study conducted in a collaboration between Helsinki University Central Hospital, Aalto University, and University of Helsinki. Up until now, a few research projects have studied the pilot version of MMM with 37 participants divided into two groups: healthy controls and mmd [35, 36]. However, the current dataset contains 164 participants to 4 groups: healthy control (N=31), mdd (N=85), bpd (N=27), and bd (N=21). Non control subjects are considered as patient group for the rest of this thesis. The participants were asked to remain in the study for up to a year, but they could drop out at any point they wanted. The duration of participation and availability of data for each subject vary hugely, ranging from a few days to a year. The participants were recruited on a rolling basis, meaning that they have entered in and exited from the study in different times.

During the first two weeks of participation, called active phase, subjects were required to answer daily questions about their mood. The data collection continued after the active phase, called passive phase, with subjects answering psychological surveys such as PHQ9.

Subjects were given two digital devices to have with them during the active phase: a bed sensor and an actigraphy device. The bed sensor, ballistocardiography-based Murata SCA11H nodes), recorded the acceleration of the bed, heart rate, respiration rate, stroke volume, and signal strength at the frequency of 1Hz. Actigraphs are wrist-worn devices which measure subjects’ sleep and activity levels. The actigraphy device used for data collection was Philips Actiwatch 2. Moreover, an application called AWARE [37] was installed on their smartphone to capture data from various built-in sensors, namely location, lock and unlock events, call and message logs, screen, battery status, application usage, and ambient noise. The subjects kept the application installed during the passive phase.

AWARE is an open source mobile instrumentation toolkit that provides a platform for capturing data from built-in sensors in mobile phones. It ensures data security and privacy by encrypting data via a one-way hashing of personal information and using secure connections between user client and dashboard [37]. Also, it enables researchers to manage multiple studies simultaneously and allows users to visualize their own personal information. These features, however, were not employed in the data collection process.

The dataset is created using a data collection platform, Non-Intrusive Individual Monitoring Architecture (NIIMA) [38], which is a platform for managing and gathering behavioral data. This platform takes into account three key design features: 1) flexibility of access control, to allow conducting simultaneous independent studies with the same users and mixing data with considering users’ privacy, 2) flexibility of data sources, to allow gathering data from various devices and seamless integration, and 3) first-order privacy protection, ensuring data privacy and preventing researchers from accident violations of privacy regulations.
3.2 Dataset properties

In this study, we focus on three data sources: message, call, and location data. In the following, the format of these data types along with the number of data points for each group is presented.

3.2.1 Message

Messages are referred to SMS (Short Message Service) which are short text messages that are exchanged between phones as a method for asynchronous communication. This data does not include the messages transferred in social media applications. The Table 1 shows the message data scheme. The Table 2 presents the number of messages people in each group have sent or received.

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>user</td>
<td>Anonymized user id</td>
</tr>
<tr>
<td>device</td>
<td>Device id from which the communication is happened</td>
</tr>
<tr>
<td>time</td>
<td>Timestamp</td>
</tr>
<tr>
<td>message_type</td>
<td>Type of the message. Could be incoming or outgoing</td>
</tr>
<tr>
<td>trace</td>
<td>Source/target of the message</td>
</tr>
</tbody>
</table>

Table 1: Properties of message data

<table>
<thead>
<tr>
<th>group</th>
<th>number of incoming messages</th>
<th>number of outgoing messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>6319</td>
<td>3887</td>
</tr>
<tr>
<td>mdd</td>
<td>12444</td>
<td>6518</td>
</tr>
<tr>
<td>bpd</td>
<td>3134</td>
<td>1186</td>
</tr>
<tr>
<td>bd</td>
<td>5685</td>
<td>3592</td>
</tr>
<tr>
<td>all</td>
<td>27582</td>
<td>15183</td>
</tr>
</tbody>
</table>

Table 2: Statistics of message dataset.

3.2.2 Call

Calls data refer to the phone call events which are synchronous voice communication between people. This data does not include the calls made through applications. The Table 3 shows the call data scheme. The Table 4 presents the number of calls happened in each group.
### Table 3: Properties of call data

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>user</td>
<td>Anonymized user id</td>
</tr>
<tr>
<td>device</td>
<td>Device id from which the communication is happened</td>
</tr>
<tr>
<td>time</td>
<td>Timestamp</td>
</tr>
<tr>
<td>call_type</td>
<td>Type of the call. Could be incoming, outgoing, or missed</td>
</tr>
<tr>
<td>call_duration</td>
<td>Duration of the call in seconds</td>
</tr>
<tr>
<td>trace</td>
<td>Source/target of the call</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>group</th>
<th>number of incoming calls</th>
<th>number of missed calls</th>
<th>number of outgoing calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>4710</td>
<td>1686</td>
<td>6974</td>
</tr>
<tr>
<td>mdd</td>
<td>9765</td>
<td>3604</td>
<td>10999</td>
</tr>
<tr>
<td>bpd</td>
<td>1308</td>
<td>986</td>
<td>2205</td>
</tr>
<tr>
<td>bd</td>
<td>2278</td>
<td>961</td>
<td>2926</td>
</tr>
<tr>
<td>all</td>
<td>18061</td>
<td>7237</td>
<td>23104</td>
</tr>
</tbody>
</table>

Table 4: Statistics of call dataset.

### 3.2.3 Location

The geographical position of the subjects is recorded in the form of location data. Timestamped latitude, longitude, and speed of the individuals are the main properties of location data which is employed in the next analysis. The Table 5 shows the location data scheme. The Table 6 presents the number of location bins recorded in each group.

### Table 5: Properties of location data

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>user</td>
<td>Anonymized user id</td>
</tr>
<tr>
<td>device</td>
<td>Device id with which the location is captured</td>
</tr>
<tr>
<td>time</td>
<td>Timestamp</td>
</tr>
<tr>
<td>double_latitude</td>
<td>Latitude</td>
</tr>
<tr>
<td>double_longitude</td>
<td>Longitude</td>
</tr>
<tr>
<td>double_altitude</td>
<td>Altitude</td>
</tr>
<tr>
<td>accuracy</td>
<td>Accuracy of the location in meters</td>
</tr>
<tr>
<td>double_speed</td>
<td>Speed (m/s) of the user</td>
</tr>
<tr>
<td>provider</td>
<td>The source of location; could be gps, network, or fused (a combination of both)</td>
</tr>
<tr>
<td>label</td>
<td>Shows if the location service of the phone is enabled or not; could be enabled or disabled or empty</td>
</tr>
</tbody>
</table>

Table 5: Properties of location data
3.3 Location preprocessing

Location data can be originated from various sources including Global Positioning System (gps) or close wireless network access points (network), each of which have their own advantages and disadvantages [39]. When the location is requested from the user’s phone, the operating system decides which source to use for finding an approximate of the phone location based on the system configuration and available resources.

