Me in the Wild: An Exploratory Study Using Smartphones to Detect the Onset of Depression

Kennedy Opoku Asare[1][0000−0003−1986−1006], Aku Visuri[1][0000−0001−7127−4031], Julio Vega[2][0000−0002−0140−3392], and Denzil Ferreira[1][0000−0002−2195−0449]

[1] Center for Ubiquitous Computing, University of Oulu, Oulu, Finland
{kennedy.opokuasare, aku.visuri, denzil.ferreira}@oulu.fi
[2] Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
vegaju@upmc.edu

Abstract. Research on mobile sensing for mental health monitoring has traditionally explored the correlation between smartphone and wearable data with self-reported mental health symptom severity assessments. The effectiveness of predictive techniques to monitor depression is limited, given the idiosyncratic nature of depression symptoms and the limited availability of objectively labelled depression sensor-driven behaviour. In this paper, we investigate the possibility of using unsupervised anomaly detection methods to monitor the fluctuations of mental health and its severity. Informed by literature, we created a mobile application that collects acknowledged data streams that can be indicative of depression. We recruited 11 participants for a 1-month field study. More specifically, we monitored participants' mobility, overall smartphone interactions, and surrounding ambient noise. The participants provided three self-reports: Big five personality traits, sleep and depression. Our results suggest that digital markers, combined with anomaly detection methods are useful to flag changes in human behaviour over time; thus, enabling mobile just-in-time interventions for in-the-wild assistance.

Keywords: Mobile Sensing · Mental Health · Depression · Anomaly Detection

1 Introduction

Today, depression is one of the most prevalent mental disorders. The World Health Organisation (WHO) reports that depression affects 300 million people globally [81]. Individuals afflicted with depression can experience recurrent episodes of sadness, feelings of worthlessness, suicidal ideation, fatigue, sleep disturbance, loss of appetite, cognitive impairments, and are prone to social and physical isolation [79, 60, 55]. Depression is also known to worsen the outcomes of many medical disorders such as Parkinson's Disease [34], heart failure [51], Alzheimer's Disease and stroke [73]. Depression does not only negatively affect individuals, but to an extent, those around them also. WHO projects that by 2030 [46], depression will be the single most significant contributor to the global

disease burden. In the US alone, the long-term medical care and lost productivity costs add to more than USD 210 billion [31].

Depression is treatable with effective medication and non-pharmaceutical treatments like cognitive behaviour therapy. The challenge, however, is that depression afflicted individuals mostly live unaware or misdiagnosed due to barriers such as social stigma and the scarcity of objective assessment methods. There is the need to extend current clinical diagnosis tools for depression with objective data collected in-the-wild, continuously, and as effortlessly as possible. For the past 30 years, the gold standard for the clinical diagnosis of depression has not changed [24, 79] and is based on subjective (self-reported) assessments such as the Patient Health Questionnaire (PHQ-9) [44], Beck Depression Inventory (BDI) [4], and Hamilton Depression Scale (HAMD) [32]. Thresholds are applied to these instruments' scores to classify the severity of depression for each individual. However, the reliability of current clinical diagnosis methods of depression is debated [24, 45], mainly owing to their subjectivity, and in some cases, because they were derived from clinical consensus with limited empirical evidence [25]. Lastly, these methods are often employed only a couple of times a year in a controlled laboratory or office setting as they require a professionally trained clinician.

Today, smartphone and wearable devices have become part of our everyday lives, and we can better understand people with them [78, 70, 63]. Instrumenting smartphones, wearable devices such as Fitbit, and smartwatches with sensor logging software [21, 45] have made it possible to passively and unobtrusively collect granular, moment by moment and in-situ datasets outside laboratory confinement. In addition, instrumenting smartphones and wearable devices provide an opportunity to actively collect subjective and self-reported data through the Experience Sampling Method - (ESM) [6, 70], instead of paper and pencil diaries or retrospective recollection. Inherent in these time series datasets are human behavioural patterns, i.e., digital biomarkers that are essential in developing interventions for mental health [60, 64, 70, 79].

2 Study Objective

The main objective of this exploratory study is to investigate the feasibility of identifying out of the ordinary human behaviours to enable early detection and monitoring of depression. More specifically, we investigate the feasibility of detecting out of the ordinary behaviours, deterioration of everyday routines, social interactions and mobility, from small datasets of digital biomarkers captured via a smartphone application, using multivariate unsupervised anomaly detection methods. Anomaly detection methods [28] find irregular or nonconforming patterns (outliers) in time series behavioural data, and have been explored in predicting schizophrenia relapses [3, 2], abnormal behaviour of the elderly in Smart Homes [50] and detecting depression in imbalanced datasets [26]. The anomaly detection approach differs from predictive analysis, which requires substantial ground truth data for accurate predictions [69, 62]. We hypothesise that

anomaly detection could be more suited for detecting the early onset and monitoring of depression, given the complex and dynamic nature of depression, the heterogeneity of depression symptoms between individuals [24], and the scarcity of objectively labelled behavioural datasets for depression.

Towards achieving the study objective, we developed an android-based sensing application for smartphones, to passively and unobtrusively collect smartphone sensor data and self-reported surveys. With this application, we collected in-the-wild behavioural data from 11 participants for 4 weeks. We analysed the data to probe deeper into anomalous human behaviours with an ensemble of multivariate anomaly detection algorithms and report our insights into the relationship between the observed anomalous behaviour and depression symptoms.

3 Related Work

Smartphones and wearables (e.g. smartwatches, rings, bracelets) are increasingly accessible to the wider population. These devices are bundled with several sensors to collect and monitor different human activities and their related physiological signals, such as heart rate, sleep quality, body temperature, among others [63]. Lastly and more importantly, most of such devices are, directly or indirectly, connected online [56]. This combination of conditions offers an unprecedented opportunity for a real-time, in-situ understanding of the users' context [22]. The highly personal nature of these devices has driven researchers' interest in investigating the role of Digital Phenotypes/Biomarkers (DPB) in monitoring human behaviour and health conditions [16]. In medicine, phenotypes/biomarkers are physiological, pathological, or anatomical characteristics that are objectively measured and evaluated as an indicator of normal biological, pathological processes or biological responses to therapeutic interventions [12]. Here, we consider a DPB as a multimodal sensory metric, i.e., the outcome of a data analysis that can be compared and measured across individuals using the same combination of data sources.

There are two primary approaches to develop DPBs, e.g., [39]: *active data collection* if a participant is prompted to perform a measurement or provide input (e.g., diaries, self-reports, Experience Sampling Method [6, 70]); and *passive data collection* if measurements occur without users' intervention or input (e.g., wrist-worn devices provide estimates of daily steps and calories autonomously, smartphones' sensor data). Active and passive approaches are collected in tandem to correlate the data points, where the source of active data collection is regarded as the ground truth for psychological and subjective measures [56].

For example, in Alzheimer's disease, there is evidence that cognitive, sensory and motor changes occur 10-15 years before their effective diagnosis by a professional with traditional neuropsychological tests [43]. Dexterity and cognitive tasks' performance have been successfully used to monitor cognitive function decline (e.g., working memory, memory, executive function, language) [14]. Parkinson's clinical scales and self-reporting of symptoms also correlated well with smartphone-based sensing: device interactions, motor activities such as going

up or downstairs, gait (using smartphone's accelerometer, barometer) and social interaction (e.g., amount of texts and calls) [77]. Actigraphy (i.e., monitoring time sequences of activity vs rest) is useful to predict symptom changes in mood disorders such as bipolar and major depressive behaviour [40]. Geolocation-based digital biomarkers such as distance travelled, the number of locations visited, time spent on location was strongly associated with bipolar disorder and schizophrenia [23, 57]. Game-based digital biomarkers such as performance data reflecting cognitive and motor processing and social context such as where, when and with whom a game takes place could predict mental health [54]. More related to our work, in depression symptoms, mobility features such as location variance correlate with depression symptom severity determined by questionnaires such as the PHQ-9 [65].

4 Experiment

4.1 Data Logging Software

We developed an Android-based smartphone sensing application, Me. As a feasibility study and not an intervention study, we aimed at a simple interface and architecture to collect relevant data to explore the use of anomaly detection-based analysis methods to identify potential digital biomarkers for depression.