Real-world data is often not clean and has a lot of missingness and location data is not an exception. Missingness in data leads to a lack of information about participants and consequently might prevent us from understanding their behavioral patterns. The location data could be missing due to different reasons. One of them is technical issues, which might be because app/phone stopped functioning or the server crashed. Another source of data missingness is semantic reasons; for example, a participant might not make or receive a call during a whole day. This kind of missingness proves to be useful for us as they contain meaningful information about the participant’s behaviors. There are roughly two approaches for imputing missing location data. The first approach is using heuristics to fill the missing gaps. For instance, in [40], if the missing interval is less than two hours and the distance between two sides of the interval is less than 500 meters, they fill the missing interval with the average of the coordinates of the missing interval ends. Modeling location trajectories is another approach. The authors in [41], for example, introduced a bidirectional location imputation method using online Gaussian Process. Figure 1 provides a big picture overview of missingness in our dataset.

Besides missingness, there might be a lot of noise and outlier data points in the datasets that need to be addressed. For instance, most of the time the coordinates captured by phone GPS service do not exactly correspond to the user’s location and it might have some error. The presence of data missingness or noise data has been observed in previous studies as well. For instance, [29] excluded 22 participants out of 40 from their dataset because of insufficient GPS location data. In another study, authors reported that in 31% of days there were not enough data recorded for each participant [42]. Therefore, before working on the raw data, we should perform a cleaning or preprocessing step to remove low-quality data points. The steps we perform for preprocessing location data are discussed in the following sections. Note that the number of data points in the dataset after each preprocessing step is shown in the Table 7.

<table>
<thead>
<tr>
<th>group</th>
<th>number of bins</th>
</tr>
</thead>
<tbody>
<tr>
<td>controls</td>
<td>101097</td>
</tr>
<tr>
<td>mdd</td>
<td>386965</td>
</tr>
<tr>
<td>bpd</td>
<td>81529</td>
</tr>
<tr>
<td>bd</td>
<td>81243</td>
</tr>
<tr>
<td>all</td>
<td>650834</td>
</tr>
</tbody>
</table>

Table 6: Statistics of location dataset.
Figure 1: Daily Missingness plot of location data. Rows are users and columns are the day starting from user’s participation date. White days show that there was at least one GPS data in that day, red days show that there were no GPS but at least one network data in that day, and black days are days without any location data recorded.

Table 7: Number of data points in each group after each preprocessing step.

<table>
<thead>
<tr>
<th>Group</th>
<th>All</th>
<th>disabled removed</th>
<th>(0,0)s removed</th>
<th>non GPS removed</th>
<th>Binned</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>1,973,250</td>
<td>1,962,902</td>
<td>1,957,435</td>
<td>810,220</td>
<td>180,339</td>
</tr>
<tr>
<td>mdd</td>
<td>5,442,343</td>
<td>5,412,469</td>
<td>5,402,509</td>
<td>2,564,333</td>
<td>650,554</td>
</tr>
<tr>
<td>bpd</td>
<td>2,608,618</td>
<td>2,582,849</td>
<td>2,582,594</td>
<td>1,14,3012</td>
<td>142,742</td>
</tr>
<tr>
<td>bd</td>
<td>1,322,007</td>
<td>1,315,470</td>
<td>1,311,173</td>
<td>614,398</td>
<td>181,561</td>
</tr>
<tr>
<td>Total</td>
<td>11,346,218</td>
<td>11,273,690</td>
<td>11,253,711</td>
<td>5,131,963</td>
<td>1,155,196</td>
</tr>
</tbody>
</table>
3.3.1 Removing (0, 0) points

Visualizing data points collected by one of researchers on a world map, we realized there are some strange locations recorded in the middle of Atlantic Ocean. The coordinates of those locations were (latitude=0, longitude=0). Since the mentioned researcher has not been in that place we concluded such data points are noise. The (0, 0) points existed in the MMM dataset for many participants and contributed to less than 1% of all data points, but neglecting them could affect heavily on subsequent results. Therefore, these data points are removed from the dataset.

3.3.2 Keeping only gps data points

To further assess the quality of data, we looked at the velocity of users. It is expected that the speed of the user should be within a reasonable range. A histogram of velocities of a user in bd group is shown in Figure 2. There are some velocities of over $10^7 m/s$ which is way higher than the speed of the airplane. This suggests that there is a serious anomaly in the dataset. By looking at the pair of points with high velocity, we noticed the location provider of all of them is network. Therefore, we remove all of the network-derived locations. The reason for such an anomaly needs further research. As it can be seen in Figure 2, the range of speeds after this cleaning stage is more reasonable. As a reference, the speed of the airplane is approximately 300 m/s.

As a final step of data quality check, we plot the histogram of daily total distance of users for each preprocessing step in Figure 3. The first row is for original data points which includes zero points and network points. Traveling for $10^8 m = 100,000 km$ in a day does not make sense at all. In the middle row, zeros are removed but network points are still present. Finally, the last row corresponds to the final preprocessing stage in which both zero and network points are removed. For sanity check, we plotted part of the travel path of a user with the highest distance traveled and it turned out that they had a flight on that day (look at Figure 4).
Figure 3: Histogram of daily total distance traveled for all users.
Figure 4: A part of travel path of a sample participant.
3.3.3 Down sampling (binning)

To further reduce noise, we down sample location data into 5-minute bins. The \textit{latitudes} and \textit{longitudes} of locations within 5-minute intervals are aggregated by taking the median of them and the result is assigned to that bin. After this stage, we end up having 1,155,196 bins in the dataset which is 9,096 bins per person. There are two main benefits of downsampling data. The first is to reduce the noise in data by ignoring outlier and error-prone locations in the intervals. Since the binning interval is relatively short, we do not lose many behavioral patterns. The second benefit is to reduce the computational and memory overhead for the subsequent analysis. Computing some of the location-related features, for instance, clustering static points, is computationally heavy, and computing them over the entire dataset without down sampling might take too much time. Further analysis of location data is done on the down-sampled data.
4 Research material and methods

4.1 Features

In this study, we focus on three types of data that are captured from participants’ mobile phones: call, message, and location GPS data. Call and message data quantify communication patterns with their social network members. Location data can be used to investigate the mobility patterns, physical movements and social isolation of individuals. For example, time spent at home can be seen as a measure of social isolation and variance in the locations visited by a subject shows an aspect of mobility pattern.

The subjects in our dataset have participated in the study in different time periods. Besides, the duration of time that each person has provided data varies hugely from person to person. Therefore, it is crucial how to design the feature extraction scheme.

One of the goals of this analysis is to understand the relation between behaviors and mood which is measured with PHQ9 scores. PHQ9 (Patient Health Questionnaire) is a 9-question survey that assesses the presence and severity of depression in individuals [43]. The questions in PHQ9 survey are meant to be answered based on how the person has been in the past two weeks. Therefore, it makes sense to look at the data from the 14 days before PHQ9 is answered until the survey day. Another advantage of this data splitting is that the length of time periods in which features are captured is equal for every data instance, which makes the task of feature comparison independent of time duration.

We extract features from 14-day periods. To be more specific, for each person we first take the dates in which they have responded to PHQ9 surveys. Then we select the data from 14 days prior to their answers up to the answer day and extract features on the data recorded within that time range. We did a literature review of the past work and picked the most common used features and we designed a number of new features which we found helpful. Note that with this approach a person might have potentially more than one feature instance. Take a look at Figure 5 for an illustration of this approach.

Figure 5: A schematic representation of how feature extraction is done on 2-week periods.
4.1.1 Location

The main properties of the data that are used for feature extraction are timestamp of the record, latitude, longitude, and speed. To capture this information from data, we implement a number of features [21, 44, 45, 40, 46, 29]:

- **dist_total**: Total distance traveled in meters. This feature is computed by calculating the summation of distances between all consecutive location data points. Euclidean distance cannot be used here and instead we use great-circle distance between two points on a sphere. The distance between the two location data points is computed as

  \[ d = R \cos^{-1}[\cos \theta_1 \cos \theta_2 \cos(\gamma_1 - \gamma_2) + \sin \theta_1 \sin \theta_2], \]

  where \( R \) is the radius of Earth, \( \theta_1 \) and \( \theta_2 \) are the latitudes, \( \gamma_1 \) and \( \gamma_2 \) are the longitudes of the locations.

- **variance**: sum of the variance in latitudes and longitudes.

- **log_variance**: logarithm of variance.

- **speed_average, speed_variance, speed_max**: statistics of one’s speeds in meters per second. In niimpy, if the speed column is not present in the dataset, it is computed by dividing the distance between two locations by the time elapsed between the two samples.

Based on the speed of the subject at any given location, we label them as either **moving** or **static** (stationary) [30]. **moving** locations are the moments in which the speed is over 0.27 m/s, showing that the person is traveling to a destination. The rest of the locations are considered **static**.

Following the approach of [21], the **static** samples are clustered using DBSCAN algorithm [47]. The reasons for choosing DBSCAN as the clustering algorithm are 1) it does not require the number of clusters to be pre determined and 2) it labels outlier points as -1 that can represent the places the person rarely visits. The resulting clusters are called Significant Places since they show the places that the person visits frequently. They could, for example, be their home, workplace, friend’s or family’s house, or their favorite restaurant.

Two parameters specify how DBSCAN clusters the samples. The first one is \( \epsilon \) which is the maximum distance between two samples to be considered as neighbors in one cluster. This parameter is set to 200 in our analysis. The second important parameter is the minimum size of a cluster. The size of a cluster is the number of samples in it. This parameter is set to 5. These parameters are set based on clustering a subset of data and manually adjusting the parameters to get the most sensible clustering.

After finding significant places, several features are extracted:

- **n_static, n_moving**: fraction of **static** and **moving** bins.
• **n_sp**: number of significant places. This is equal to the number of clusters found by DBSCAN.

• **n_rare**: fraction of time visiting rarely visited places (outlier clusters).

• **n_transitions**: number of transitions between significant places.

• **entropy**: entropy of time spent in different significant places. Higher entropy shows that the person has spent time in the significant places more evenly and the stay time distribution is less skewed.

• **normalized_entropy**: calculated as entropy divided by the number of significant places. This is done so that the effect of number of significant places on the entropy is reduced.

We consider the home of a person to be the place the person has been mostly between midnight and 6 AM. In order to find home, we cluster all the samples which are recorded in the aforementioned time period. Then, the center of the biggest cluster is considered as home. When a sample location is within the radius of 50m of the home, we consider the person to be at home. Based on these information we collect two features:

• **n_home**: fraction of time spent at home.

• **max_dist_home**: maximum distance from home in meters.

### 4.1.2 Message

The features that can be potentially extracted form messages are limited since we do not have access to the content of the text messages. The main data properties used for feature extraction are timestamp of the messages and the ID of the person who is contacted. The list of features are as follows [46]:

• **incoming, outgoing**: number of messages sent and received.

• **inout_ratio**: ratio of incoming to outgoing messages.

• **unique_correspondent** number of unique contacts that the person has messaged to or received message from.

### 4.1.3 Call

Call data is similar to message data in that it is a one to one communication tool. The main data properties that are used for feature extraction are timestamps the calls, ID of the source/destination of the call, type of the call, and call duration. The call features we have implemented are [46, 21]:

• **incoming, outgoing, missed**: number of incoming, outgoing, and missed calls.
• **inout_ratio**: ratio of number of incoming to outgoing calls.

• **unique_correspondent**: number of unique contacts that the person has called to or received call from.

• **duration_total, duration_mean, duration_median, duration_std**: total, average, median, and standard deviation of duration of all phone calls.

• **duration_total_incoming, duration_mean_incoming, duration_median_incoming, duration_std_incoming**: total, average, median, and standard deviation of duration of all *incoming* phone calls.

• **duration_total_outgoing, duration_mean_outgoing, duration_median_outgoing, duration_std_outgoing**: total, average, median, and standard deviation of duration of all *outgoing* phone calls.

• **entropy**: entropy of total duration of calls with each unique contact. It measures how the person distributes his/her call effort among their contacts.

• **entropy_incoming, entropy_outgoing**: entropy of total duration of incoming and outgoing calls with each unique contact.

### 4.2 Feature Analysis

#### 4.2.1 Welch’s t-test

In order to find variations between people’s behavior in different groups, we aim to examine the significance of difference between features extracted from each group. In cases where there are more than 2 groups, we consider one of the groups as baseline to which other groups are compared.