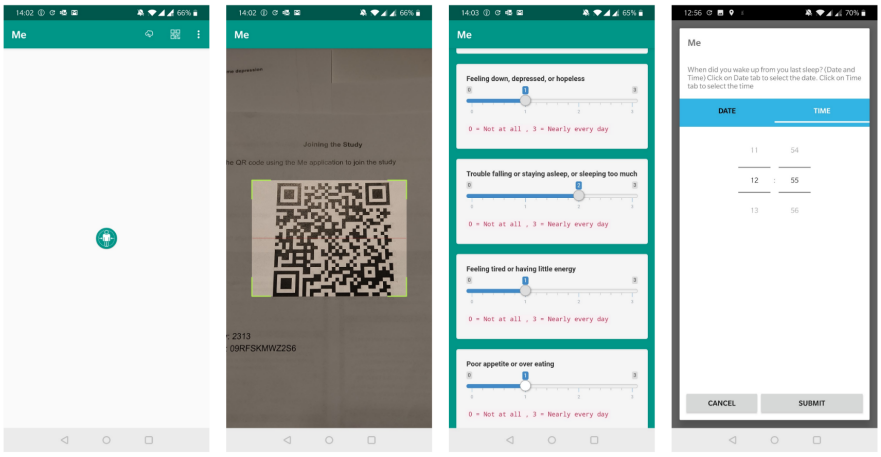

Fig. 1. Four screenshots (A, B, C, D) of the Me application. **A.** is the home screen, B. is the QRCode scanner, **C.** and **D.** are screens of self-reported questionnaires.

Smartphone Sensing Application and Online Data Storage: Our application, Me, is compatible with Android 7.0 and above. The application was

created using the AWARE Framework [21] Android library. Figure 1 shows different views of Me. Figure 1 A is the initial screen with two top-right menu icons: *on-demand data sync* and a QR-Code scanner. By scanning a study-specific QR Code, as seen in Figure 1 B, Me connects to the *Online Data Storage*, i.e., our study database, which is a MySQL server instance running on an Amazon EC2 with encryption enabled. All data collected with Me is primarily stored locally on the device. A background data sync service sends the data over a secure encrypted connection to the study's *Online Data Storage* at 30-minute intervals. In other words, the data transmission to the *Online Data Storage* is encrypted with HTTPS. Alternatively, the data sync can be initiated on-demand by tapping on the *on-demand data sync* menu icon. Me also prompts participants to complete scheduled self-reported surveys at specific times using a notification. Participants can click on this notification to open the survey, as shown in Figure 1 (C and D).

Study Management and Data Analysis: The Study Management dashboard was developed using R Shiny and is used for compliance and data quality monitoring of our participants. The dashboard automatically updates every five minutes with visualisations, including the last time each sensor or survey data was received from each participant. We also developed the *Data Analysis Pipeline* using the R programming language for data pre-processing, visualisation and analysis pipeline for mobile sensing and behavioural analysis.

Data Description and Data Privacy: Me passively collect smartphone sensor data, in addition to self-reported surveys. Based on existing literature [65, 45, 79, 64], several potential digital biomarkers were implemented to detect and monitor mental health. Table 1 details the sensor data that Me collects. In addition to the passively sensed data, Me has built-in self-reported questionnaires for sleep duration, personality traits using the 50-item Big Five personality trait questionnaire [27], and depression assessment using the PHQ-9 [44].

Me inherits AWARE frameworks' privacy-aware features [21]. The study application does not collect any personally identifiable information or sensitive data such as the content of texts, phone calls, visited websites and notifications. We only log metadata such as the time a text was received, the state of the phone screen, i.e., locked or unlocked at a given time. For calls and texts, we anonymise the identity, i.e., *phone number and name* of the other party, into a single alphanumeric trace value, using a one-way SHA-1 hash, on the participant's device. With the one-way SHA-1 hashing method, we retain the same trace value for the same contact and prevent anyone from re-identifying the communicating party. Likewise, Me does not store the actual characters typed using the phone's keyboard, but rather package name, timestamp, and the masked text before and after each keystroke. Text masking is done by replacing all uppercase characters with the letter 'A', lower case characters with the letter 'a', and digits (0,1,...,9) with the digit '1'. Finally, the data collected is not tagged with any personally identifiable labels such as name or email. For each new installation of Me on a smartphone, a Universally Unique Identifier (UUID) is randomly generated on the participant's device and used as the sole identifier of the data

Table 1. A summary of the description, frequency of the data collected with the Me. Each data point is timestamped with the time of the sampling or event.

Sensor Data	**Frequency**	**Sensor Data Description**
GPS Location	5 minutes	Latitude, Longitude coordinates
Physical Activity		Walking, Running, Biking, and In-Vehicle
Light		Intensity of ambient light
Noise	3-second sample every 5 minutes	Intensity of ambient noise
Screen	event-based	Screen locks, unlocks
Touch		Touch interaction type(tap, long tap, scroll)
Battery		Battery level changes, battery charges and discharges
Application		Name of applications that are launched
Notifications		Name of the application that triggered the notification
Calls		Call type (incoming, outgoing, missed), trace of caller
Messages		Message type (received, sent), trace of sender
Keyboard		Masked text before and masked text after a keystroke
Timezone		Device's timezone
Self Reports	**Frequency**	**Description**
BIG-5	1 per participant at the beginning of the study	50 item personality trait questionnaire
Sleep	1 per day in the morning	Start and end date and time of sleep session
PHQ-9	Beginning of the study and 1 per week	9 item depression test questionnaire

from a specific smartphone; thus, all sensor data entries are tagged with this UUID. This UUID is reset if the participant removes and re-installs the app to avoid cross-study matching.

4.2 Recruitment

We conducted a call for participation using a campuswide mailing list, in addition to posting advertisement posters on various faculty and student notice boards. The advertisement had URLs to the study information website that contained a detailed explanation of the purpose of the study, duration, the data collected, participant requirements to be eligible to join the study, and the reward participants would receive after completing the study - a 50 Euro Amazon voucher. A total of 11 participants joined the study, six female and five male, with ages ranging from 19 to 37 years (*Mean: 26.55, Median: 25, SD: 6.02*). Six were undergraduates, three graduate students, and two vocational students. Six out of eleven participants reported a history of clinical diagnosis with depression, anxiety, and other related mental disorders. All participants used a smartphone with Android 8.0 or higher as their primary device. We followed all ethical procedures required by our institution. According to the local ethical board guidelines [75] in the conduct of research, our study is compliant: 1) the study does not deviate from the informed consent; 2) the research does not intervene in the physical integrity of the participants; 3) all our participants are above 15 years old; 4) our study does not expose participants to strong stimuli; 5) there is no intervention, nor there is a foreseeable potential for mental harm

to the participants that exceed the limits of participants' normal daily life or those around them.

4.3 Study Protocol

We conducted a 4 weeks study with three stages: *Onboarding*, *Data Collection and Monitoring*, and *Post Study Debriefing and Exit*.

Onboarding: To avoid cross-participant interaction and bias, we invited each participant to a private meeting. At the meeting, we explained the study's purpose, data collection schedule, their right to quit the study at any time and get their data deleted, and the reward for participation. Afterwards, we asked participants to sign a Consent Form and provide their basic demographic information. We then installed the Me application on their Android-based smartphone, and with a short tutorial, we instructed participants how to fill in the study's surveys visible within the app. We showed the participants a snapshot of the actual data that is collected with the application. We configured the Me to bypass the battery optimisation (Doze) feature of the Android OS. This configuration is to allow the Me to run continuously in the background without interference. Next, using Me, the participant joined the study by scanning the enrolment QR Code. Lastly, participants were then asked to fill the baseline questionnaires, that is, the BIG Five personality traits [27] and PHQ-9 [44], also collected in-app. The onboarding meeting took approximately 30 minutes.

Data Collection and Monitoring: Me (see Figure 1 A) does not present any visualisations, i.e. feedback to the participants during the data collection period, as we do not want to intervene or influence participant behaviour at this stage. We designed Me to prompt participants with notifications when self-reported surveys are due. However, answering ESMs in longitudinal studies poses a considerable burden for participants [6] mainly because the received EMS prompts may be triggered at inopportune moments, or the prompts may go unnoticed, especially when the device contains numerous notifications pending from other applications. To mitigate these challenges and ensure compliance, we monitor the study using the *Study Management* dashboard. When we observed gaps in the data collected, for instance, because the participant phone is not syncing data to the online study database, the ESM surveys were not answered, or the GPS sensor was turned off, we proactively contacted participants with recommended actions over the phone or by email.