A statistical test hypothesis is a method to find out whether a particular hypothesis holds in a dataset or not. Here, there are two populations (same features from two groups, for example time spent at home in control group and mdd group) and we want to see if they are significantly different or not. The type of test we use depends on how “difference” is defined. This study utilizes **Welch’s t-test**, which is applied to examine the difference through the lens of sample’s mean. Welch’s t-test, or unequal variances t-test, is used to compare two independent populations with normal distribution and unequal variances. It is assumed that underlying data comes from normal distribution or at least it is not highly skewed. Given two independent set of data, \(X\) and \(Y\) with variances \(\sigma_X^2\) and \(\sigma_Y^2\), a \((1 - \alpha)100\%\) confidence interval for the difference between population means is

\[
\bar{X} - \bar{Y} \pm t_{\alpha/2,r} \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}},
\]
where \( r \) is the degree of freedom, \( n \) and \( m \) are the sample sizes, and \( s_X^2 \) and \( s_Y^2 \) are sample standard deviation. The degree of freedom in \( t \) statistics is calculated as

\[
r = \frac{(s_X^2/n + s_Y^2/m)^2}{(s_X^2/n)^2/n-1 + (s_Y^2/m)^2/m-1}.
\]

The \( r \) and \( t \) are used with the Student’s t-distribution to test the null hypothesis that the two populations have equal means. The distribution of most of the populations we study in this project are similar to normal distribution; so the usage of Welch’s t-test is valid.

### 4.2.2 Comparing groups

There are two different settings in which samples are grouped. In the first setting, **disorder based grouping**, samples are grouped based on their disorder diagnosis. This label is persistent in the course of the study for each individual. For example, if a person is diagnosed to suffer from major depressive disorder, all the samples from that person throughout their participation are labeled as mdd. However, there might be some cases where the person has recovered from their disorder but since such data is not available in our dataset, their disorder group is considered to be persistent.

To take the dynamic nature of people’s mood and state into account, we group them based on their mood score. In the second setting, called **mood based grouping**, samples are grouped based on their PHQ9 score as a proxy for their mood. PHQ9 score is the sum of the answers of each PHQ9 question that participants have responded to. Based on PHQ9 scores, individuals are grouped as follows: PHQ9 between 0-4 are **minimal**, 5-9 are **mild**, 10-14 are **moderate**, 15-19 are **moderately severe**, and 20-27 are **severe**. An advantage of such grouping is that it takes the variations of mood in the long-run into consideration. A person, independent of their disorder group, can have samples belonging to different depression mood groups over time. A boxplot of PHQ9 scores of people based on their disorder groups is illustrated in Figure 6. As expected, people in control group have lower PHQ9 scores compared to others. However, there are some people with a mental disorder who have low PHQ9 scores. That is either because they were temporarily experiencing a better mood or they have recovered.

### 4.3 Correlation Analysis

This section aims to answer this question: “Which behavioral features are strongly correlated with self-reported mood scores?” To answer this question we compute the correlation between the PHQ9 scores and the behavioral features extracted from the sensor data within the 14 days prior to the survey. A correlation coefficient is a number that measures a statistical relationship between two variables. The possible range of values for a correlation is a number between -1 and 1, where +1, -1, and 0 indicate the strongest possible agreement, strongest possible disagreement, and no relationship, respectively. In other words, the correlation -1 between two variables...
means that as one variable gets larger the other one gets smaller, and the correlation 1 between two variables means that as one variable gets larger the other one gets larger too.

Studying the correlation coefficient between PHQ9 score and a particular behavioral feature gives insight into the link between a behavioral trait and depressive state. This enables us to find behavioral patterns which are associated with depressed or healthy moods. However, one must consider not to make conclusions about causal relationship between a behavioral feature and mood scores. In other words, the presence of correlation between behavior X and PHQ9 scores does not justify the claim that the behavior X leads to depression or vice versa.

In this study we employ Pearson’s product-moment correlation coefficient, which measures the linear relationship between two quantitative variables. Since we are analyzing a sample of population, we calculate sample pearson correlation coefficient. Given two pair of paired datasets, \( x \) and \( y \), consisting of \( n \) samples, sample correlation coefficient is

\[
 r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},
\]

where \( \bar{x} \) and \( \bar{y} \) are sample means of each data set.

As it can be seen in Figure 6, the PHQ9 scores of subjects from control group do not vary very much and most of them are 0. Since the variation of PHQ9 and its relationship with behavioral features is our primary interest, the data from control subjects are excluded for the correlation analysis.

After calculating the correlation coefficients, our interest is to find either positive or negative correlations. However, most of the calculated correlations are not very strong which means the correlations, if any, are weak. To solve this issue, very weak correlations should be removed. To do so, we conduct a statistical test with the null hypothesis that the correlation between two data sets is zero. We reject the null hypothesis with \( P < 0.05 \) to end up with significant correlations.

Figure 6: Boxplot of PHQ9 scores in each disorder group. People in control group have lower PHQ9 score.
4.4 Signature Analysis

4.4.1 Social Signature

Social Signature quantifies how people allocate their communication effort among the members (alters) of their social network. Researchers have found that people seem to have a persistent, robust social signature in their phone calls communication [48]. A few close, top-ranked alters receive a large amount of communication. Although this pattern varies between individuals, the distribution of social signature for each individual tends to be persistent over time, despite turnovers in their social network due to changing their environment, i.e., going to university or work. In another study with population of over 500,000, it has been shown that the persistence in social signature holds not only for calls but also for phone messages and a combination of calls and messages [49].

Social Signature of each person, or ego, is defined as the distribution of the interactions with their alters. To construct the social signature of a person, we first calculate the amount of interaction this person has with their alters. This can be the total number of messages they sent or the total duration of their phone calls. Then we rank the alters based on the amount of interaction and normalize them to come up with a probability distribution. Figure 7 shows an example of this procedure.

Our goal is to compare the similarity of two groups using their social signatures. First, we compute the social signature of all samples in each group. The social signatures represent a probability distribution. For instance, \( P_i = (p_i(1), p_i(2), \ldots, p_i(n)) \), is the social signature of the ego \( i \) and \( P_i(k) \) is the fraction of interactions of the ego with the rank \( k \) alter. A possible way for measuring the similarity between two social signatures is methods for comparing the shape of probability distributions. A suitable choice is Jensen Shannon Divergence (JSD), as is done in [48]. The JSD is a
generalized form of the Kullback–Leibler divergence and it is symmetric and always finite. The JSD between two probability distributions, \( P_1 \) and \( P_2 \), is defined as

\[
\text{JSD}(P_1, P_2) = H\left(\frac{P_1 + P_2}{2}\right) - \frac{H(P_1) + H(P_2)}{2}.
\]

\( H(P) \) is Shannon entropy and is defined as \( H(P) = -\sum_{i=1}^{n} p(i) \log p(i) \), where \( n \) is the maximum rank (i.e. number of alters).