Post Study Debriefing and Exit: During this last stage, we again invited each participant individually for a debriefing meeting. The goal of the debriefing was not to present to the participant an opinion on whether they are depressed or otherwise. This goal was made clear to the participants, and we strictly avoided discussing the participant's depressive state. The goal of the debriefing was to assess with participants whether our algorithms could detect out of the ordinary events that are familiar to participants. At this meeting, we presented participants with statistics and visualisations of out of the ordinary events flagged by our algorithm. Figure 2 shows an example of such visualisation. With a semi-structured interview, we discussed the statistics and visualisations and collected

participant's reflections and feedback on the Me application deployment during the study. Finally, we performed a last manual data sync to complete the dataset and uninstalled Me.

5 Analysis Protocol

5.1 Behavioural Analysis

Smartphones are a powerful behavioural observation tool in psychological science [33]. With a focus on finding behaviour indicative of depression, we explore smartphones' sensor data that capture the following behaviours: mobility patterns, daily activities, social interactions [57, 59]. Concretely, we investigate how much a participant transit during the day. This analysis provides an overview of mobility, outdoor activities (location accuracy is significantly better for outdoor locations) and the likelihood of social interactions without necessarily exposing visited locations. We also investigate the most used applications (top-10), allowing us to understand whether a participant is socially interacting non-physically and whether sports, music, browsing the internet play a role in someone's daily activities. Lastly, we look into the screen usage length to understand how engaged the participant was with the applications.

5.2 Feature Extraction

We converted the timestamps of each sensor data into a human-readable date and time format using the timezone data of each participant. From the hour of the day, we determined the day segment as follows; *morning* - 06-11 hours, *afternoon* - 12-17 hours, *evening* - 18-23 hours, and *night* - 00-5 hours. Except for the Location features, which were computed at a daily level only, we aggregated all other features at the daily and day segment level. The aggregation on the day and day segment level allows for the emphasis of behavioural patterns during specific segments of the day, for example, typing speed and typing error rate at night, and the number of unique applications during the morning.

In addition to the minimum, maximum, mean, median, sum, and standard deviation (SD) aggregation of the computed features, we also captured the degrees of complexity and irregularity of features with Shannon Entropy estimation [71, 67] and Normalised Shanon Entropy [65], using the 'entropy' R package [72]. Furthermore, we computed other estimators that are robust to outliers in the computed features, including robust estimator for mean (Huber's M) [38] , and variance (VarQn) [13] using the *robustbase* R package [53]. We summarise the features extracted from the sensor data in Table 2. We explain the feature extraction process in more detail next.

Keyboard: We defined a typing session as all keystrokes while the user is using an application. For every two successive keystrokes per typing session, we determine the keystroke transition as *Character-Character* - from a character to a character, *Character-Backspace* - from character to backspace,

Table 2. Summary of extracted features from the study dataset. Feature_name* denotes that multiple estimations; sum, minimum, maximum, median, mean, standard deviation, entropy, normalised entropy, robust mean and variance (Huber's M and VarQn) were computed for that feature.

Feature Group	**Feature Metrics**
Keyboard	interkey_interval*****, count_session, count_keystrokes, speed, pauses_ratio error_ratio
Location	distance*****, speed*****, nclusters, location_variance
Call	call_count, distinct_contact, duration*****
SMS	sms_count, distinct_contact
Phone usage	unlock_duration, unlock_time_interval*****, count_tap, count_long_tap touch_time_interval*****, usage_sec*****, usage_count, unique_apps
Ambient Noise	episode_sec *****, silent_episodes, loud_episodes

Character-Number - from character to number, *Character-Punctuation* - from character to punctuation and *Other*, and also compute *interkey_interval*, i.e. the time difference in seconds between two successive keystrokes. A transition is a *pause* if the *interkey_interval* exceeds the 95th quantile of all *interkey_interval* per the typing session. The quantile method we used does not assume normality for the distribution of the *interkey_interval*. Also, we aggregate all typing sessions as; (1)*count_session*, i.e. count of typing session, (2) count_keystrokes, i.e. count of keystrokes, (3) speed i.e, count_keystrokes divided by sum of the *interkey_interval*, (4) *pauses_ratio*, i.e. count of all pauses divided by count_keystrokes, (5) *error_ratio*, i.e. count of *Character-Backspace* divided by count_keystrokes, and compute additional estimation as shown in Table 2.

Location: We computed the location features [79, 65] at the day level only. First, we computed the Haversine distance and speed between two successive GPS coordinates. Next, with stationary GPS coordinates [65], we apply DBSCAN [19] clustering to identify the *nclusters*, i.e., the number of significant places participants dwell per day. Furthermore, with the stationary GPS coordinates, we also compute the variability [65] in the GPS coordinates as *location_variance*. Aggregating at the day level, we compute additional estimates for distance and speed, as shown in Table 2.

Call and SMS. For each call type; *missed*, *incoming* and *outgoing*, we aggregated the count, i.e.*call_count*, the number of distinct contacts, i.e. *distinct_contact*, and additional estimates of the call duration as shown in Table 2. Likewise, for SMS, for each SMS type; *sent* and *received*, we aggregated the count and number of distinct contact/trace.

Phone Usage. The phone usage features comprise features extracted from the screen interaction, touch interaction and foreground applications. With screen interaction, we defined a screen episode as the time between two screen states (lock, unlock) changes. We used only the screen unlock episodes to compute screen episode *duration* and additional estimates, as shown in Table 2, of the time between two successive unlocks, i.e., *unlock_time_interval*. Likewise, for touch interactions, aggregated count of tap, scroll, long tap interactions, and additional estimation, in Table 2, of the time between two successive touch interactions, i.e., *touch-_time_interval*. For features from foreground applications, we added

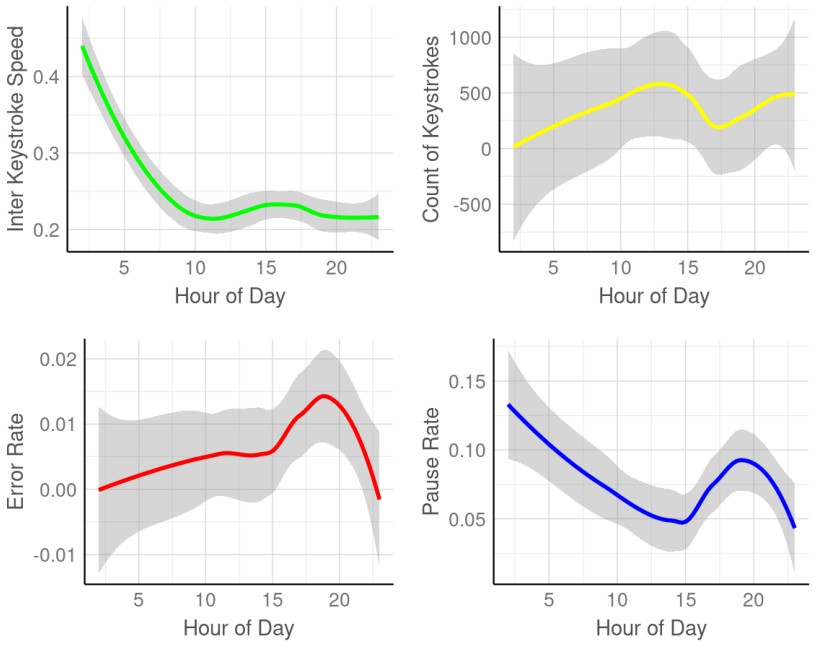

Fig. 2. A visualisation of the typing features of a participant on an anomalous day. The typing speed (top left) gradually declines during morning and evening hours, with a slight increase during afternoon hours. The error rate (bottom left) and pause rate (bottom right) sharply rises and declines between the 15th and 20th hour of the day. The participant confirmed during a debriefing session and explained that this typing feature pattern was due to an emotional mental state and fighting period with a friend.

to each app their corresponding category (e.g., social network) based on an external dataset [71] and the Google Play Store [29]. We defined application use episodes as the time during which a particular application is in the foreground. For each application use episode, we computed *usage_sec* as the usage duration in seconds. We then aggregated the *usage_sec* , *usage_count* - count of all application launches, and *unique_apps*- the count of unique application launches at the daily and day segment level for all application categories, and *messaging*, *calling*, *tvvideoapps*, *musicaudioradio*, *email*, *socialnetworks*, *eating*, *healthselfmonitoring*, *datingmating*, and top 1, 2 and 3 used application categories. Furthermore, for *usage_sec*, we computed additional estimates listed in Table 2.

Ambient Noise. We used 50 decibels as a threshold [82, 1] to determine the noise state; thus, whether the ambient noise was *silent*- less than or equal to 50 decibels or *loud* otherwise. We determined noise episodes as all intermittent samples until the noise state changes. For each noise episode, we computed the *episode_sec* - total duration of the episodes in seconds after discounting sampling time intervals and maximum, minimum and median decibels. Finally, we aggre-

gate the noise episode features into *silent_episodes* - count of silent episodes, the sum of silent episode_sec, *loud_episodes* - count of loud episodes, sum of loud episode_sec, minimum of all minimum decibels of noise episodes, maximum of maximum decibels of noise episodes, mean of mean decibels of noise episodes, median of median decibels of noise episodes. We also compute additional estimations for *episode_sec*, as shown in Table 2.

5.3 Anomaly Detection

Anomaly detection is a process of finding nonconforming, unexpected or irregular patterns or behaviours in a time series dataset, taking into account only the intrinsic properties of the dataset [28]. Depressive behaviour in an individual may manifest as anomalous - an out of the ordinary behaviour in the context of the individual's routine behaviour [47]. For instance, the depressive behaviours may manifest in unusual sleep changes - inferred from changes in sleep duration, wake up time, screen interactions, touch interactions, and typing during the night or physical and social isolation - inferred from changes in calling, texting, use of social media applications, and reduced physical activity [79, 55, 3, 2]

While different anomaly detection methods exist [28, 47], the methods applied in this paper are unsupervised multivariate anomaly detection methods. In contrast to supervised classification methods, unsupervised anomaly detection methods do not need to be trained with labelled datasets of depressive behaviour of the individual, which in practice, may not be available or are scarce at the onset of depression. The applicability of anomaly detection in detecting and monitoring depression is that the growth of anomalous behaviours often translates into critical, significant and actionable information prompting just-in-time interventions. For example, in [3], 71 % higher anomalous behaviour rates were found in the two weeks before relapse of schizophrenia than other times. In healthcare, unsupervised anomaly detection algorithms have been useful to predict health information about smart home residents [48].

We implemented four anomaly detection algorithms; K-Nearest Neighbours (KNN) [28, 7], Isolation Forest (ISOFOR) [49, 18], Local Outlier Factor (LOF) [28, 10, 37], and Connectivity-Based Outlier Factor (COF) [28, 74, 52] to detect anomalies separately in each participant's features (see Table 2), quantified from their 4 weeks dataset. The participant's features computed on the day level and the day segment level were concatenated, into one feature matrix, with each roll representing a particular day. We applied z-score normalisation to the feature matrix before applying the anomaly detection algorithms.

Unlike classification methods that predict a specific class or label, unsupervised anomaly detection algorithms output continuous anomaly scores. Generally for LOF, COF, ISOFOR and KNN, higher anomaly scores indicate a higher likelihood of an anomalous data point [7, 18, 37, 52]. We used the 95th quantile of the anomaly score, without assuming a normal distribution, as a threshold for determining whether the feature matrix point is an anomaly ($>$ the threshold) or non-anomaly otherwise. For each detected participant's anomalous day, we

then compute a *weight*; thus, a simple count of the anomalies detected for the day. Figure 3 shows a high-level overview of the anomaly detection process.

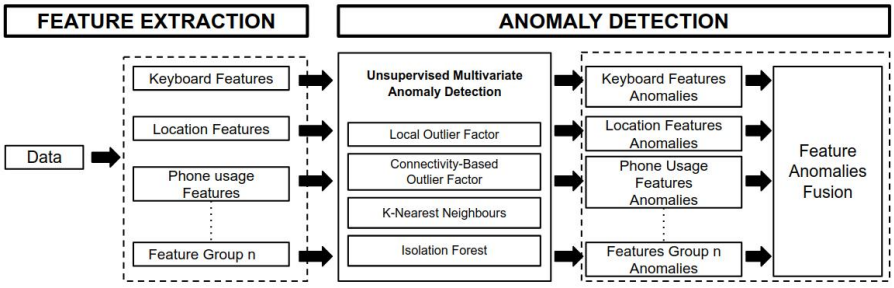

Fig. 3. An overview of the unsupervised multivariate anomaly detection process.

6 Results

6.1 Quantitative Results

We summarise the behavioural data, anomaly detection and self-reported questionnaires collected from 11 participants during our 4 weeks deployment of [OUR-_APP] in Table 1. By applying a threshold [44] on the baseline PHQ-9 responses collected during the *Study Onboarding*, we grouped the participants into *depressed* (PHQ-9 score ≥ 10) and *non-depressed* (PHQ-9 score < 10). Out of the 11 participants, 3 participants (**P1**, **P2**, and **P10**) were *depressed* with a mean PHQ-9 score of 16 (*sd* 7.0) at *Study Onboarding*, and 9 participants were *non-depressed* with a mean PHQ-9 at of score 6 (*sd* 2.27) at *Study Onboarding*. Six participants (**P1**, **P4**, **P6**, **P8**, **P9**, **P11**) mentioned they have at some point in their life been clinically diagnosed of depression, anxiety, and other related mental disorders, and we grouped them as *with history*.

Behavioural Analysis: We did not find a statistical difference between our groups. Our participants' sample is modest (N=11), the groups are not balanced in sufficient numbers. It was incredibly challenging to recruit participants for this pilot. Hence we report our findings using descriptive statistics. As our goal was to pilot the methods and software to assess the usefulness of such in monitoring depression remotely, we conducted this feasibility study with the recruited individuals.

We took the top-10 most used applications and investigated their daily usage patterns (Figure 4 as an example). This figure shows which app, time of day, and how frequently is the app used at a given time of the day.

Across all the participants, the average application usage time was approximately 8 minutes. Comparing across groups, the *depressed* and *with history*

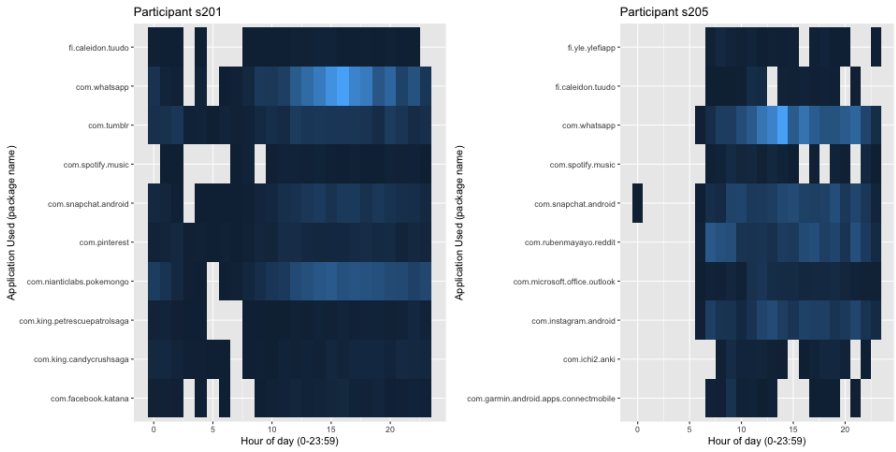

Fig. 4. Example of the most used applications (top-10) for two participants P1 and P5 (depressed and non-depressed, respectively).

group had an average application usage time of approximately 9 minutes. The *non-depressed* group average application usage time was approximately 5 minutes. In the depressed group, we find application usage reveals a pattern in social application usage (e.g., WhatsApp, Facebook, Snapchat, Tumblr) throughout the day and especially late in the evening (23h00 onward) until the early hours of the day (see Figure 4, left). The participants in the *depressed* group show more usage of social apps than within the other groups, for longer periods of time – a normalised daily median frequency and app session length: 231 daily sessions, median session length: 12m31s vs 34, 57 daily launches, 5m26s, 7m32s session length in *non-depressed* and *with history*, respectively. We do not find this pattern in the non-depressed group, with a clear break in application usage during the early hours (e.g., between 0h and 6h00, Figure 4, right). We also find the depressed pattern in P11 (*with history*), but not others in the same group.

Next, we investigated engagement with the smartphone using the screen status (being on or off). This allows us to see when the engagement occurs in the day and for how long (in Figure 5, the longer the engagement the brighter it is).

Across all the participants, the average engagement time was approximately 8 minutes. Comparing across groups, the *depressed* and *with history* group had an average engagement time of approximately 9 minutes. The non-depressed group average engagement time was approximately 5 minutes (see Figure 5). Following the application usage pattern, the engagement of the depressed group is highlighted in the early hours (see Figure 5, left). The non-depressed group shows a diluted engagement pattern throughout the day (i.e., more frequent, more brief).