We measure Social Signature Distance (SSD) between two groups \( G_1 \) and \( G_2 \) of social signatures as the set of JSD distances between each pair of social signature in \( G_1 \) and \( G_2 \). To be more precise, if \( G_1 = (P_{11}, P_{21}, P_{31}, \ldots, P_{n1}) \) and \( G_2 = (P_{12}, P_{22}, P_{32}, \ldots, P_{m2}) \) are the set of social signatures of groups, \( G_1 \) and \( G_2 \), respectively the SSD between the groups \( G_1 \) and \( G_2 \) is calculated as

\[
\text{SSD}(G_1, G_2) = \{\text{JSD}(P_{i1}, P_{j2}) ; i = 1..n, j = 1..m\}.
\]

When computing the SSD of a group with itself (for example control vs control) we omit calculating the difference between a social signature with itself so as to reduce bias since the JSD distance of a distribution to itself is zero. In other words,

\[
\text{SSD}(G_1, G_1) = \{\text{JSD}(P_{i1}, P_{j1}) ; i = 1..n, j = 1..n, i \neq j\}.
\]

The closer the social signature of the two groups, the lower the distance between those groups will be.

A schematic illustration of this method for comparing groups of distributions is shown in Figure 8.

### 4.4.2 Location Signature

We generalize the idea of social signatures to location trajectory of participants. A person spends time in different venues, for example, at home; workplace; friends’ houses and so on. The fraction of time spent in the 5 top rank places forms a probability distribution for each individual which we call Location Signature. In other words, location signature of participant \( i \) is \( P_i = (p_i(1), p_i(2), \ldots, p_i(5)) \), where \( p_i(k) \) is the fraction of time spent in the top \( k \) place.

Like SSD, we define Location Signature Distance (LSD) between two groups as the JSD distance between each pair of location signatures in the two groups.

### 4.5 Rhythm Analysis

Human activities follow circadian or daily rhythms which can be observed in various psychological, physiological and biochemical aspects [50, 51]. Daily rhythms are generally synchronized with the day-night cycle [52]. However, there are significant variations between daily rhythms of individuals. Particularly, this difference is clear between morning persons (people who tend to wake up early) and evening persons (people who tend to go to bed late at night) [53].
In this study, we investigate the daily and weekly rhythms of outgoing calls. Daily rhythm of a person $i$ is defined as a vector $R_{di} = (r_{d1}, r_{d2}, ..., r_{d24})$, where $r_{dj}$ is the number of calls made in the hours $j-1$ and $j$ divided by total number of calls on any day [54]. For example, $r_{d16}$ is the fraction of calls made during [3PM, 4PM) on any day. The weekly rhythm is defined in a similar fashion, $R_{wi} = (r_{w1}, r_{w2}, ..., r_{w168})$, where $r_{wj}$ is the fraction of calls made on day $\lceil j/24 \rceil$ of the week (week starting on Mondays) and hour $j \% 24$ of the day on any week [55]. For example, $r_{w38}$ is the fraction of calls made on Tuesdays ($\lceil 38/24 \rceil = 2$) during [1PM, 2PM). Since the values in the rhythm vector add up to one, they can be considered as probability distribution and JSD could be used as a similarity measure between rhythms. Both the daily and weekly rhythms are computed over periods of 3 months. The daily call rhythms of 3 subjects from the control group are depicted in Figure 9.

This analysis mostly concerns answering two questions. First, we aim to answer the question “is the daily or weekly rhythm of control people different than each of patient groups?”.

The second question revolves around finding a relationship between changes in PHQ9 and changes in rhythms. This hypothesis is based on the assumption that changes in mood might be reflected in changes in previous habits or behavioral patterns.

To answer the first question, we compute the rhythm distance between each pair of groups and conduct Welch’s t-test to examine if their average is the same or not. This method is the same as the method done for comparing social signature and
location signature of groups discussed in the previous sections.

The second problem requires a different treatment. We compute daily rhythms of calls during two periods prior to answering PHQ9 questions: 2 weeks and 90 days. The rhythm obtained from the latter is a more accurate estimation of the true daily rhythm distribution as more data is available. In both approaches, there is no overlap among the time intervals. After computing the rhythms in this way, we compute the changes in both PHQ9 and daily rhythms between consecutive PHQ9 intervals of a subject. Figure 10 shows the way the pairs of change in rhythms and changes in PHQ9s are computed.

After computing changes, the presence of linear correlation between changes in rhythms and changes in PHQ9s is examined by computing the correlations between them. In this analysis, the data of control subjects are excluded.
5 Results

This chapter presents the results of applying the methods explained in the previous chapter on the MMM dataset. First, in Section 5.1, the behavioral features of people in different disorder groups or mood score categories are studied. Next, in Section 5.2, the correlation of behavioral features with PHQ9 score is analyzed. Section 5.3 investigates the differences in social and location signatures of different disorder groups. Finally, Section 5.4 investigates daily and weekly rhythms of people in control and other patient groups as well as the link between changes in rhythms and changes in mood.

5.1 Feature Analysis

In this section, message, call, and location features of people in different disorder groups are examined. We consider differences with p-value < 0.01 to be statistically significant for this section.

5.1.1 Messages

A first look at the pair plot of message features, illustrated in Figure A1, gives little hint about the difference in features between groups. In order to assess the difference between features of each group, we perform Welch’s t-test on the means of each group compared to the control group.

The number of incoming messages in all patient groups (except for mdd) is significantly higher than the control group (Figure 11a). However, the number of outgoing messages, only in the bd group is significantly higher than the control group (Figure 11b). As it can be seen from the Figures 11a and 11b, the individuals in bd group tend to send and receive more messages compared to other groups. To be more specific, on average, the people in bd group have sent about 2.5 times and received about 2.6 times more messages in periods of two weeks compared to the control group, respectively. In terms of number of unique contacts, people in bpd and mdd groups tend to have fewer unique contacts than the control group. Nonetheless, bd group have contacted a higher number of unique individuals compared to the control group (Figure 11c). The difference in ratio of incoming to outgoing messages between patient and the control groups is only significant in mdd group. This suggests that subjects in the mdd group have higher incoming to outgoing message ratio compared to those in the control group (Figure 11d).

Figure 12 illustrates the CDF of message features, while samples are grouped based on their depression categories. People with mild and severe depression moods tend to receive more messages than people with minimal depression (Figure 12a). There is not much difference in the number of outgoing messages between minimally depressed individuals with other groups (Figure 12b). Finally, people in moderate and moderately severe moods have significantly higher ratio of incoming to outgoing messages compared to minimal depression group (Figure 12d).

Putting all this information together, it is difficult to come up with a clear and consistent pattern between depression score and message features. In other words, we
cannot find a pattern about the level of depression and value of features in message data.

Figure 11: The CDF of different computed features extracted from message data. Groups are disorders. The people in the bd group communicated more using message and they had a higher number of unique contacts compared to other groups.