Lastly, we probed the users' daily mobility patterns by the median distance travelled in a given hour (Figure 6). Across all participants, the median daily total mobility was 7km. Comparing across groups, the *depressed* and *with history*

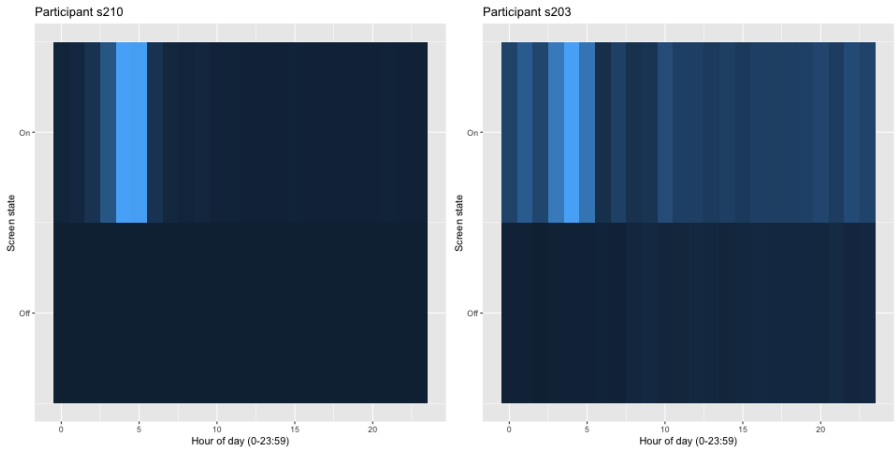

Fig. 5. Example smartphone engagement for two participants P10 and P3 (depressed and non-depressed, respectively).

groups had a median daily total mobility of 7,4km and 3.6km, respectively. The *non-depressed* group median daily total mobility was 12km. Participants in the *depressed* group show a pattern of more mobility in the afternoon hours (11h00 until 18h00, see Figure 6 left). For the *with history* and *non-depressed* group, the mobility starts earlier in the day, often at 8h00. We also find peaks of mobility for the non-depress-ed group around lunch time (11h00-12h00) (see Figure 6, right).

Anomaly Detection: We applied four unsupervised multivariate anomaly detection algorithms separately on each participant's data. We computed at least 74 features per participant per day, from the captured behavioural dataset. We did not include features from calls and SMS since the data contained only a few records from 2 participants. Using Spearman's correlation coefficient, we analysed the relationship between the weekly PHQ-9 responses and anomalies detected within two weeks leading to the PHQ-9 survey date.

We found no statistically significant correlation between PHQ-9 scores and anomalies detected. We found pairwise correlations between the individual PHQ-9 question ratings and anomalies detected. However, the correlations were not statistically significant when their *P-values* were adjusted for multiple testing using the Holm-Bonferroni method [35]. We highlight the pairwise correlations where *P-values* < 0.05 for all participants, depressed and non-depressed group.

For all participants, we found a negative correlation (r= -0.351, p=0.009) between typing or keyboard feature anomalies and PHQ-9 question seven [*Trouble concentrating on things, such as reading the newspaper or watching television*], a negative correlation (r= -0.273, p=0.044) between typing anomalies and PHQ-9 question eight [*Moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual*]. In addition, we found a negative correlation

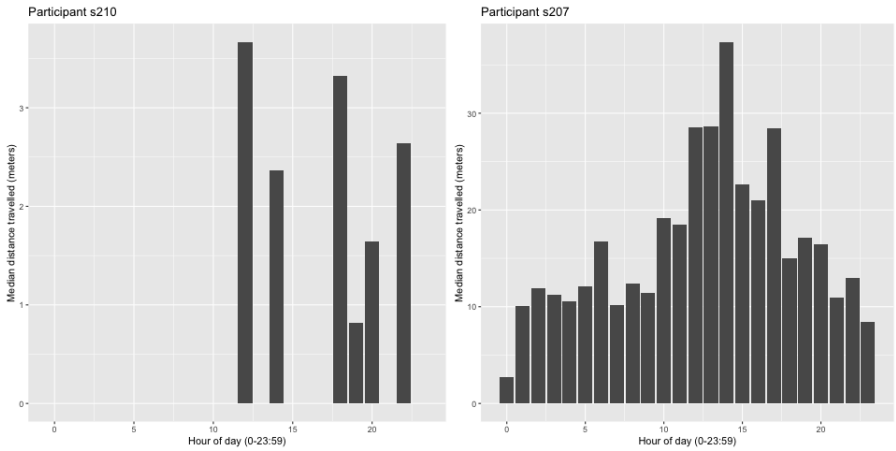

Fig. 6. Example mobility for two participants P10 and P7 (depressed and non-depressed, respectively)

(r= -0.283, p=0.037) between PHQ-9 question eight, and *total anomalies* (the sum of anomalies in keyboard, location, phone usage, ambient noise features, see Table 2), *total anomalies* and PHQ-9 question seven (r= -0.369, p=0.006). These results suggest that participants with difficulties in concentrating on doing routine activities are more likely to exhibit out of the ordinary and anomalous typing behaviours. Additionally, the results also suggest that participants with restless or fidgety mobility patterns are more likely to exhibit anomalous typing behaviours.

For the *depressed* group, we found a negative correlation (r= -0.532, p=0.041) between phone usage feature anomalies and PHQ-9 question three [*Trouble falling or staying asleep, or sleeping too much*], and a negative correlation (r=-0.545, p=0.036) between *total anomalies* and PHQ-9 question seven. This suggests that depressed participants experiencing insomnia or hypersomnia at night, are more likely to use their phones (screen interaction, touch interaction and application launches), in an unusual manner compared to other phone usage patterns.

Similarly, for the *non-depressed* group, we found a positive correlation (r= 0.394, p=0.012) between phone usage feature anomalies and PHQ-9 question four [*Feeling tired or having little energy*], a positive correlation (r= 0.356, p=0.024) between phone usage feature anomalies and PHQ-9 question five [*Poor appetite or overeating*] and a positive correlation (r=0.334, p=0.0035) between ambient noise features anomalies and PHQ-9 question four. This suggests that participants who feel tiredness, stress, poor appetite, and have low energy, are more likely to launch phone applications and interact with their phone through unlocks and scroll in an unusual manner as compared to other phone usage patterns, and might prefer to spend time in low noise environments.

6.2 Qualitative Results

We used the Grounded Theory [11] open coding approach to analyse the data collected at the *Post Study Debriefing*. We read all the responses and reflections from participants and coded them based on similar concepts. We then systematically connected the similar concepts into themes, as follows:

Challenges in the wild: Generally, the Me application functioned as expected, however some participants (N=3) noticed a different behaviour of [OUR-_APP] from what was explained to them during the Study Onboarding. *"I think that PHQ-9 questionnaire didn't come through every Thursday. So I had to remind myself to do that. But that was just a minor thing"* (P4) *"I have maybe two or three times where the app was probably not running properly. So I didn't get the notification in the morning to put the sleeping hours"* (P6). In addition, some participants (N=4) also reported that running Me impacted their phone's storage and battery life. *"I had problem with the flash memory in the device and I'm not sure if it was related to the app or not. Because I was travelling at the time. I took a lot of pictures"* (P6). *"I think the more relevant things I've noticed was the battery because it was going down always."* (P10) *"yeah, the battery drain a bit faster, especially after I got an Oura ring. After that, so, after I stopped using the app, I quickly noticed how my battery lasted longer than usual."* (P4)

Early Feedback: While our research design did not show participants any feedback during the study period, some participants (N=9) were of the view that early feedback on Me would have been more useful, rather than after 4 weeks. Some participants could not recall what happened, or sounded surprised, when they were asked to reflect on statistics and visualisations of their own data. *"I didn't expect it, for example, like a personal feedback. So that was nice. And also to see, for example, that I make more errors, for example, when I am tired or when I am emotionally in trouble"* (P10). *"7th of November? no I don't remember. I can check, I have my calendar, everything is in there. No I don't have anything here"* (P3) *"Well, it's quite interesting. Maybe it's a bit more apps than I thought it would be. But regarding the number of uses, I'm really wondering what happened on that day"* (P4). *"Well, I think it would be interesting to just look at my own history. What have I done on these days? Because I don't remember right now, what happened during that time? "* (P4).

Plausibility of detected anomalies: Whereas some participants could not recall or explain what happened when asked to reflect on visualisations on anomalous behaviours, most participants (N=6) could quickly explain, relate to and confirm the out of the ordinary behaviours detected by the anomaly detection algorithms. *"I had a party on the 26th and then I remember I really used my phone on the 27th, I was emotional, I've been in a bad mood and yeah, fighting over text"* (P10). *"I don't remember if that was the same day I was travelling. Because if it was, I was in the car as a passenger"*(P4). *"I usually sleep quite well. But some nights I haven't slept so good. And during this period of using the app, I had to change my medication and increase the dosage and that could be the cause"* (P1). *"I've been realising I use my phone the most between 11 and 12 at night, 10.30 or 11 I'm in my bed, I take the phone, I have the*

phone for an hour so and then I go to sleep" (P3). *"Maybe some looking out at the game results or something like that. Because I follow Brazilian soccer and then probably looking at the results of the Sunday games. On Mondays because the games are late on Sunday in Brazil, and it's the night here. So on Monday morning, I wake up and check the game scores. And then there are lots of news about it. probably spend some time on that. Yeah."* (P6).

7 Discussion

Our results give us a bulk of smartphone-based user behaviours that may be relevant in monitoring depression unobtrusively. Participants in the *depressed* group went to bed late (i.e., after midnight) - using screen status data - which also provided evidence that it makes it challenging for them to wake up early (first app usage, first time the screen turned on). Moreover, routinely application usage is until late (Figure 4) and more engaged in the evening/night (Figure 5), and daily mobility is usually after 10h00 (Figure 6). The participants in the *depressed* group show more usage of social apps than within the other groups for more extended periods. Participants in the *non-depressed* group followed a device off time night (23h00 onward), picking up the phone at around 8 am. Surprisingly, P11 is a *with history* participant who follows the depressed pattern, but not the others in the same group (they exhibit the non-depressed group pattern). Such information could be beneficial for a mental health professional to investigate further. Participants in the *non-depressed* group are more active (higher daily median mobility) throughout the day, with a peak around lunchtime (11h00-12h00), while others are less mobile.

As depression symptoms are, human behaviour is heterogeneous and varies between individuals. Previous research has established that smartphone usage behaviours and depression varies among different demographics [8, 71, 61, 9]. For example, certain personality traits are associated with compulsive use of YouTube [42], specific application categories can predict personality traits [61], and some applications are more likely to be used at specific hours of the day [8]. Monitoring inherent patterns in these behaviours at the individual level, over time, to detect changes and out-of-the-ordinary behaviours, is helpful in detecting the early onset of depression [3, 2, 26]. The findings in this study revealed some relationship between anomalous human behaviour and questions 3, 4, 5, 7 and 8 of the PHQ-9 depression scale, which suggests that it is feasible to passively and unobtrusively collect datasets from smartphones, to detect out of the ordinary behaviours related to depression. These findings corroborate the findings of previous studies [79, 55, 3] with regards to the relationship between biomarkers and depressive symptoms, particularly of typing patterns [79, 55], sleep disturbances and fatigue [79], trouble concentrating and psychomotor agitation or retardation [79].

In a hypothetical early detection of depression system, for example, the anomalous behaviour detection system will monitor the increase of detected anomalies in a participants dataset that correlate with depressive symptoms. If

a reasonable threshold is exceeded, the system will then prompt the participant to provide a response to a self-reported depression scale such as PHQ-9 [44], Beck Depression Inventory (BDI) [4] and Depression Anxiety and Stress Scales (DASS-21) [58]. The score of the depression scale will determine the requisite course of action.

While it was not the goal of the current study to provide depression interventions to participants, our results suggest that participants expected to see feedback on the study's application. Equally important is the impact of such feedback on participants' behaviour. Previous research [45, 20, 66] have investigated various feedback mechanisms when providing intervention in mental health systems. Careful design of the feedback mechanism [66] regarding timing, frequency and personalisation is required to prevent a negative impact [20] on the participants' mental health state.

In addition, to prompt actionable feedback [3, 66], the feedback should be specific and personalised, interpretable, meaningful to participants. Model interpretability is one advantage anomaly detection methods used in this study have over predictive analysis, whose output interpretation is sometimes challenging [68, 17]. Our results suggest that, while some participants could not always recall if something of significance happened on anomalous days, the detected anomalous behaviours (example in Figure 2) were generally plausible and meaningful to the participants. With prompt and interpretable feedback on detected anomalous behaviours, participants could provide additional context to the dataset by annotating detected behaviours.

Our results also highlighted some application deployment challenges with some participants. In recent years, application distribution platforms such as the Google Play Store [29] have enabled mobile health researchers and application developers to reach millions of people using different smartphone devices and OS versions globally. Not only do these application distribution platforms bring opportunities, they also bring some challenges to application development and deployment [36, 5] due to ever evolving device platforms and application deployment guidelines. For instance, recent changes in Google Play Store's application publishing policies [30] restricted the types of applications that can access specific permissions, including SMS and call logging permissions. Consequently, our attempt to publish Me on the Google Play Store was rejected since Me is not a replacement for calls or SMS applications, yet it needs SMS and call permissions to access the call and SMS behaviour data. Consequently, we hosted Me in-house using Jenkins [41]. However, self-hosting applications pose challenges in scaling and updating the application on multiple devices, which are otherwise handled by the application distribution platform. Lastly, Android's Doze and background processing limitations negatively interfere with passive and continuous sensing applications like in Me, justifying why at certain times, the application did not remind the participant.

8 Limitations and Future Work

Notwithstanding the results in this study, the number of participants in the study is limited and were recruited from a general population. With the study being exploratory, evidence of clinical diagnosis with depression was not a recruitment requirement. Secondly, all correlations reported in this study are not statistically significant when their *P values* are adjusted for multiple testing.

In the future work, we replicate the study with a larger cohort drawn from a clinical population to explore further the relationship between out of the ordinary behaviours and depression. With a larger sample size, future research could review the statistically significant relationship between detected anomalies and depression. We could further explore the causal relationship between detected anomalies and depression. Since correlations only reveal linear relationships. In addition, we could investigate both linear and non-linear relationships between depression and detected anomalies, using information theory methods such as Mutual Information (MI) [15]. Additionally, future work could expand the features extracted from the smartphone dataset to include additional measures of routines and variability in human behaviour using methods such as Regularity Index [76, 80]. Future work could improve the anomaly detection system with other contextual information such as the Big Five Personality traits, and explore the relationship between anomalous human behaviour and depression using other depression scales such as the BDI [4], DASS-21 [58].

The current study is an exploratory first step towards creating a system for just-in-time depression intervention with anomaly detection methods. As such, the current study does not predict whether an individual is depressed or not. The findings from this study naturally lead to questions such as; how much data will be collected from individuals to constitute a ground truth for their baseline behaviour, how would the system adapt to changing human behaviours such as seasonal changes(e.g winter and summer), situational changes such as changing jobs, graduating from college, and other environmental changes that may change human behaviour?, how many or what kind of anomalous behaviours will be statistically significant with depression scores, What threshold of detected anomalies will trigger an intervention from the application. These are questions we seek to investigate in future work, ultimately leading to the creation of a system of personalised and labelled datasets of anomalous individual behaviours that are indicative of depression, and personalised models for just-in-time depression interventions.

9 Conclusion

In this study, we investigated the feasibility of using unsupervised multivariate anomaly detection methods to detect at the early onset and monitor the progression of depression. Our quantitative and qualitative findings show that anomalies detected in participants' behaviour collected via smartphone sensing over a 4 weeks period, represented specific and meaningful out of the ordinary behaviour. Our findings also show non statistically significant correlation between

the detected anomalous human behaviour and various symptoms of depression under the PHQ-9 depression scale. In spite of our study's limitations, our findings demonstrate a step forward towards detecting and monitoring depression with anomaly detection methods. Further research is needed to replicate these findings in larger population studies, potentially leading to creating just-in-time interventions for depression using anomaly detection methods.

Acknowledgments

The *Me in the Wild* study is supported by the Academy of Finland SENSATE (Grant Nos. 316253, 320089), 6Genesis Flagship (Grant No. 318927), and the Infotech Institute University of Oulu Emerging Project. We thank all the participants of the *Me in the Wild* study.