Figure 12: The CDF of different computed features extracted from message data. Groups are depression categories.
5.1.2 Call

The pair plot of a subset of call features is illustrated in the Figure 13. To examine if the average of features in the control group differs from other groups, we perform Welch’s t-test.

The difference in the average number of incoming calls between the control group and both bd and bpd groups is significant. The control, bd, and bpd group received 22, 34, and 15 calls, on average, over periods of 2 weeks (Figure 13a). The average number of incoming calls for mdd group, however, is close to the control group and is not significantly different. The bd group made more outgoing calls compared to the control group, and mdd and bpd group made fewer calls (Figure 13b). Regarding the number of missed calls, the control group had the fewest missed calls among the groups while bd had the highest number of missed calls (Figure 13c). The average duration of calls (outgoing, incoming, and both combined) for bd group is fewer than the control group and for bpd is higher than the control group (Figures 13j, 13h, 13e). People in bpd and mdd groups made call contacts with fewer unique contacts than control and bd groups (Figure 13m).

Now we focus on the relationship between call behaviour and depressed moods (Figure 14). People with severe and moderately severe depressed moods tend to have fewer incoming, outgoing, and missed calls (Figures 14a, 14b, 14c). Besides, they have a lower entropy of duration of calls (Figure 14l) and fewer unique contacts (Figure 14m) compared to other groups. On the other hand, the mentioned groups have a higher average and standard deviation of duration of calls. From these observations, it might be concluded that strong depressive symptoms correlate with less call contact with other people. However, the average duration of calls for depressed people is higher than people with minimal depression, which suggests that they might have stronger ties with their peers.
Figure 13: The CDF of different computed features extracted from call data. Groups are disorders. People in the bd group had more communications via call (number of incoming/outgoing/missed calls).
Figure 14: The CDF of different computed features extracted from call data. Groups are depression categories. Individuals with more severe levels of depression tend to have fewer number of incoming and outgoing calls as well as fewer unique contacts. However, they seem to have higher average duration of calls.
5.1.3 Location

The Figures 15 and 16 illustrate the CDF of a subset of computed features on location data. There is not statistically significant difference between the total distance traveled in the control group with other groups (Figure 15a). However, the normalized entropy of time spent in SPs in the control group is higher than bd and bpd groups (Figure 15b). People in bd group tend to spend more time in home compared to other groups and there is not much difference between time spent at home among control, bd, and bpd groups (Figure 15c). Persons in the control group have the fewest number of SPs and transition between SPs compared to other groups (Figures 15g and 15h). People in groups bd and bpd have more SPs as well as transitions between them (Figures 15g and 15h). Combining this with the observation that people in bd group spend more time in home suggests that when they go out, they tend to go to more diverse places compared to the control group.

Figure 15: The CDF of different computed features extracted from location data. Groups are disorders. Subjects in the bd group spent more time at home but have more SPs.

Figure 16 shows the difference between location features in different depression groups. People in severe or moderately severe depression have a lower total distance traveled compared to people with minimal or no depression (Figure 16a). This
shows that depressed people tend to move and travel less compared to people in normal mood. Looking at the Figure 16c, we observe that people with severe and moderately severe depression tend to spend more time in home compared to minimally depressed individuals. However, the difference between minimal and severe groups is not significant (p-value = 0.156). This weakens the claim that the more depressed a person, the more time they spend in home. In terms of the number of SPs, there is not much difference between depression groups except for mild depression group which is higher than minimal group (Figure 16g). Figure 15f shows that people with severe, moderately severe, or moderate levels of depressed moods had significantly lower location variance compared to people with no or minimal depression.

To summarize, bd group communicate more (both in call and message communication) compared to other groups and they spend more time in home. Depressed moods are associated with fewer unique call contacts and depressed people have longer calls with their contacts compared to people in normal mood. This shows that depressed people limit their social circle to a set of close friends or family and tend to spend more time with them.

5.2 Correlation Analysis

We investigate the correlation between total PHQ9 score of subjects with the behavioral features they had within the 2 weeks before answering PHQ9 questions. Only correlations with $P < 0.05$ are kept for the rest of this section. The Figure 17 shows the correlations with acceptable significance. It can be observed that total duration of calls; number of both incoming and outgoing calls; and number of unique contacts have negative correlation with PHQ9 score. In other words, the higher the number of calls a person made, the lower their PHQ9 (and hence better mood) they had. On the other hand, average duration of calls has a positive correlation with PHQ9. It can be concluded that individuals who are going through phases of depressed states are likely to have longer calls.

Each of the features represent some part of subjects’ behavioral pattern. The number of unique contacts can represent the size of a person’s social network and average duration of calls can be considered as the strength of bonds with peers. Following this assumption, it can be concluded that depressed people tend to have smaller social networks with more strong connections compared to less depressed people. However, less depressed states are associated with more diverse social networks and higher total communication effort.

The same analysis is done on location data (Figure 18). The fraction of time spent at home is the only feature which is positively correlated with PHQ9, while logarithm of variance, total distance traveled, fraction of moving states, and number of transitions between SPs are negatively correlated with PHQ9. Fraction of time spent at home can represent the level of social isolation and the other mentioned features can describe the level of mobility and physical activities. With this in mind, it can be inferred that depressed states are associated with higher social isolation and more time spent at home. On the other hand, less depressed people tend to move more to different places.
Figure 16: The CDF of different computed features extracted from location data. Groups are depression categories. Individuals with more severe levels of depression tend to have a lower total distance traveled and location variance. On the other hand, they spend more time at home.

5.3 Signature Analysis

The aim of this section is to employ the concept of social signature to compare the behaviour of control samples with patients. In each analysis, we compute the SSD or LSD of group control with groups 1) control itself, 2) major depressive disorder (mdd), 3) border line personality disorder (bpd), 4) bipolar disorder (bd), and 5) mdd, bpd, and bd combined (patient). The next sections discuss how we do so on the MMM dataset.

5.3.1 Social Signature in calls

In call egocentric networks, we consider the type of interaction between ego and alters to be the number of calls between them, like in [48]. Since participants have different amounts of data, to make the comparison more fair, we only look at the call data for 90 days for each individual. In the Figure 19a, the CDF of SSDs is
provided. The Figure 19b depicts the average of SSDs. While in the first glance, the social signature of controls are more similar to each other compared to other groups, this question arises that to what extent this difference is significant. To answer this question, we perform Welch’s t-test for comparing means of the SSDs of “control vs control” and “control vs patients”. All the p-values are smaller than 0.001 which suggests the difference is more likely to be significant rather than due to chance.