References

1. Acoustics, I.: Comparitive Examples of Noise Levels | Industrial Noise Control (Jan 2020), https://www.industrialnoisecontrol.com/comparative-noise-examples.htm
2. Adler, D.A., Ben-Zeev, D., Tseng, V.W., Kane, J.M., Brian, R., Campbell, A.T., Hauser, M., Scherer, E.A., Choudhury, T.: Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. JMIR mHealth and uHealth **8**(8), e19962 (2020). https://doi.org/10.2196/19962
3. Barnett, I., Torous, J., Staples, P., Sandoval, L., Keshavan, M., Onnela, J.P.: Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology **43**(8), 1660 (2018). https://doi.org/10/gdrks3
4. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. Archives of general psychiatry **4**(6), 561–571 (1961). https://doi.org/10.1001/archpsyc.1961.01710120031004
5. Ben-Zeev, D., Schueller, S.M., Begale, M., Duffecy, J., Kane, J.M., Mohr, D.C.: Strategies for mhealth research: lessons from 3 mobile intervention studies. Administration and Policy in Mental Health and Mental Health Services Research **42**(2), 157–167 (2015)
6. van Berkel, N., Ferreira, D., Kostakos, V.: The experience sampling method on mobile devices. ACM Comput. Surv. **50**(6), 93:1–93:40 (Dec 2017). https://doi.org/10.1145/3123988, http://doi.acm.org/10.1145/3123988
7. Beygelzimer, A., library), S.K.a.J.L.c.t., approach), S.A.a.D.M., Li, S.: FNN: Fast Nearest Neighbor Search Algorithms and Applications (Feb 2019), https://CRAN.R-project.org/package=FNN
8. Böhmer, M., Hecht, B., Schöning, J., Krüger, A., Bauer, G.: Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In: Proceedings of the 13th international conference on Human computer interaction with mobile devices and services. pp. 47–56 (2011)
9. Bonful, H.A., Anum, A.: Sociodemographic correlates of depressive symptoms: a cross-sectional analytic study among healthy urban ghanaian women. BMC public health **19**(1), 50 (2019)
10. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: ACM sigmod record. vol. 29, pp. 93–104. ACM (2000)

11. Charmaz, K., Belgrave, L., et al.: Qualitative interviewing and grounded theory analysis. The SAGE handbook of interview research: The complexity of the craft **2**, 347–365 (2012)
12. Coravos, A., Khozin, S., Mandl, K.D.: Developing and adopting safe and effective digital biomarkers to improve patient outcomes. npj Digital Medicine **2**(1), 1–5 (Mar 2019). https://doi.org/10.1038/s41746-019-0090-4
13. Croux, C., Rousseeuw, P.J.: Time-efficient algorithms for two highly robust estimators of scale. In: Computational statistics, pp. 411–428. Springer (1992)
14. Dagum, P.: Digital biomarkers of cognitive function. npj Digital Medicine **1**(1), 1–3 (Mar 2018). https://doi.org/10.1038/s41746-018-0018-4
15. Dionisio, A., Menezes, R., Mendes, D.A.: Mutual information: a measure of dependency for nonlinear time series. Physica A: Statistical Mechanics and its Applications **344**(1-2), 326–329 (2004)
16. Dorsey, E.R., Papapetropoulos, S., Xiong, M., Kieburtz, K.: The first frontier: Digital biomarkers for neurodegenerative disorders. Digital Biomarkers **1**(1), 6–13 (2017). https://doi.org/10.1159/000477383
17. Elshawi, R., Al-Mallah, M.H., Sakr, S.: On the interpretability of machine learning-based model for predicting hypertension. BMC medical informatics and decision making **19**(1), 146 (2019)
18. Eric, G.: iForest: Isolation Forest Anomaly Detection (Aug 2019), https://rdrr.io/github/Zelazny7/isofor/man/iForest.html
19. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996)
20. Faurholt-Jepsen, M., Frost, M., Ritz, C., Christensen, E.M., Jacoby, A., Mikkelsen, R.L., Knorr, U., Bardram, J., Vinberg, M., Kessing, L.V.: Daily electronic self-monitoring in bipolar disorder using smartphones–the monarca i trial: a randomized, placebo-controlled, single-blind, parallel group trial. Psychological medicine **45**(13), 2691–2704 (2015)
21. Ferreira, D., Kostakos, V., Dey, A.K.: Aware: mobile context instrumentation framework. Frontiers in ICT **2**, 6 (2015)
22. Ferreira, D., Kostakos, V., Schweizer, I.: Human Sensors on the Move, p. 9–19. Springer International Publishing (2017). https://doi.org/10.1007/978-3-319-25658-0_1
23. Fraccaro, P., Beukenhorst, A., Sperrin, M., Harper, S., Palmier-Claus, J., Lewis, S., Van der Veer, S.N., Peek, N.: Digital biomarkers from geolocation data in bipolar disorder and schizophrenia: a systematic review. Journal of the American Medical Informatics Association **26**(11), 1412–1420 (Nov 2019). https://doi.org/10.1093/jamia/ocz043
24. Fried, E.I., Nesse, R.M.: Depression is not a consistent syndrome: an investigation of unique symptom patterns in the star* d study. Journal of affective disorders **172**, 96–102 (2015). https://doi.org/10.1016/j.jad.2014.10.010
25. Fried, E.I., Nesse, R.M.: Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC medicine **13**(1), 72 (2015)
26. Gerych, W., Agu, E., Rundensteiner, E.: Classifying depression in imbalanced datasets using an autoencoder- based anomaly detection approach. In: 2019 IEEE 13th International Conference on Semantic Computing (ICSC). p. 124–127 (Jan 2019). https://doi.org/10.1109/ICOSC.2019.8665535
27. Goldberg, L.R.: The development of markers for the big-five factor structure. Psychological assessment **4**, 26 (1992)

28. Goldstein, M., Uchida, S.: A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one **11**(4), e0152173 (2016)
29. Google: Google Play (Dec 2019), https://play.google.com/store?hl=en%5FGB
30. Google: Use of SMS or Call Log permission groups - Play Console Help (Dec 2019), https://support.google.com/googleplay/android-developer/answer/9047303?hl=en
31. Greenberg, P.E., Fournier, A.A., Sisitsky, T., Pike, C.T., Kessler, R.C.: The economic burden of adults with major depressive disorder in the united states (2005 and 2010). The Journal of Clinical Psychiatry **76**(2), 155–162 (Feb 2015). https://doi.org/10.4088/JCP.14m09298
32. Hamilton, M.: A rating scale for depression. Journal of neurology, neurosurgery, and psychiatry **23**(1), 56 (1960)
33. Harari, G.M., Lane, N.D., Wang, R., Crosier, B.S., Campbell, A.T., Gosling, S.D.: Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. Perspectives on psychological science : a journal of the Association for Psychological Science **11**(6), 838–854 (Nov 2016). https://doi.org/10.1177/1745691616650285
34. Hemmerle, A.M., Herman, J.P., Seroogy, K.B.: Stress, depression and parkinson's disease. Experimental neurology **233**(1), 79–86 (2012). https://doi.org/https://doi.org/10.1016/j.expneurol.2011.09.035
35. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics pp. 65–70 (1979)
36. Holzer, A., Ondrus, J.: Mobile application market: A developer's perspective. Telematics and informatics **28**(1), 22–31 (2011)
37. Hu, Y., Shan, W.M.a.Y., Australia: Rlof: R Parallel Implementation of Local Outlier Factor(LOF) (Sep 2015), https://CRAN.R-project.org/package=Rlof
38. Huber, P.J.: Robust statistics. Springer (2011)
39. Huckvale, K., Venkatesh, S., Christensen, H.: Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. npj Digital Medicine **2**(1), 1–11 (Sep 2019). https://doi.org/10.1038/s41746-019-0166-1
40. Jacobson, N.C., Weingarden, H., Wilhelm, S.: Digital biomarkers of mood disorders and symptom change. npj Digital Medicine **2**(1), 1–3 (Feb 2019). https://doi.org/10.1038/s41746-019-0078-0
41. Jenkins.io: Jenkins and android (Jan 2019), https://jenkins.io/solutions/android/index.html
42. Klobas, J.E., McGill, T.J., Moghavvemi, S., Paramanathan, T.: Compulsive youtube usage: A comparison of use motivation and personality effects. Computers in Human Behavior **87**, 129–139 (2018)
43. Kourtis, L.C., Regele, O.B., Wright, J.M., Jones, G.B.: Digital biomarkers for alzheimer's disease: the mobile/wearable devices opportunity. npj Digital Medicine **2**(1), 1–9 (Feb 2019). https://doi.org/10.1038/s41746-019-0084-2
44. Kroenke, K., Spitzer, R.L., Williams, J.B.: The phq-9: validity of a brief depression severity measure. Journal of general internal medicine **16**(9), 606–613 (2001), https://doi.org/10.1046/j.1525-1497.2001.016009606.x
45. Lee, J., Lam, M., Chiu, C.: Clara: Design of a new system for passive sensing of depression, stress and anxiety in the workplace. In: Cipresso, P., Serino, S., Villani, D. (eds.) Pervasive Computing Paradigms for Mental Health. pp. 12–28. Springer International Publishing, Cham (2019)
46. Lépine, J.P., Briley, M.: The increasing burden of depression. Neuropsychiatric disease and treatment **7**(Suppl 1), 3 (2011). https://doi.org/10.2147/NDT.S19617