When people are in the work environment, several external factors affect their communication behaviour which might not represent their internal mood states. Therefore, we excluded working times (between 8 AM and 6 PM Monday to Friday) from the communication data and studied the resulting features. The aim is to examine whether the difference between healthy control and patients amplifies in this time period or not. We study the call social signature through the lens of duration of calls and number of calls. In both cases, the difference in the social signature of controls and patients remained significant (p-values < 0.001) both in working and non working times. This leads us to the conclusion that working times did not have
a significant effect on the call behavior of individuals.

5.3.2 Social Singnature in messages

We analyze the social signatures of messages in MMM dataset. Here, the egocentric network is created from text messages (SMS) made between ego and alters, and the interaction between ego and alters is defined as the number of messages transferred between them. We perform the same analysis as the previous section. Only the messages in the first 30 days for each individual are considered here. Figures 19c and 19d represent the CDF of SSD between groups and the average of SSD, respectively. Looking at the p-values of Welch’s t-test on the means of the differences suggests that there is no evidence of significant difference between the groups.

5.3.3 Location Signature

In this section, we aim to examine if the distribution of location signature for the control groups is different from other groups or not. First, LSD of group control with itself is considered as the reference LSD and we compare that with LSD of control and other groups. The CDF of LSD of control with other groups is shown in the Figure 19e and their means are visualized in the Figure 19f. Welch’s t-test for comparing the means of LSD distributions shows that average LSD of controls with themselves is lower than that of controls with patients (p-value < 0.01).
Figure 19: The CDF and average of SSD and LSD between groups. SSD is the set of JSD distance between each pair of social signatures in two groups. LSD is defined in the same manner, except that instead of social signature, we use location signature. Welch’s t-test show that average of SSD and LSD between control vs control is significantly higher than control vs patient both in call and location data.
5.4  Rhythm Analysis

As discussed in the section 4.5, there are two questions that we aim to answer. First, we are interested in investigating whether rhythm of the patient groups is different from the control group. Second, the relationship between change in PHQ9 and change in rhythm is studied. The next few subsections present the results of these analysis.

5.4.1  Group comparison

Before computing the group distances, it is worthwhile to visualize the aggregated rhythms of each group. Figures 20 and 21 illustrate the averaged daily and weekly rhythms of participants in each group, respectively.

Looking at the daily rhythms, several points about daily rhythms can be seen. During the night (between midnight and 6 AM) almost no calls have been made. Strangely, there has been a few calls by people in bpd group in the night. At 7 in the morning, which can be considered as the start of the day, fraction of calls for each group increases, with the control group having the highest increase. Most of the calls are happened during afternoon (between noon and 6 PM). After 6 PM the fraction of calls faces a decreasing trend until the end of the day.

The weekly rhythms shown in the Figure 21 represents the daily rhythm of groups which is being repeated (to some degree) each day. Generally, there are fewer calls made during the weekend (Saturday and Sunday) compared to other days.

We conduct the analysis based on the procedure explained in 4.5. The results are shown in the Figure 22. According to the Figures 22b and 22d, the daily and weekly rhythm distances between control vs itself is lower than compared to distance between control and other patient groups. Interestingly, the average JDSs in weekly rhythms have higher values than average JSDs in daily rhythms, the former being between 0.45 and 0.55 and the latter being between 0.25 and 0.3. That is because weekly rhythm vector is longer than daily rhythm vector (168 vs 24) which allows more variation and degree of freedom and hence higher distances. The next step is to conduct the statistical test on the average of distances to test the null hypothesis that if there is no significant difference between averages. The results of Wilch's t-test indicate except for bd group all of the p values are lower than 0.001 which rejects the null hypothesis of no difference in distances. It can be said that call rhythm of subjects in the control group is similar to subjects in bd group but it is different than subjects in other patient groups or all the patient groups combined.

5.4.2  Relationship between change in PHQ9 and change in rhythm

The link between change in mood score and rhythm is studied in short-term and long-term approaches. In the short-term approach, the daily rhythm is obtained from the calls made before answering the PHQ9 up until 2 weeks prior to that. In the long-term approach the length of this interval is 3 months. In the Figures 23 and 24 the scatter plot of absolute value of change in PHQ9 along with the change in rhythm distributions is illustrated. The points are color coded to represent the
Figure 20: Daily rhythm of calls averaged over daily rhythm of all individuals within a group. It seems that people in the control group start their days sooner that other groups and most of their calls are happening in the evening (after 4PM).

Figure 21: Weekly rhythm of calls averaged over weekly rhythm of all individuals within a group. During the weekend (Saturday and Sunday) fewer calls are made compared to the weekdays for all the groups.

The patent group the associated subject belongs to. Interestingly, there has been several cases of increasing PHQ9 by over 10 points in around two weeks, all of which belong to mdd and bpd groups.

It is difficult to find a clear link between change in PHQ9 and change in rhythm from the plots at the first sight. We proceed by computing the correlation coefficient between those two variables. The correlation for the short-term approach is 0.034
Figure 22: The CDF and average of rhythm distances between groups. Rhythm of the control group is more similar to itself than to each of the other patient groups. This holds for both daily and weekly rhythms.

and for the long-term approach is -0.029, both of which give no indication of a linear relationship between change in mood and change in daily rhythm.
Figure 23: Scatter plot of change in PHQ9 score and daily rhythm in different patient groups over 2-week periods. No correlation between change in mood and change in daily rhythm was found ($r = 0.034$).

Figure 24: Scatter plot of change in PHQ9 score and daily rhythm in different patient groups over 90-day periods. No correlation between change in mood and change in daily rhythm was found ($r = -0.029$).
6 Conclusions and discussion

6.1 Feature analysis

In this part, all the behavioral features computed from message, call, and location data were analyzed independently. Based on two grouping methods, disorder-based and mood-based groupings, the difference between features of each group were studied.

The people in the bd group spent more time at home, and communicated more with their phones compared to the other disorder groups. They had statistically significant higher number of incoming, outgoing, and missed calls along with higher incoming and outgoing messages. However, the average duration of calls in bd group is lower than other groups. Assuming the average duration of calls to be the strength of bonds of people with their peers, this might suggest that that people in the bd group have weaker bonds with their social network. Also, people in control group had the fewest number of missed calls, number of SPs, and number of transitions between SPs compared to patient groups.

Looking at the depressed mood based grouping, people with severe and moderately severe depressed moods had fewer call communications and fewer unique call contacts, while having higher average of call duration compared to people with no depression. Besides, people with depression had lower total distance traveled and they tend to stay more at home. This suggests that people with depression have lower physical activity and tend to stay more at home compared to healthy control group.

6.2 Correlation analysis

This part summarizes the findings about the correlation of behavioral patterns with PHQ9 scores. It is shown that the number of calls, total duration of calls, and number of unique call contacts have a negative correlation with PHQ9 scores. On the other hand, average duration of calls positively correlates with PHQ9, showing that depressed individuals tend to have longer calls. These findings suggest that depressed people have a smaller social circle with stronger bonds. However, they make less communication effort in total compared to less depressed people.

The fraction of time spent at home is the only location feature with positive correlation with PHQ9. On the contrary, logarithm of variance, total distance traveled, fraction of moving states, and number of transitions between SPs negatively correlates with PHQ9. This gives indication of lower levels of physical activity in depressed people, which is in line with the findings from previous section.

6.3 Signature analysis

This analysis showed that call SSD and LSD of control group with itself is significantly lower than that of control group with each of the other patient groups. This means that the social signature of calls and location signatures of the control group are more similar to that of itself compared to the other groups. On the other hand, no
difference between message social signatures of the control group with other groups were found. In conclusion, the way the control people distribute their call efforts among their social contacts is different compared to other groups. But why such a difference exists and the mechanisms behind this observations remain issues that call for further research.

6.4 Rhythm analysis
In this analysis, the daily and weekly call rhythm of subjects were studied. The method for comparing the rhythms of control group with other group is similar to the method for comparing social signature and location signature. It was found that both daily and weekly rhythm of control group is more similar to itself than that of patient group. However, the daily and weekly rhythm of control and bd groups looks similar to each other. An initial experiment on all the daily rhythms was done by clustering them using JSD as a distance metric to find the types of daily rhythms (look at the Figure 25). However, like the signature analysis, the cause and implications of the difference in rhythms between control and patient groups need further research.

Another analysis done on rhythms was investigating a link between changes in rhythms and changes in PHQ9 scores. It was shown that there is no linear relationship with the the change in the distribution of call rhythms and magnitude of change in PHQ9 scores of the subjects. This suggests that the variations in rhythms might not be a suitable proxy for tracking change in mood levels.

6.5 Limitations
The first limitation of this work is that there were huge variations in the data availability among participants. Therefore, we had to remove some of the subjects with few data points from our study. It might be worthwhile to find why some subjects had few data points. It could be because there were technical issues with the data collection server or the digital devices, which should be addressed in the future data collection projects.

Second, the features calculated from the smartphones might not fully represent the actual behavior of individuals. That is because 1) they might not be carrying their phones with themselves, for example when they go outside or 2) there are some errors and inaccuracies in the data collected from sensors. While the former problem is more challenging to resolve, the latter is addressed to some extent by the preprocessing step on location data. Not all of the network-originated location data points were highly inaccurate; therefore, a more detailed study on the source of inaccuracy in those data points and finding methods for keeping high-quality network locations is needed.

Finally, the occurrence of SARS-CoV-2 (Covid-19) pandemic in December 2019 [56], which happened during the data collection, might have changed the behaviors of participants. For instance, the widespread social isolation during the pandemic increased the feeling of loneliness and the risk of depression [57], and also increased
Figure 25: Average of rhythms in each rhythm cluster using DBSCAN clustering algorithm. In cluster 2, the highest call activity is done around morning and noon and there are fewer calls after 2PM compared to other clusters. Cluster 3 follows the opposite of this pattern: during the morning (between 7AM and 11AM), fewer calls happened compared to other clusters and most of the call activity is done in the afternoon around 5PM. The calls in cluster 1 are distributed more evenly across the waking time compared to other clusters. Almost half of the rhythms were clustered as outliers, all of which were excluded from this plot.

6.6 Future work

The first possible future work worth exploring is studying the data from sensors other than message, call, and location. The MMM contains data from various sensors, including accelerometer, screen, notification, battery, actigraph, and bed sensor which carry potential insight into the behaviors of people with mental disorders. Each of those sensors captures some behavioral traits of individuals. It might be fruitful to find out which set of sensors better discriminate disorder types or better reflect the mood score. Future data collection projects and research projects could focus more on the resulting sensors to collect more data or extract more representative features from such sensors.

Second, the demographics of the subjects could be taken into account in the
future studies. Although the majority of individuals in MMM dataset were adults, there were imbalance in the gender ratio. Also, living alone; with a partner; or with family, employment status, and being foreigner or native are other factors that could be taken into considerations in further analysis.

Third, the cause of the difference in social signature, location signature, and rhythms of control group with patient group needs to be studied. It would be worthwhile to group individuals based on their signature and rhythms (similar to Figure 25) and to find common characteristics of people in the same group.

All of the analysis in this thesis were done on a group level (either disorder or depression groups) and the variations in people in a particular group were not considered. The difference in personality and life situations of subjects might affect the link between behavior and mood. For example, for a subgroup of subjects, the time spent at home might have a stronger correlation with mood score compared to the rest of the group, which a group-level modeling cannot characterise that.

In all of the analysis in this thesis, each behavioral feature were examined apart from other features, in which the relationship between features are not reflected. The interaction between data from different sensors or behavioral features captured from a specific sensor might carry insightful information. Therefore, another area of future work is to employ more advanced methods for modeling the complex interaction between behavioral features. Machine Learning methods and statistical algorithms such as Linear Mixed Models are capable to do so. However, interpreting the Machine Learning models and coming up with useful explanations for clinical settings might be challenging.

The main focus of this work was to find behavioral markers for detecting mental disorders and mood variations. The final avenue for research direction is to use behavioral data from digital sensors to predict mood score and to classify patients into mental disorder types. Predicting current depression score of subjects will help them and clinicians to keep track of the treatment or progression of depression in a passive manner. The severity of depression can be forecasted for the near future, for instance the next weeks, in order to provide enough time in advance to intervene. The task of prediction and classification can be done using supervised Machine Learning methods.
References


Figure A1: The pair plot of message features.
Figure A2: The pair plot of call features.
Figure A3: CDF of call features before (left column) and after (middle column) covid along with the whole duration of study (right column). The number in front of the groups shows the number of samples. The p value in the title shows the p value of Welch’s t-test on the difference between average of groups.
Figure A3: CDF of call features before (left column) and after (middle column) covid along with the whole duration of study (right column). The number in front of the groups shows the number of samples. The p value in the title shows the p value of Welch’s t test on the difference between average of groups.

Figure A4: CDF of location features before (left column) and after (middle column) covid along with the whole duration of study (right column). The number in front of the groups shows the number of samples. The p value in the title shows the p value of Welch’s t test on the difference between average of groups.
Figure A4: CDF of location features before (left column) and after (middle column) covid along with the whole duration of study (right column). The number in front of the groups shows the number of samples. The p value in the title shows the p value of Welch’s t-test on the difference between average of groups.