47. Liang, Y., Zheng, X., Zeng, D.D.: A survey on big data-driven digital phenotyping of mental health. Information Fusion **52**, 290–307 (2019)
48. Liao, Z., Kong, L., Wang, X., Zhao, Y., Zhou, F., Liao, Z., Fan, X.: A Visual Analytics Approach for Detecting and Understanding Anomalous Resident Behaviors in Smart Healthcare. Applied Sciences **7**(3), 254 (Mar 2017). https://doi.org/10.3390/app7030254, https://www.mdpi.com/2076-3417/7/3/254
49. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD) **6**(1), 3 (2012)
50. Lotfi, A., Langensiepen, C., Mahmoud, S.M., Akhlaghinia, M.J.: Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour. Journal of ambient intelligence and humanized computing **3**(3), 205–218 (2012)
51. Macchia, A., Monte, S., Pellegrini, F., Romero, M., D'Ettorre, A., Tavazzi, L., Tognoni, G., Maggioni, A.P.: Depression worsens outcomes in elderly patients with heart failure: an analysis of 48,117 patients in a community setting. European journal of heart failure **10**(7), 714–721 (2008)
52. Madsen, J.H.: Connectivity-based Outlier Factor (COF) algorithm in DDoutlier: Distance & Density-Based Outlier Detection (May 2019), https://rdrr.io/cran/DDoutlier/man/COF.html
53. Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M.: robustbase: Basic Robust Statistics (May 2019), https://CRAN.R-project.org/package=robustbase
54. Mandryk, R.L., Birk, M.V.: The potential of game-based digital biomarkers for modeling mental health. JMIR Mental Health **6**(4), e13485 (2019). https://doi.org/10.2196/13485
55. Mastoras, R.E., Iakovakis, D., Hadjidimitriou, S., Charisis, V., Kassie, S., Alsaadi, T., Khandoker, A., Hadjileontiadis, L.J.: Touchscreen typing pattern analysis for remote detection of the depressive tendency. Scientific Reports **9**(1), 1–12 (Sep 2019). https://doi.org/10.1038/s41598-019-50002-9
56. Meister, S., Deiters, W., Becker, S.: Digital health and digital biomarkers – enabling value chains on health data. Current Directions in Biomedical Engineering **2**(1), 577–581 (2016). https://doi.org/10.1515/cdbme-2016-0128
57. Moshe, I., Terhorst, Y., Opoku Asare, K., Sandar, L.B., Ferreira, D., Baumeister, H., Mohr, D., Pulkki-Råback, L.: Predicting symptoms of depression and anxiety using smartphone and wearable data. Frontiers in Psychiatry **12** (2021). https://doi.org/10.3389/fpsyt.2021.625247
58. Norton, P.J.: Depression anxiety and stress scales (dass-21): Psychometric analysis across four racial groups. Anxiety, stress, and coping **20**(3), 253–265 (2007)
59. Opoku Asare, K., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., Ferreira, D.: Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: Exploratory study. JMIR mHealth and uHealth **9**(7), e26540 (Jul 2021). https://doi.org/10.2196/26540
60. Opoku Asare, K., Visuri, A., Ferreira, D.S.T.: Towards early detection of depression through smartphone sensing. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers. pp. 1158–1161. UbiComp/ISWC '19 Adjunct, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3341162.3347075
61. Peltonen, E., Sharmila, P., Opoku Asare, K., Visuri, A., Lagerspetz, E., Ferreira, D.: When phones get personal: Predicting big five personality traits from application usage. Pervasive and Mobile Computing **69**, 101269 (2020)

62. van der Ploeg, T., Austin, P.C., Steyerberg, E.W.: Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC medical research methodology **14**(1), 137 (2014)
63. Rodarte, C.: Pharmaceutical Perspective: How Digital Biomarkers and Contextual Data Will Enable Therapeutic Environments. Digital Biomarkers **1**(1), 73–81 (2017). https://doi.org/10.1159/000479951
64. Rohani, D.A., Faurholt-Jepsen, M., Kessing, L.V., Bardram, J.E.: Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: Systematic review. JMIR mHealth and uHealth **6**(8), e165 (2018). https://doi.org/10.2196/mhealth.9691
65. Saeb, S., Zhang, M., Kwasny, M., Karr, C.J., Kording, K., Mohr, D.C.: The relationship between clinical, momentary, and sensor-based assessment of depression. In: 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth). pp. 229–232. IEEE (2015)
66. Schembre, S.M., Liao, Y., Robertson, M.C., Dunton, G.F., Kerr, J., Haffey, M.E., Burnett, T., Basen-Engquist, K., Hicklen, R.S.: Just-in-time feedback in diet and physical activity interventions: systematic review and practical design framework. Journal of medical Internet research **20**(3), e106 (2018)
67. Shannon, C.E.: A mathematical theory of communication. Bell system technical journal **27**(3), 379–423 (1948)
68. Shmueli, G., Koppius, O.R.: Predictive analytics in information systems research. MIS Quarterly **35**(3), 553–572 (2011), http://www.jstor.org/stable/23042796
69. Sordo, M., Zeng, Q.: On sample size and classification accuracy: a performance comparison. In: International Symposium on Biological and Medical Data Analysis. pp. 193–201. Springer (2005)
70. Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., Rentfrow, J.: Passive mobile sensing and psychological traits for large scale mood prediction. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare. pp. 272–281. PervasiveHealth'19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3329189.3329213
71. Stachl, C., Au, Q., Schoedel, R., Gosling, S.D., Harari, G.M., Buschek, D., Völkel, S.T., Schuwerk, T., Oldemeier, M., Ullmann, T.: Predicting personality from patterns of behavior collected with smartphones. Proceedings of the National Academy of Sciences **117**(30) (2020). https://doi.org/10.1073/pnas.1920484117
72. Strimmer, J.H.a.K.: entropy: Estimation of Entropy, Mutual Information and Related Quantities (Nov 2014), https://CRAN.R-project.org/package=entropy
73. Strober, L.B., Arnett, P.A.: Assessment of depression in three medically ill, elderly populations: Alzheimer's disease, parkinson's disease, and stroke. The Clinical Neuropsychologist **23**(2), 205–230 (2009)
74. Tang, J., Chen, Z., Fu, A.W.C., Cheung, D.W.: Enhancing effectiveness of outlier detections for low density patterns. In: Advances in Knowledge Discovery and Data Mining. pp. 535–548. Springer Berlin Heidelberg, Berlin, Heidelberg (2002)
75. TENK: Guidelines for ethical review in human sciences, https://tenk.fi/en/advice-and-materials/guidelines-ethical-review-human-sciences
76. Tseng, V.W.S., Sano, A., Ben-Zeev, D., Brian, R., Campbell, A.T., Hauser, M., Kane, J.M., Scherer, E.A., Wang, R., Wang, W., et al.: Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. Scientific reports **10**(1), 1–17 (2020)

77. Vega, J., Jay, C., Vigo, M., Harper, S.: Unobtrusive monitoring of parkinson's disease based on digital biomarkers of human behaviour. In: Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility. p. 351–352. ASSETS '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3132525.3134782
78. Wagner, D.T., Rice, A., Beresford, A.R.: Device analyzer: Large-scale mobile data collection. SIGMETRICS Perform. Eval. Rev. **41**(4), 53–56 (Apr 2014). https://doi.org/10.1145/2627534.2627553
79. Wang, R., Wang, W., daSilva, A., Huckins, J.F., Kelley, W.M., Heatherton, T.F., Campbell, A.T.: Tracking depression dynamics in college students using mobile phone and wearable sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(1), 43 (2018)
80. Wang, W., Harari, G.M., Wang, R., Müller, S.R., Mirjafari, S., Masaba, K., Campbell, A.T.: Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(3), 1–21 (2018)
81. WHO: Depression (Mar 2018), https://www.who.int/news-room/fact-sheets/detail/depression
82. Wright, B., Peters, E., Ettinger, U., Kuipers, E., Kumari, V.: Understanding noise stress-induced cognitive impairment in healthy adults and its implications for schizophrenia. Noise and Health **16**(70), 166–176 (2014)